Interpretation of variation in omics data

Applications in proteomics for sustainable agriculture

Willforss, Jakob

2020

Interpretation of variation in omics data

Applications in proteomics for sustainable agriculture

JAKOB WILLFORSS | DEPARTMENT OF IMMUNOTECHNOLOGY
FACULTY OF ENGINEERING | LUND UNIVERSITY

New technologies can measure thousands of molecules in cells. These are used to solve among the biggest challenges facing us today in fields such as agriculture and medicine. The first part of this work introduces two new computer programs which make it easier to draw accurate conclusions from this kind of complex data. The second part of this work study how proteomics – the measurement of proteins in cells – can be used to speed up the breeding of important agricultural traits.

Printed by Media-Tryck, Lund 2020 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

Faculty of Engineering
Department of Immunotechnology

ISBN 978-91-7895-641-8

Interpretation of variation in omics data

Interpretation of variation in omics data

Applications in proteomics for sustainable agriculture

by Jakob Willforss

LUND
UNIVERSITY

DOCTORAL DISSERTATION
by due permission of the Faculty of Engineering, Lund University, Sweden
To be defended at Hörsalen, Medicon Village, Scheelevägen 2, Lund
Friday December 11th at 9:00.

Faculty opponent

Prof. Laura Elo
Turku Bioscience Centre, University of Turku
Turku, Finland

Organization
LUND UNIVERSITY

Department of Immunotechnology
Medicon Village (building 406)
SE–223 87 LUND
Sweden

Document name
DOCTORAL DISSERTATION

Date of disputation
2020-12-11

Sponsoring organization

Author(s)
Jakob Willforss

Title and subtitle
Interpretation of variation in omics data: Applications in proteomics for sustainable agriculture

Abstract
Biomarkers are used in molecular biology to predict characteristics of interest and are applied in agriculture to accelerate the breeding of target traits. Proteomics has emerged as a promising technology for improved markers by providing a closer view to the phenotype than conventional genome-based approaches. However, a major challenge for biomarker development is that the identified biological patterns often cannot be reproduced in other studies. One piece of the puzzle to alleviate this problem is improved software approaches to distinguish biological variation from noise in the data.

In this work, two new pieces of software are introduced to facilitate interpretation of data from omic experiments. NormalyzerDE (Paper I) helps the user to perform an informed selection of a well-performing normalization technique, presents a new type of normalization for electrospray intensity variation biases and gives a user-friendly approach to performing subsequent statistical analysis. OmicLoupe (Paper II) provides interactive visualizations of up to two omics datasets, introduces novel approaches for the comparison of different datasets and provides the ability to rapidly inspect individual features. These pieces of software were applied together with existing methods to study three agricultural organisms. Firstly, a proteogenomic approach was used to study *Fusarium* head blight in oat. This study provided the deepest proteomic resource to date in this organism (Paper III) and identified proteins related to a differential resistance towards *Fusarium* head blight. It can contribute towards the development of commercial varieties with improved resistance towards this pathogen. Secondly, bull seminal plasma was studied to identify proteins correlated with fertility, which are also robust to seasonal variation (Paper IV). This study contributes towards ensuring maintained high fertility in livestock. Finally, potato grown at sites in northern and southern Sweden (Paper V) were studied to identify proteins linked to the different growth conditions at the two locations. This study contributes towards a better understanding of the molecular physiology in the agricultural field and the selection of varieties better adapted to the different growth conditions.

In conclusion, these results contribute towards improved analyses of omics data and to biomarkers with potential applications in accelerated breeding in the studied organisms. Together, this could provide tools for the development of a more sustainable agriculture.

Key words
agriculture, proteomics, omics, biomarker, normalization, batch effect, visualization, software

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title

ISBN
978-91-7895-641-8 (print)
978-91-7895-640-1 (pdf)

Recipient's notes

Number of pages
232

Price

Security classification

Signature 

Date 2020-11-02

Interpretation of variation in omics data

Applications in proteomics for sustainable agriculture

by Jakob Willforss

LUND
UNIVERSITY

Cover illustration front: The cover art is drawn by Zuzanna Sadowska.
Funding information: This work was financially supported by Mistra Biotech.

Part of Figure 7 produced using biorender.com

Faculty of Engineering, Department of Immunotechnology

ISBN: 978-91-7895-641-8 (print)
ISBN: 978-91-7895-640-1 (pdf)

Printed in Sweden by Media-Tryck, Lund University, Lund 2020

Media-Tryck is a Nordic Swan Ecolabel certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se
MADE IN SWEDEN

The first principle is that you must not fool yourself –
and you are the easiest person to fool
Richard Phillips Feynman

Contents

List of publications

This thesis is based on the following publications, referred to by their Roman numerals:

I **NormalyzerDE: Online Tool for Improved Normalization of Omics Expression Data and High-Sensitivity Differential Expression Analysis**

J. Willforss, A. Chawade, F. Levander
Journal of Proteome Research (2019), 18 (2), pp. 732–740
https://doi.org/10.1021/acs.jproteome.8b00523

II **OmicLoupe: Facilitating biological discovery by interactive exploration of multiple omic datasets and statistical comparisons**

J. Willforss, V. Siino, F. Levander
Submitted manuscript with preprint available on bioRxiv
https://doi.org/10.1101/2020.10.22.349944

III **Interactive proteogenomic exploration of response to *Fusarium* head blight in oat varieties with different resistance**

J. Willforss, S. Leonova, J. Tillander, E. Andreasson, S. Marttila, O. Olsson, A. Chawade, F. Levander
Journal of Proteomics (2020), 218:103688
https://doi.org/10.1016/j.jprot.2020.103688

IV **Stable bull fertility protein markers in seminal plasma**

J. Willforss, J.M. Morrell, S. Resjö, T. Hallap, P. Padrik, V. Siino, D.J. de Koning, E. Andreasson, F. Levander, P. Humblot
Submitted manuscript in review

V **Comparative proteomic analyses of potato leaves from field-grown plants grown under extremely long days**

S. Resjö*, **J. Willforss***, A. Large, V. Siino, E. Alexandersson, F. Levander, E. Andreasson (*Shared first authors)
Manuscript

My contributions to papers

Paper I: NormalyzerDE: Online Tool for Improved Normalization of Omics Expression Data and High-Sensitivity Differential Expression Analysis

Further developed previously outlined study design, carried out the software development, performed data analysis and interpreted the data, drafted the manuscript.

Paper II: OmicLoupe: Facilitating biological discovery by interactive exploration of multiple omic datasets and statistical comparisons

Designed study, carried out the software development and the majority of the data analysis, shared the data interpretation, drafted the manuscript.

Paper III: Interactive proteogenomic exploration of response to *Fusarium* head blight in oat varieties with different resistance

Performed majority of data analysis and interpreted the data, carried out the software development, shared the biological interpretation, drafted the manuscript.

Paper IV: Stable bull fertility protein markers in seminal plasma

Participated in the study design, analysed and interpreted the data, shared the biological interpretation, drafted the manuscript.

Paper V: Comparative proteomic analyses of potato leaves from field-grown plants grown under extremely long days

Participated in the study design, analysed and interpreted the data, took part in writing the manuscript.

List of publications not included

I **Patient-Derived Xenograft Models Reveal Intratumor Heterogeneity and Temporal Stability in Neuroblastoma**

N. Braekeveldt, K. Stedingk, S. Fransson, A. Martinez-Monleon, D. Lindgren, H. Axelson, F. Levander, **J. Willforss**, K. Hansson, I. Øra, T. Backman, A. Börjesson, S. Beckman, J. Esfandyari, A. Berbegall, R. Noguera, J. Karsson, J. Koster, T. Martinsson, D. Gisselsson, S. Påhlman, D. Bexell
Cancer Research (2018), 78 (20), pp. 5958–5969
https://doi.org/10.1158/0008-5472.CAN-18-0527

II **Identification of genes regulating traits targeted for domestication of field cress (*Lepidium campestre*) as a biennial and perennial oilseed crop**

C. Gustafsson, **J. Willforss**, F. Lopes-Pinto, R. Ortiz, M. Geleta
BMC genetics (2018), 19 (1), 36
https://doi.org/10.1186/s12863-018-0624-9

III **RNA seq analysis of potato cyst nematode interactions with resistant and susceptible potato roots**

A.J. Walter, **J. Willforss**, M. Lenman, E. Alexandersson, E. Andreasson
European journal of plant pathology (2018), 152 (2), 531–539
https://doi.org/10.1007/s10658-018-1474-z

Abbreviations and explanations

DDA ..... Data Dependent Acquisition

DIA ..... Data Independent Acquisition

DON ..... Deoxynivalenol (Toxin produced by *Fusarium* species)

eQTL ..... Expression Quantitative Trait Locus

ESI ..... Electrospray ionization (Technique to ionize peptides)

FDR ..... False Discovery Rate

FHB ..... *Fusarium* Head Blight (Disease caused by *Fusarium* species)

LC ..... Liquid Chromatography

m/z ..... Mass-to-charge ratio

MS ..... Mass Spectrometry

PCA ..... Principal Component Analysis

PTM ..... Post Translational Modification

QTL ..... Quantitative Trait Locus (Region in genome linked to trait)

RT ..... Retention Time

Introduction

The agriculture of today faces challenges of sustaining the world's food production for a growing number of people during a changing climate (Ruane et al. 2018). Biomarkers, biological characteristics that can be measured to predict traits of interest in organisms of interest, have emerged as a valuable tool for accelerating agricultural and medical research. In recent years, molecular biomarkers have played an important role in molecular breeding (Nadeem et al. 2018), where it is used to predict traits and guide breeding decisions. Here, these biomarkers can help solve the sustainability challenges facing the agriculture by accelerating our ability to shape our food.

DNA-based markers are currently the most established technology for molecular breeding and have been used in various applications (Xie and Xu 1998). These markers are relatively easy to measure but are hindered by the complexity of cellular biology, where the information in the DNA needs to be translated through transcripts into proteins before having a function in the organism. Both transcripts and proteins have shown the potential to improve DNA marker-based predictions (Langridge and Fleury 2011; Holloway and Li 2010), with proteins having the biggest potential predictive ability since they are closest to the function. Still, proteomics, the large-scale study of proteins, is less established and needs to overcome many challenges before being widely used in molecular breeding. Included in these challenges is the fact that proteomics has a more complex workflow both in terms of laboratory procedure and data analysis which leads to a higher degree of variation between samples and experiments, making reproducibility between studies more challenging (Piehowski et al. 2013). Many of these challenges are being addressed, and recent publications such as those presented in this work have shown the potential of proteomics for further improving current breeding techniques (Ma, Rahmat and Lam 2013; Sandin, Chawade and Levander 2015).

Bias caused by technical variation (unwanted noise introduced by variation in laboratory procedures) in proteomics experiments often makes it difficult to find biological patterns of interest. To solve this, two pieces of software were developed to reduce the impact of technical variation and are presented in this work. Here this is achieved by directly reducing

technical variation by using a technique called normalization and by providing visualizations that help the user identify the best performing analysis methods for their dataset (**Paper I-II**). These pieces of software and approaches were then applied in three proteomic studies on three different agricultural organisms (**Paper III-V**) to identify proteins linked to their respective traits of interest. A summary of these studies is shown in Figure 1.

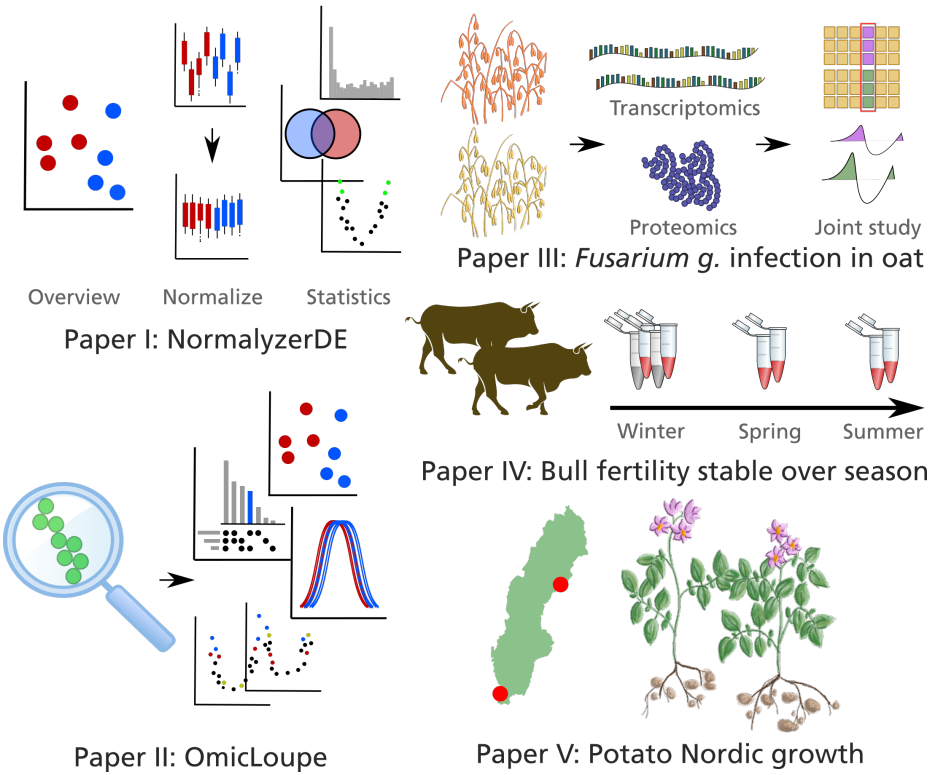


Figure 1: Overview of projects presented in this thesis. Papers I-II introduce software to improve existing analysis approaches in omics. Papers III-V present applied studies investigating the proteome response related to important agricultural traits.

For the software studies, in **Paper I** the software NormalyzerDE was developed. It helps the user carry out an optimally performing normalization of their dataset, a procedure to reduce certain types of unwanted variation. Furthermore, NormalyzerDE presents a new normalization approach reducing variation caused during the ionization of peptides in the mass spectrometer. Finally, it conveniently provides tools for executing and visualizing the downstream statistical analysis. NormalyzerDE has gained a wide userbase, and simplifies the selection of an optimal analysis approach by being accessible on a web server and as a Bioconductor R package. **Paper II** introduces the newly developed software OmicLoupe, an interactive and easy-to-use software for rapid visualization of omics data. It provides

visualizations for sample quality and statistical aspects, and introduces new approaches to compare data from different experiments or types of omics, revealing shared trends potentially missed using conventional methods. OmicLoupe can help the user understand limitations and see opportunities in the data at hand, thus guiding better analysis decisions and the identification of proteins or other features of interest.

For the agricultural studies presented in this work, in **Paper III** we used proteomics together with transcriptomics based references to study the molecular response in oat when infected by the fungal pathogen *Fusarium graminearum*. This pathogen causes the disease *Fusarium* head blight (FHB) and upon infection emits a toxin called deoxynivalenol (DON) which when ingested affects the health of both human and livestock (Alshannaq and Yu 2017; Wu, Groopman and Pestka 2014). The response to the disease was investigated in two varieties of oat with different resistance to FHB. Our study confirms the differential response to infection between the oat varieties and identifies proteins affected upon infection. In **Paper IV** we study bull fertility by analysing the proteomic profile in seminal bull plasma from a set of individuals with different fertilities. Estimating bull fertility by traditional means is slow and costly as it requires awaiting fertile age and performing enough inseminations to get reliable estimates (Humblot, Decoux and Dhorne 1991). The seminal plasma proteome has been shown to play a role in the fertility in bulls (Druart et al. 2019). Here we identified proteins consistently correlated across three separate measurements and seasons, contributing to the identification of markers of bull fertility, which could be used to detect bulls with low fertility at an early stage, saving considerable resources. Finally, in **Paper V** we study the proteome of potatoes grown at different latitudes in Sweden. In northern Sweden, the days are longer and the growing season is shorter. Here we study how growth location impacts the proteome of different varieties of potato. This identified proteins with consistently different abundances across three years in one potato variety and between groups of varieties with varying yields at the two sites.

In conclusion, this work introduces new software to improve the analysis of omics datasets, providing a foundation for better analysis decisions. The three agricultural proteomics studies identify proteins linked to different phenotypes, contributing to potential biomarkers and accelerated breeding in the three organisms. In the present thesis, I have based on these studies chosen to highlight the limitations and considerations one needs to consider when carrying out proteomics biomarker discovery studies. Many of these apply generally when working with omics-data. These considerations are, in my view, among the most valuable insights gained through this work. By contributing improved methods to work with omics data and by increasing the molecular knowledge about important agricultural traits, this work aims to increase our ability to shape our food towards a more sustainable agriculture.

Thesis aims

The aim of the work presented in this thesis is to improve the methodologies available for interpretation of omics data to allow for the implementation of omics analysis methods within the field of sustainable agriculture. To accomplish this aim, there are two objectives:

1. Identify limitations in existing omics data processing workflows and develop new methods to overcome these limitations.
2. Apply existing and newly developed omics-methodologies to identify proteins linked to traits of interest in three diverse agricultural organisms, and by doing this, contribute towards new biomarkers.

Chapter 1: From experiment to proteins

Proteomics biomarker discovery studies are long journeys consisting of many steps that influence the reliability of the final result. This type of study typically involves scientists with different expertise, including experimentalists, mass spectrometrists, data analysts, and biologists. These scientists need to work together and communicate about what opportunities and issues have appeared during the project. Doing so efficiently requires an understanding of both the upstream and downstream steps of the performed work. This chapter summarizes the initial experimental and computational steps, including the experimental design, while highlighting potential limitations it may cause on the subsequent data analysis.

The journey of a proteomics biomarker study begins at the drawing board, where the structure of the experiment is outlined - the experimental design. Then the experiment starts - the field trials are grown, or the tissue samples are collected. A sequence of experimental steps is carried out, starting from protein extraction, followed by protein digestion into peptides and finally measuring these peptides in the mass spectrometer. These measurements produce large amounts of mass spectra - accurately measured mass over charge ratios of peptides and peptide fragments. These spectra are computationally processed using specialized software, piecing them back together into a comprehensive view of what proteins were originally present in each sample and in what amount. A schematic of this workflow is illustrated in Figure 2. Each step of this workflow is addressed in this chapter.

Different types of variation will appear during this journey, causing uncertainties to the final estimates of which protein were originally present and in which abundance. The variation can be biological - caused before sampling by differences during the biological experiment itself, or technical - caused by inaccuracies in the sample handling and in the mass spectrometer. These variations can systematically influence groups of samples causing what is called batch effects, or randomly influence individual samples differently, called random effects. In this thesis, this collection of undesirable biases is jointly called *unwanted variation*. Ideally, these variations should be accounted for already at the design

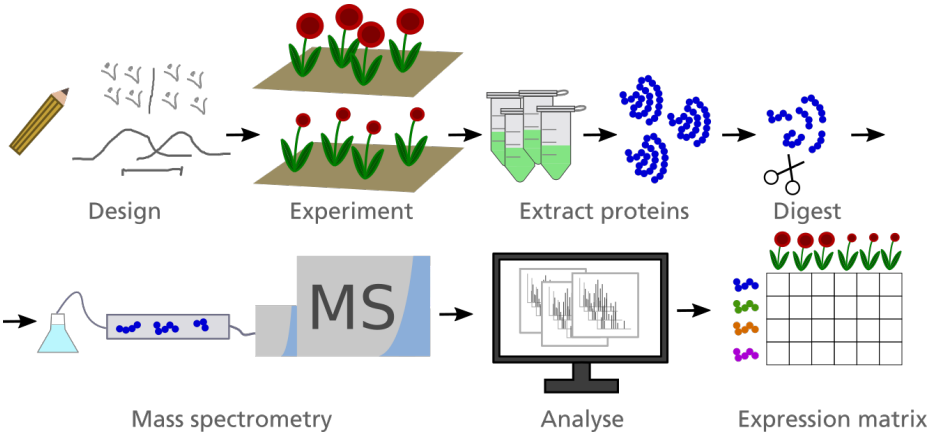


Figure 2: Schematic illustration of the steps from drawing board to measured protein abundances.

stage by planning the experiment such that if issues occur they can be corrected for during the data analysis, and during the experiments by maximizing the reproducibility of the laboratory procedure. Realistically, even with the best intentions unwanted variation will appear during experiments and needs to be carefully monitored and understood so that the computational biologist can consider it and draw reliable conclusions from the data despite its presence.

Navigating technical limitations in experiments has been an important aspect through all of the presented studies. The aim of **Papers I-II** is to make it easier to identify trends of interest in omics expression data and to provide tools to help draw optimal findings. These were subsequently used during the data analysis in **Papers III-V**, where different types of analysis decisions had to be made to draw reliable conclusions from the data, further discussed in Chapter 3. This first chapter aims to act as a stepping stone to understanding the issues that may appear during a proteomics study and the downstream challenges they can cause.

Designing an experiment

How an experiment is structured has far-reaching consequences to what conclusions can be drawn from the study. These consequences are primarily seen during the data analysis and biological interpretation towards the end of the project but require careful consideration already at the start. A poor experimental design will limit the potential of an experiment and can make it more difficult to adjust for laboratory work errors during the statistical analysis, thus wasting precious resources, time and research opportunities.

One of the main challenges of statistics in omics data is its multidimensionality, where potentially many thousands of variables (peptide abundances in the case of mass spectrometry) are measured simultaneously. Experimental design in omics has been extensively discussed for high-throughput experiments such as for microarray studies - one of the first established techniques for comprehensive profiling of gene expressions (Yang and Speed 2002; Churchill 2002; Dobbin, Shih and Simon 2003; Simon, Radmacher and Dobbin 2002). Much of this also applies to the current mass-spectrometry based measurements of proteomics (Oberg and Vitek 2009; Hu et al. 2005a). The three key experimental design questions discussed here are the number of replicates, randomization of samples, and blocking of samples.

Replicates are repetitions of the experimental workflow and are used to quantify sources of variation present in the experiment and to increase the accuracy of its measurements (Blainey, Krzywinski and Altman 2014). There are two types of replicates - biological and technical (illustrated in Figure 3). Biological replicates run the full biological experiment for additional cells, tissues or organisms. There is always a biological variation present between individuals, and biological replicates are needed to see beyond this. Technical replicates use the same biological material and runs of all or parts of the subsequent laboratory steps, for instance, by running the same biological sample twice on a mass spectrometer. A higher number of replicates gives a more reliable estimate of the variability, increasing the power of subsequent statistical tests, but require more resources. A study using RNA-seq in yeast showed that three biological replicates, a typical number in expression-based studies, only detected 20%-40% of the regulated genes compared to what was identified when using a high number of replicates (Schurch et al. 2016). The expected depth of a study per number of replicates can be calculated beforehand by considering the heterogeneity of the sample, allowing one to make trade-offs between resources and depth during the design stage. Finally, technical replicates are useful to quantify and to reduce the impact from the technical variation of an experiment. Both biological and technical replicates are valuable tools for understanding the variance in the experiment and increasing the sensitivity of the subsequent statistical tests.

Randomization and blocking are strategies for organizing the processing of samples in order to minimize the risk of technical variation, which disrupts the later statistical analyses (Suresh 2011). Here, samples from different biological conditions of interest are balanced across possibly disturbing factors, such as run day or reagent batch (illustrated in Figure 4). This balancing allows the statistical test to consider the technical condition as a disturbance by including the it as a so-called *covariate*. Including a condition as a covariate gives the statistical test the ability to independently model variation from that condition, and can thus separate it from the condition of interest. In the worst case, a technical effect is completely overlapping with the studied effect, making them inseparable, a concept called *confounding* (top row in Figure 4). In a randomized experiment, the order of the samples

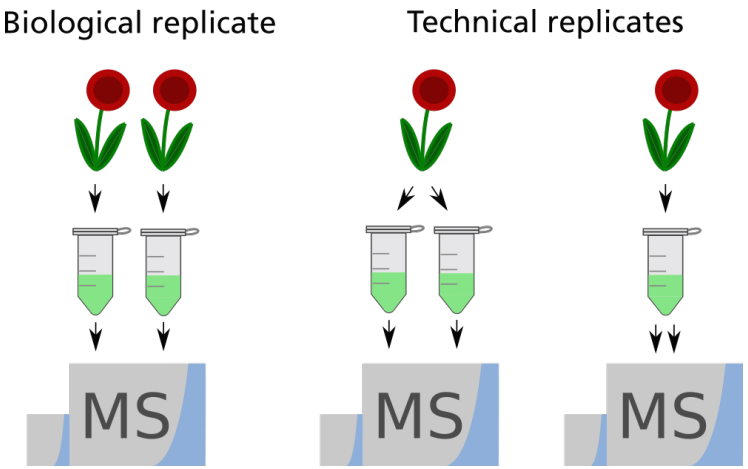


Figure 3: The two types or replicates: Biological use different organisms, cells or tissues for each sample, while technical replicates rerun parts of the experimental workflow for a sample taken from the same individual.

is shuffled to reduce the risk of confounding. Here, there is still a risk that conditions purely by chance are distributed unevenly across the technical conditions, interfering with the statistical analysis (middle row in Figure 4). Blocking extends randomization by evenly distributing biological conditions across groups of samples known to later cause variation, ensuring that the condition is evenly balanced (lower row in Figure 4) (Burger, Vaudel and Barsnes 2020).

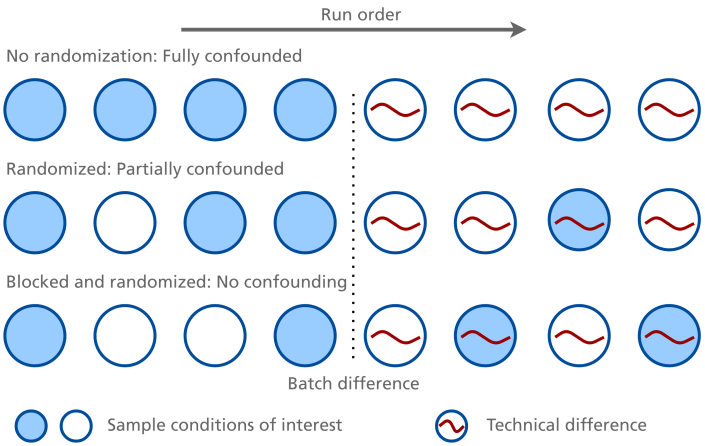


Figure 4: Types of randomization, and how it can lead to overlap (confounding) between conditions of interest and unwanted conditions.

The structure of the experiment defines its potential and its robustness to technical issues.

A good design gives the resources put into the project their best shot of coming to good use and allows accounting for expected and unexpected unwanted variation appearing during the statistical analysis. Poor design may severely limit the value of an experiment or even make it impossible to draw conclusions from it.

Sample handling for proteomics

The sample handling process starts with extracting proteins from the biological samples and ends with inserting the processed sample into the mass spectrometer. The sample handling steps have been found to be the most susceptible to technical variation in the proteomic workflow (Piehowski et al. 2013). Some aspects that can cause systematic bias are variation in chemical reagents, instrument calibrations, differences in liquid chromatography columns, temperature changes or differences in human handling (Karpievitch, Dabney and Smith 2012). This variation can partially be adjusted computationally using algorithms such as normalizations and batch effect corrections, as carried out by software such as NormalyzerDE (**Paper I**). Still, they can never be fully adjusted for, and the exact impact on the subsequent analysis is often uncertain. Therefore, the experiments need to be carried out with the utmost care, potentially using sample handling robots to automate steps to reduce variation caused by human handling (Krüger, Lehmann and Rhode 2013), as well as having a good maintenance routine for the mass spectrometer. The main sample handling steps in bottom-up label-free proteomics (the type of proteomic approach used in the work presented in this thesis) are illustrated in Figure 5. Briefly, the proteins are extracted from the tissues or cells while also cleaning away substances such as salts and surfactants, unfolded by a process called denaturation, having their cysteines reduced to break their sulphide bonds, digested to peptides using a protease that cleaves the proteins adjacent to specific amino acids, and finally optionally cleaned again prior to injection into the mass spectrometer (Kulak et al. 2014).

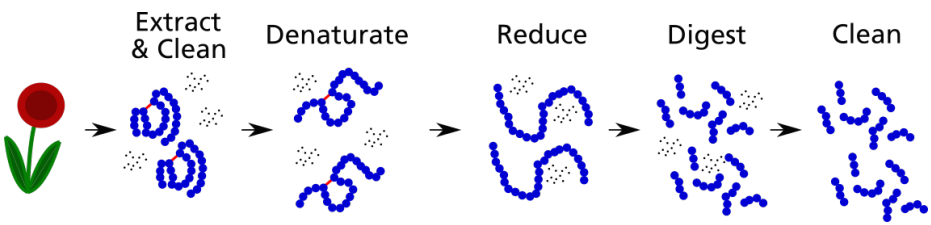


Figure 5: The main sample handling steps in bottom-up label free proteomics.

During the extraction steps, the proteins are retrieved from the cells or tissues of interest. Variations in the original material and how the extraction is performed have been shown to impact the protein yield and the structural integrity of the target proteins (Simpson 2003;

Piehowski et al. 2013). Different types of tissue require different considerations (Wang et al. 2018; Dittrich et al. 2015), further complicating the procedure. In bottom-up proteomics (the approach used in **Papers III-V** and outlined in Figure 5), proteins are cleaved into peptides at specific sites using a protease, commonly trypsin, before analysis in the mass spectrometer. This process results in a mixture of peptides masses mostly fitting into the detection range of the mass spectrometer. During digestion, cleavage points are sometimes missed, leading to a mix of fully and partially digested peptides. Undigested peptides have been shown to constitute around 20% of the resulting peptides (Burkhart et al. 2012; Picott, Aebersold and Domont 2007), thus causing considerable variation in the downstream analysis if the degree of missed cleavages is not constant within the analysed set of samples.

The studies presented in **Papers III-V** use a label-free approach. The alternative is to use labelled approaches where labels are inserted either chemically or metabolically into the proteins (Ong et al. 2002; Thompson et al. 2003; Gygi et al. 1999), allowing for mixing of multiple samples up to the maximum number allowed by the type of labelling used (commonly 10 or 16 samples per set for chemical labels). These labels are then used by the mass spectrometer to distinguish proteins coming from different samples. This approach can reduce the number of mass spectrometry runs and consequently the variation caused during the mass spectrometry processing, but risks causing batch effects when the number of labels are exceeded and additional samples need to be run with a separate set of labels. Furthermore, labelled proteins are often analysed across multiple mass spectrometry runs after fractionation, where proteins with different characteristics are separated to allow a deeper study of the proteome. Both labelled and label-free methods have strengths and weaknesses. In this work, we use the label-free approach due to its laboratory simplicity, in particular in light of running sets of samples exceeding the labelling sizes.

If the sample preparation is handled well, the chance of an accurate view of the underlying biology in the experiment is maximized. The automation of sample preparation has gradually started gaining more widespread use. Automation can reduce variability caused by human handling of samples while making it possible to process samples in parallel, increasing throughput (Fu et al. 2018; Krüger, Lehmann and Rhode 2013). Furthermore, thorough documentation of parameters such as reagents, temperatures, and personnel performing the experiments is critical as it makes it possible to assess the limitations of the data during the data analysis. The day where sample handling in proteomics is without challenges still seems far away. Thus, potential sources of variation need to be carefully managed, documented and considered during the data analysis steps.

Measuring peptides using bottom-up mass spectrometry

A mass spectrometer is a complex machine used to measure the mass-to-charge (m/z) ratio of molecules with high accuracy. In bottom-up proteomics, these measurements are performed on cleaved proteins (called peptides), and the measured intensities are used to calculate protein identities and abundances. Similarly to the process of sample handling, technical variation can be introduced during the steps performed using the mass spectrometer due to performance changes in its components or drift in its calibration. These changes should ideally be accounted for by careful handling of the experiments and maintenance of the equipment, but will, in practice, need to be evaluated and adjusted for during the computational processing using techniques such as normalization and batch effect correction. To further add to the complexity, large scale experiments can involve dozens or hundreds of samples being run sequentially over days. If any parameter in the instrument changes during this time, it will lead to technical variation. An overview of the mass spectrometry workflow is illustrated in Figure 6. In this section, I will discuss common sources of variation and their impact on the subsequent analyses.

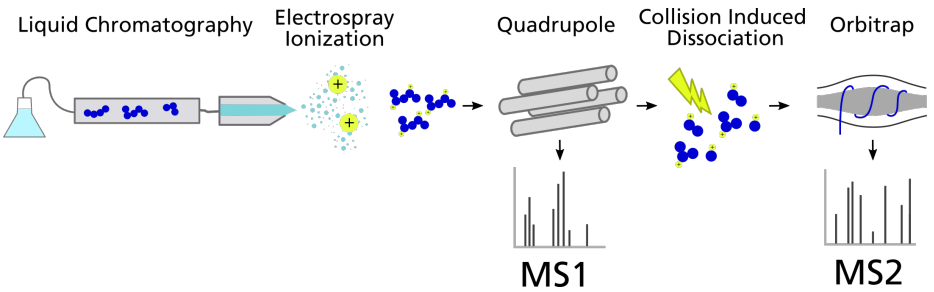


Figure 6: Schematic illustration of the main steps in the mass spectrometry workflow.

In the studies presented in this thesis, the mass spectrometry has been preceded by liquid chromatography (LC) separation. In this technique, peptides are sent under pressure through a chromatographic column packed with a material, commonly C18, able to interact with peptides based on their chemical characteristics. Thus, peptides are separated, traveling with different speed towards the ion source. The time it takes for a peptide to pass by the column and reach the instrument is called retention time (RT). This separation gives the mass spectrometer more time to measure the incoming peptides, providing a deeper view of the proteome. In a typical experiment with a 40-90 minutes gradient, individual peptides will spread out to mostly less than minute-long distributions, meaning that many peptides are continuously being measured by the mass spectrometer for this duration. The chromatographic column is a common source of variation, making it difficult to directly compare samples ran at different times or in other mass spectrometers. This phenomena was seen in **Paper V** where the column was replaced midway through the sample processing,

causing considerable variation to the dataset and prompting reruns of samples.

Next, the peptides are passed from the narrow tip of the column into an electrospray (Fenn et al. 1989), where they are ionized by applying a high voltage and emitted as a rapidly evaporating mist of peptide droplets, sending charged peptides into the mass spectrometer. This ion intensity may fluctuate over time, which means that peptides measured at certain retention times in specific samples will have higher or lower ion intensities which consequently will influence the measured abundances. This abundancy variation is often unaccounted for in downstream normalization procedures, but attempts to correct for this have been made (Van Riper et al. 2014; Zhang, Käll and Zubarev 2016). **Paper I** introduces a new generalized approach to normalize time-dependent intensity fluctuations, compatible with a range of existing normalization techniques.

In the studies carried out here, the initial peptide selection in the mass spectrometer is performed using a quadrupole mass analyser consisting of four metal rods that produces an electric field, carefully controlling that only peptides' with a specific mass-to-charge ratio enter the mass spectrometer (Yost and Enke 1978). The selected peptides are fragmented in a collision cell where high energy particles under high pressure collide with the peptides. These fragmented ions are fed into another mass analyser measuring their mass over charge ratios. In the work presented in **Papers III-V**, the final mass analyser did in most cases consist of an orbitrap (Hu et al. 2005b), a mass analyser using an electric field to rapidly spin the peptides around an electrode and using the frequency of their movement across it to calculate their mass over charge ratios. These measurements of fragmented peptide ions give what is later referred to as the MS2-spectrum, a highly accurate fingerprint of the masses of the peptide fragments.

Two common modes of using the mass spectrometer are data-dependent acquisition (DDA) (the approach used in **Papers III-V**) and data-independent acquisition (DIA). In DDA, the peptides with the highest intensity entering the mass spectrometer are selected for further analysis. On the other hand, in data-independent acquisition (DIA) (Purvine et al. 2003), a newer technique rapidly gaining traction, the mass spectrometer performs fragmentation for predefined ranges of mass-to-charge values, stepwise going through the full m/z range. This selection produces an unbiased and comparably more complex spectrum as wider m/z ranges are used and selected regardless of which incoming peptides are present. Using DIA has been shown to reduce the challenges with missing values compared to the DDA approach while requiring more complex algorithms for processing the spectra. Software have lately been developed with this purpose, thus reducing the barrier of entry for analysing this type of data (Gillet et al. 2012; Röst et al. 2014; Tsou et al. 2015; Teleman et al. 2015; Searle et al. 2018). In this work, DDA was used due to its relative simplicity and its ability to identify and quantify thousands of proteins. However, this approach causes a selection bias as peptides with low abundance or low ionization ability may never get selected for identification, leading to missing values in the subsequent data analysis.

In conclusion, the mass spectrometer can give a comprehensive view of which proteins are present in a sample, but it requires a complex workflow that needs to be carefully tuned to ensure reliable results. Similarly to during the sample handling, variations during these steps will impact the subsequent data analysis and should be carefully documented such that they can be visualized, understood and accounted for statistically during the data analysis.

Computational processing of mass spectra to protein abundances

The computational processing of mass spectra starts with the data obtained from the mass spectrometer. Here, the aim is to use the measured masses of peptides and their fragments to build a comprehensive view of the proteins present in the original samples. In this step the challenge changes from avoiding causing technical variation to making optimal choices of software, algorithms, and parameter settings. Each will influence the results and potentially impact the final interpretations.

The choice of software has been shown to have a considerable impact on the analysis results (Bell et al. 2009; Chawade et al. 2015). In some studies, the skill and experience in using the tools even more so (Navarro et al. 2016; Choi et al. 2017), demonstrating the importance of understanding the mass spectrometry principles. Proteomics users have the choice between using a single piece of software to carry out all the analysis steps or using a modular workflow with different software for each step. Popular examples of singular software able to carry out the full proteomics workflow are MaxQuant (Tyanova, Temu and Cox 2016) and Progenesis (http://www.nonlinear.com/progenesis), which require comparably less technical knowledge, while in many cases still performing well (Välikangas, Suomi and Elo 2017). On the other hand, modular approaches such as OpenMS (Röst et al. 2016), Proteios (Häkkinen et al. 2009), DeMixQ (Zhang, Käll and Zubarev 2016) or custom workflows allow for selection of best-performing tools for each step and do, in many cases, allow for automation of the analysis, making the analysis and later reanalyses easier for technical users. Critical steps of the workflow are outlined in Figure 7.

The first computational step is to use the mass spectra to find abundances and identities of the measured peptides. In label-free proteomics, the MS1 spectra measuring peptides with different charge states over time are typically used to calculate peptide abundances (Teleman et al. 2016; Cox et al. 2014; Röst et al. 2016). As the ionization ability of the peptides varies with their sequence, it is difficult to make other comparisons than between the same peptide across samples. The differences in ionization properties also make it challenging to calculate absolute abundances of proteins using mass spectrometry.

The parallel step is to identify peptide sequences based on their MS2 measurements. The mass-to-charge ratios of the peptide fragments are used as fingerprints and are matched to

simulated fragments from databases with known protein sequences (Eng, McCormack and Yates III 1994). The peptide identification performance depends on both the algorithm, the search settings, and which database is used. If proteins are not included in the database, their peptides cannot be detected using this strategy. If using a large database, the statistical strategy commonly used to ensuring a low false positive rate (the identification of an incorrect peptide sequence) will lead to a high number of false negatives (failing to identify an existing peptide with enough confidence). Approaches to improve the false discovery rate have been proposed, such as combining multiple search engine results (Shteynberg et al. 2013) or using machine learning strategies to better separate real from false matches (Käll et al. 2007). Recently, new techniques using MS2 peak intensities in addition to the m/z values have emerged, with the potential of reducing limitations from using large databases (Barton and Whittaker 2009). If successful, this would reduce the burden of false negatives by increasing the accuracy of peptide spectrum matches, which would be particularly useful when working with large databases such as when looking for additional modifications of the peptides called post-translational modifications (PTMs), or in metaproteomics where many organisms are studied simultaneously.

Next analysis challenge is to reduce the problem of missing values. A common issue caused by the data-dependent acquisition strategy is values missing due to only measuring highly abundant peptide ions selected for MS2-fragmentation. Still, if present, the ions are observed on MS1-level and their identity can be shared across samples, partially remedying the issue. There are numerous algorithms for this purpose, as reviewed (Smith, Ventura and Prince 2013), which successfully reduce the number of missing values, but may suffer from false matches, particularly when the number of samples is large. The types of missing values also need to be distinguished, as values systematically missing in one biological condition may indicate biological effects rather than technical variation. Approaches to consider missing values and their relationship to potential biological effects are discussed further in Chapter 2 and **Paper II**, and are applied in **Paper III**.

The final step going from spectrum to protein is to infer protein identities and abundances from the peptides. Many approaches have been proposed for this purpose (Nesvizhskii and Aebersold 2005; Huang et al. 2012). The protein inference is challenging as one peptide frequently matches to many variants of the proteins, such as isoforms, close homologues or, when modified, different post-translational modifications. Each variant may have different functions and abundances. These are jointly called *proteoforms* (Smith and Kelleher 2013). In bottom-up proteomics we often miss these proteoform-specific differences as the measured peptides are present in multiple variants of the protein, giving us measurements of groups of proteoforms.

In the studies presented here, a modular workflow was used, starting with the Proteios software environment (Häkkinen et al. 2009) to carry out MS2 searches using two search engines. For MS2 searches, X!Tandem (Craig and Beavis 2004) was used together with

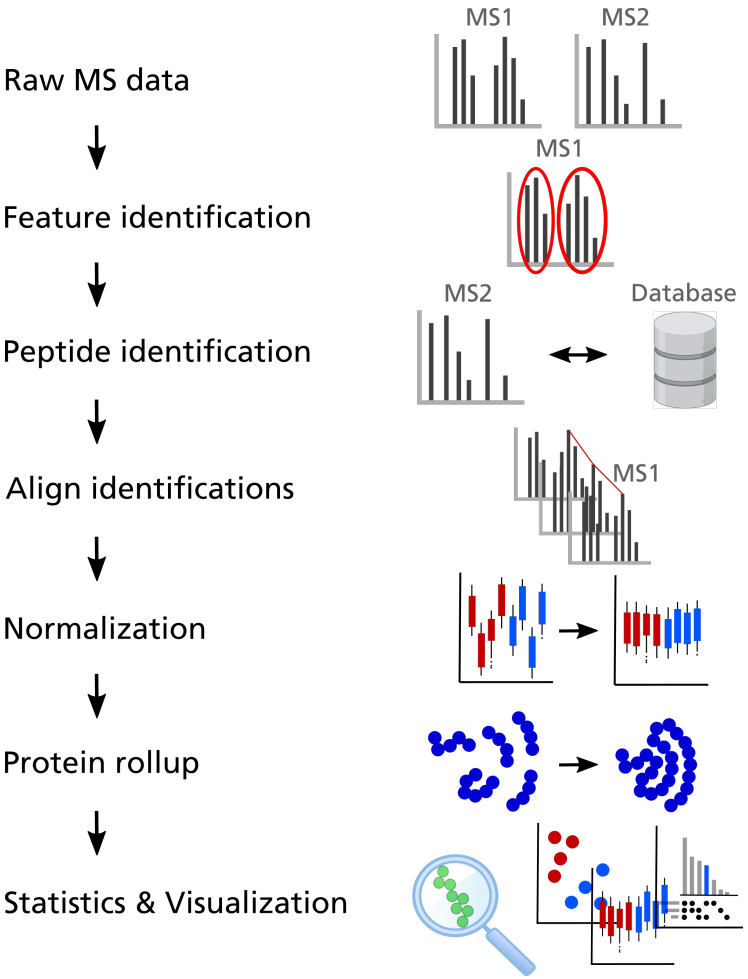


Figure 7: Schematic illustration of the main steps in the computational workflow used for bottom-up proteomics with the data-driven acquisition workflow.

either MS-GF+ (Kim and Pevzner 2014) or Mascot (Perkins et al. 1999). For feature detection, Dinosaur (Teleman et al. 2016) was used, an open-source extension of the MaxQuant algorithm for label-free quantification (Cox and Mann 2008). Approaches for alignment were explored during the projects (Scott 2019), with features in the present studies aligned and combined using an algorithm built into Proteios (Sandin et al. 2013). NormalyzerDE (**Paper I**) was used to identify a robustly performing normalization technique, here using the cyclic Loess normalization, found to consistently perform well in the datasets analysed in **Papers III-V**. No batch effect correction was performed at this stage, but later during the

statistical calculations by setting the condition as a covariate. The RRollup algorithm from DanteR (Polpitiya et al. 2008) was used for protein rollup using a Python (Møller 2017) or an R implementation (`github.com/ComputationalProteomics/ProteinRollup`), a strategy that selects a peptide with few missing values and uses it as a reference for scaling of the remaining peptides to a similar intensity level before calculating averages in each sample. No imputation was performed, keeping missing values as missing. Software choices were kept constant throughout the subsequent analyses of follow-up data in the studies presented in **Papers IV-V** to avoid introducing additional variation between the datasets due to different software choices.

The choices made during the computational processing of mass spectra into protein abundances significantly impact the resulting values and may influence the downstream interpretations of the data. A sufficient understanding of underlying principles has been shown important to obtain optimal results, both when using modular software and a single software solution. Still, many challenges in the computational processing of proteomics remain to be met and more are coming up in light of new methods developments. To meet these challenges, both new algorithms and robust and user-friendly software need to be developed.

Concluding thoughts

As discussed in this chapter, many sources of variation influence the proteomic data at each step, from the laboratory parts to choice of software and analysis methods. Some of these can be controlled by carefully designing and carrying out experiments, and using robust software to perform the analysis. Still, due to the complexity of the experiments, technical variation is still often inevitable. In the next chapter we will see how this can be accounted for using algorithms to reduce the unwanted variation in the data, and how informed data analysis choices based on visualizations help us bringing out the best of the data, even when limited by technical variation.

Chapter 2: From proteins to biological insight

Data analysis in biomarker discovery aims to identify persistent biological patterns that can be used to understand biology better and predict useful traits. This identification is challenging due to the complexity of the data. One challenge is the many sources of unwanted variation that may obscure the biological signal or even introduce signal which might be interpreted as biological. Beyond this, the inherently random nature of the data and the flexibility of the computational analysis pose other challenges, making it difficult to know what tools and statistical approaches are most appropriate for each task. Venet *et al.* explored how well random gene-expression signatures correlated to breast cancer outcomes and found that the published signatures, in most cases, did not perform significantly better than random signatures (Venet, Dumont and Detours 2011). The issue with the often limited reproducibility for published biomarker signatures have been discussed at multiple occasions (Chibon 2013; Bustin 2014; McShane 2017; Scherer 2017) and indicate that many published signatures are likely to be unreliable, spurious patterns seen only in one dataset. If so, this is consequential for how the data analysis should be approached, indicating that great care needs to be taken when interpreting this type of data. This chapter discusses how to use normalizations, batch effect correction and statistical approaches to increase the robustness of the analysis, and the use of data visualizations to guide the approach and the conclusions of the analysis. The overall analysis workflow as discussed here is illustrated in Figure 8. These steps and related challenges are discussed in this chapter. **Paper I** is closely

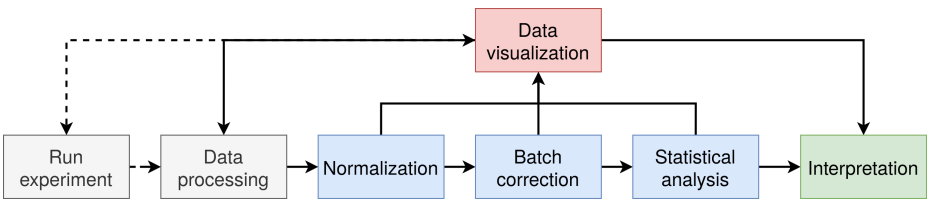


Figure 8: The data analysis workflow.

related to the normalization and statistical analysis steps, while **Paper II** almost exclusively focuses on the use of data visualization for better decisions on how to analyse the data. These pieces of software were subsequently used in **Papers III-V** to guide and carry out analysis decisions that are further discussed in Chapter 3.

Managing unwanted variation

Even in an experiment with no technical disturbances, systematic biological differences need to be distinguished from individual variation. In reality, there will generally be an unwanted variation present in the data. This variation can be caused by differing conditions in the experiment before sampling (for instance, if an older batch of seeds is used for some plants) and by technical differences from the sample handling itself (different reagents are used for protein extraction), as illustrated in Figure 9.

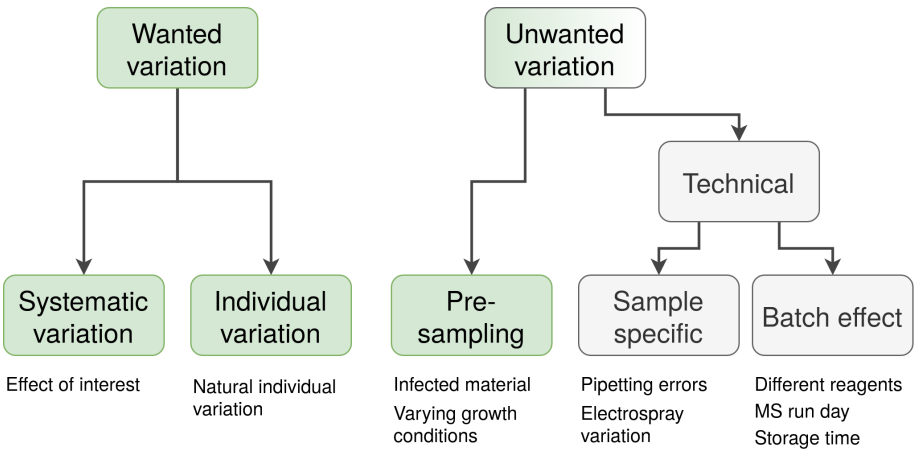


Figure 9: Breakdown of different types of variation. "Wanted variation" is what is present if only the variation intrinsic to the organism is measured with no additional variation caused by the experiment and the sample handling. "Unwanted variation" is any disturbance caused either during the experiment prior to the sampling or in the subsequent sample processing steps.

One example of pre-sampling variation was shown in a recent study where HeLa cell lines from different laboratories were compared, showing different gene expression profiles (Liu et al. 2019). This difference is likely due to gradual mutation over time, making them more diverse. This diversity means that when comparing results from studies based on the HeLa cells in different labs, this additional source of variation will be present and needs to be considered while interpreting the data. The technical difference is often further divided into sample-specific effects (such as pipetting errors or electrospray variation) and batch effects (such as run-days on the mass spectrometer and reagent batches). The difference

between random effects and batch effects is illustrated in Figure 10, showing how batch effects systematically either shift samples by a fixed effect or along a gradient. In contrast, the random effect is not linked to a specific set of samples.

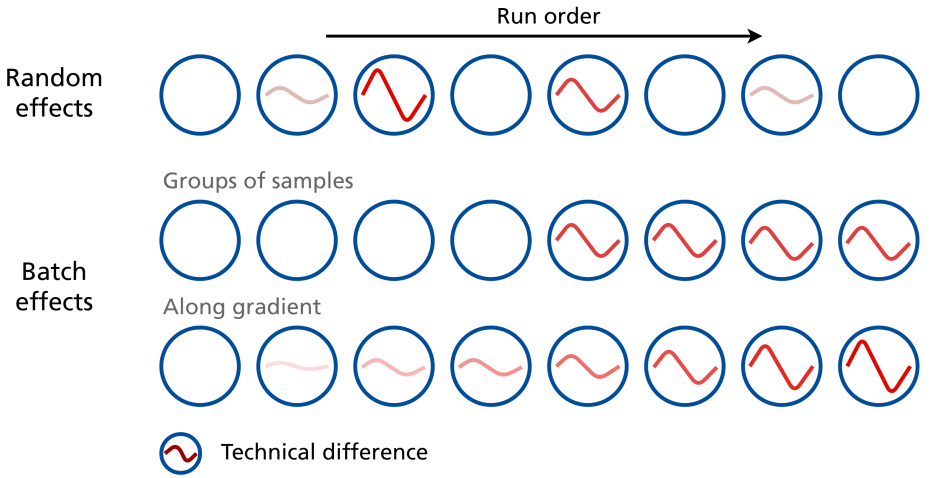


Figure 10: Illustration of random effects and batch effects. Random effects influence samples in a non-predictable way, while batch effects impact samples systematically, either as a group or along a gradient.

Strategies to correct for sample-specific technical variation are called *normalizations*. These strategies are the main focus of **Paper I**. Batch effects, the second type of technical variation, have been a central point throughout the omic studies presented in **Papers III-V**. Batch effects can sometimes be corrected for by using batch effect correction strategies. Both normalizations and batch correction methods need to be applied with care, as they will introduce new structures in the data and may risk removing biological variation while attempting to compensate for the technical variation. If the reduction of technical variation is greater than the disturbances introduced by the normalization, the normalization procedure can help give a clearer view of the variation of interest. Visualizations are also important for providing guidance on how to analyse the data to get the most reliable result. In this chapter, the two main types of technical variations and strategies to handle them during the data analysis are discussed.

Normalization

Normalizations aim to adjust for sample-specific technical differences to reduce technical variation in order to make the samples more comparable and get a clearer view of the biology. If applied correctly, this can increase the ability to draw conclusions from the data. Still, if applied in a way that breaks the assumptions of the normalization technique, this

1. Calculate the median for each sample
2. Calculate the average of these medians
3. Use this to calculate a scaling factor for each sample
4. Scale all the values within each sample with this factor

Figure 11: The procedure of median normalization.

can cause incorrect and misleading results by introducing false signals into the data. Thus normalization is an important step in omics analysis but needs to be applied with care. It can be performed at many stages of the processing of the proteomics samples. Here I focus on normalization techniques carried out as a post-processing of the peptide abundance matrix (Rourke et al. 2019).

To explain how normalization works, I will start by demonstrating a commonly used normalization technique called *median normalization*, available in **Paper I** and outlined in Figure 11. Here the assumption is that technical differences will equally shift all values within each sample, for instance, if pipetting a higher concentration in one sample leading to overall higher measured protein abundances in that sample. Median normalization also assumes that the median protein abundances are similar in the original cells or tissues. Thus, the normalization procedure evenly scales peptide abundances within the samples so that the median peptide intensity of all samples is the same. This procedure applied to four simulated proteins in four samples is illustrated in Figure 12, where sample s2 is systematically shifted towards higher abundances and protein P4 is differentially expressed in the underlying simulation. We can see that P4 will not be identified as differentially expressed without normalization, but after normalization, it will. For the median normalization to work well, its assumptions need to be met. For instance, if three of the proteins were differentially expressed, this would break the normalization assumption that most proteins are kept constant, as illustrated in Figure 13, causing the protein originally present in similar abundances across samples to appear downregulated. Another assumption of median normalization is that the technical disturbances at low intensities are shifted as much as those of high intensity. This is sometimes not the case, breaking the assumptions of the median normalization, while other more flexible normalizations allow for this.

Many normalization approaches have been proposed for use in label-free proteomics, each with different assumptions and limitations. Often techniques developed for microarray are directly applied to proteomics. Examples of this include: quantile normalization (Bolstad et al. 2003), which adjusts all samples to have the same overall distribution of values, with recent variations allowing different distributions within different provided groups of

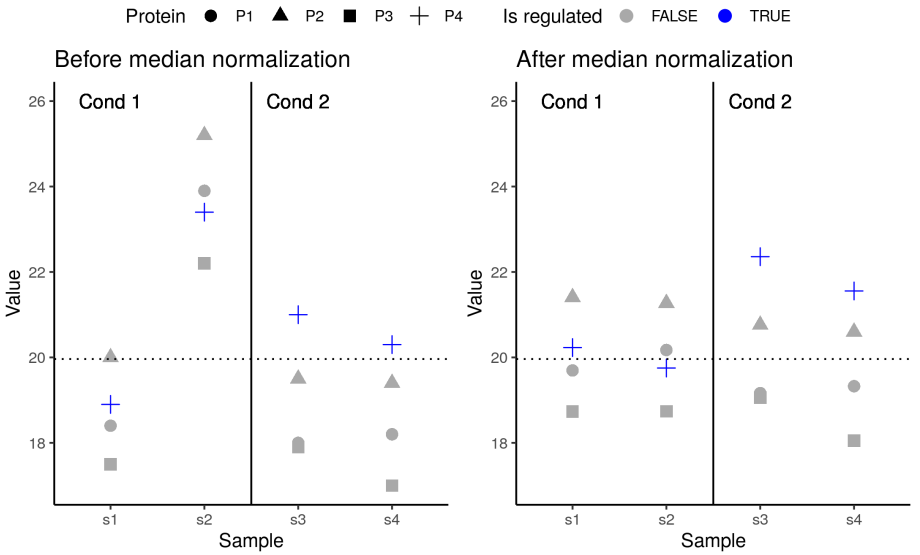


Figure 12: Illustration of median normalization. One protein (blue) is present in different abundance between the two conditions. One sample (s2) is systematically shifted compared to the rest, shifting all four proteins. After normalization the trend for P4 becomes visible. The average median is marked with a horizontal dotted line.

samples (Hicks et al. 2018); Cyclic Loess (Ballman et al. 2004), which attempts to compensate for shifts in intensity at different overall intensity levels; VSN normalization (Huber et al. 2002), which tries to compensate for any relationship between the variance and the mean. A different approach, EigenMS, looks for eigenvectors in the data and transforms the datasets based on these to remove unwanted variation (Karpievitch et al. 2009; Karpievitch et al. 2014). NormFinder identifies sets of stable features across samples, which subsequently are used to rescale the data (Andersen et al. 2004). Further, group-wise normalizations can be made, conserving variation between biological replicates groups such as provided in some normalization software (Chawade, Alexandersson and Levander 2014; Hicks et al. 2018). Here the results need to be handled carefully to not introduce artificial signals in subsequent statistics, which is likely if comparisons are performed between the groups after the normalization step.

With this range of normalizations available, selecting the best performing method can be a challenging task. Several studies have shown that which normalization method is used can have a considerable impact on the outcome (Webb-Robertson et al. 2011; Walach, Filzmoser and Hron 2018; Cook, Ma and Gamagedara 2020; Kultima et al. 2009; Callister et al. 2006; Välikangas, Suomi and Elo 2018; Yang et al. 2019). Among the normalization techniques, some methods including Cyclic Loess and VSN have shown a consistently high performance across multiple studies including **Paper I** (Välikangas, Suomi and Elo 2018;

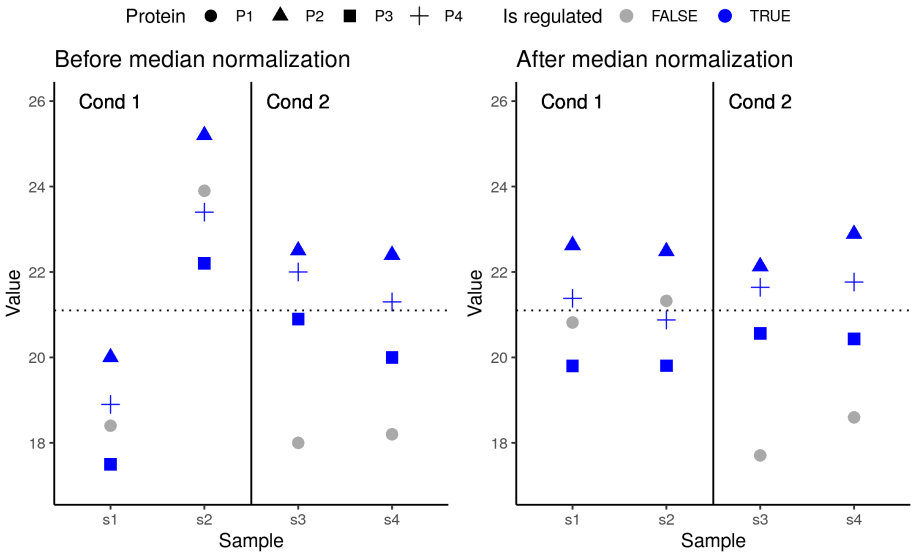


Figure 13: Illustration of median normalization when the majority of proteins are regulated. Here, the normalization artificially pushes the proteins present in different abundances (blue) to the same level, making the only protein present in the same abundance (grey) appear shifted downwards in the second condition. The average median is marked with a horizontal dotted line.

Walach, Filzmoser and Hron 2018). Still, these normalizations will not be well suited for all datasets, and careful evaluation of whether they perform well in the dataset at hand is needed.

Existing software for assessing the performance of normalization methods includes Normalyzer (Chawade, Alexandersson and Levander 2014) and NOREVA (Li et al. 2017; Yang et al. 2020), both providing normalizations and visual evaluation of performance measures. Ideally the software would automatically detect the best performing method. One example of software providing automatic method detection is quantro (Hicks and Irizarry 2015), but which provides a comparably less comprehensive assessment of the method performance. **Paper I** makes further improvements to Normalyzer and introduces the software NormalyzerDE, which extends the available normalization techniques with a retention time-based approach. The software is made accessible as a Bioconductor R package and as a web application where the user is given access to important input parameters. Furthermore, the software extends the analysis with an integrated statistical analysis step, which provides the ability to calculate statistical values and to generate statistical visualizations. **Paper I** thus provides a straight-forward and comprehensive tool for informed normalization selection and for performing the subsequent statistical analysis.

Currently, most established techniques, often developed for microarray data, do not use the

inherent structure of the proteomics when performing normalization. Some exceptions exist, but they have yet to obtain widespread use (Wang et al. 2006; Karpievitch et al. 2009; Van Riper et al. 2014). One type of bias unique to the mass spectrometer is intensity fluctuations caused during the peptide ionization in the electrospray (discussed in Chapter 1), which has been shown to vary in intensity on the scale of minutes (Lyutvinskiy et al. 2013). Methods to attempt countering this bias have been proposed, including the normalization method PIN (Van Riper et al. 2014) and a method integrated into DeMixQ (Zhang, Käll and Zubarev 2016). **Paper I** introduces a new generalized approach (illustrated in Figure 14 where it is applied to a dataset with artificial time-dependent biases present in one sample), applicable to use in conjunction with any normalization technique relying directly on the measured values and applied to mass spectrometry-based data with a time-based bias. The algorithm slices up the data across retention time (or any given analyte-specific numeric value) and applies the selected normalization technique on this subset before piecing the subsets together again. The subsets can be overlapping, allowing data points to be part of multiple normalization windows to reduce variability. In **Paper I**, it outperformed other normalization techniques, particularly in combination with Cyclic Loess normalization. Further validations could verify its performance and identify for which types of datasets its use would be particularly beneficial.

In conclusion, normalization is a critical step in the proteomics data analysis workflow. **Paper I** helps making an informed selection of a well-performing normalization technique. Furthermore, most established normalization methods do not use the unique structures of the proteomics data. **Paper I** proposes a new generalized approach to apply existing normalization methods to subsets of the data along with a moving retention time window, aiming to reduce the impact from retention-time dependent biases such as the electrospray intensity variation.

Batch effects

Batch effects are caused by systematic differences in experimental conditions influencing groups of samples. They have repeatedly been shown to have a substantial impact on omics studies, often overshadowing biological effects (Hu et al. 2005a; Gilad and Mizrahi-man 2015; Leek et al. 2010; Ransohoff 2005) and negatively influencing the ability to use the data in machine learning applications (Hilary and Jeffrey 2012; Leek and Storey 2007; Goh, Wang and Wong 2017). Ideally, batch effects should be considered both before and after the experiments are carried out. They can be considered during the experimental design with strategies such as randomization, blocking (discussed in Chapter 1), or control samples - samples with known contents later used as a reference (Cuklina, Pedrioli and Aebersold 2020). During the data analysis, the batch effects can be studied by visualization and sometimes adjusted for (Mertens 2017) using different correction strategies. The effectiveness of

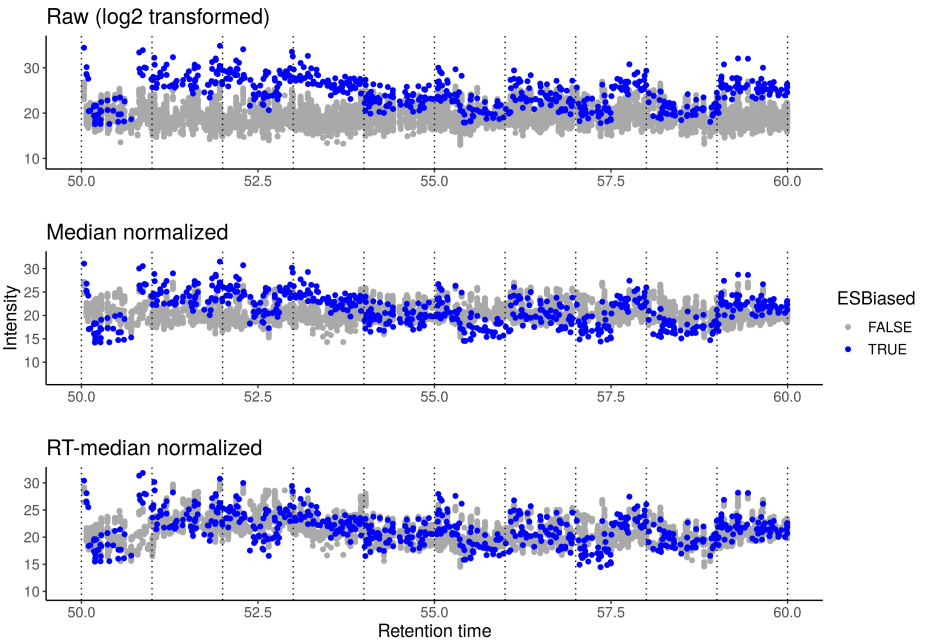


Figure 14: Illustration of retention time-based normalization approach, showing observed peptide intensities over retention time. A time-dependent bias was added to one of the samples (blue), emulating the electrospray bias. Median normalization (middle row) cannot fully compensate for this bias as it adjusts the intensity values globally. RT-median normalization (lower row) applies median normalization for time window-segmented data (dotted lines), and can better account for this type of bias.

these correction strategies have been debated and depends on the design of the experiment (Nygaard, Rødland and Hovig 2016). Still, batch effects are often unavoidable despite good experimental design and experimental procedures, such as when the samples are acquired over multiple days with potential instrument drift, or when processed in multiple laboratories (Irizarry et al. 2005). In mass spectrometry-based workflows, this is further inflated by the current trend of a growing number of samples used in studies (Cuklina, Pedrioli and Aebersold 2020). Here, I discuss strategies to understand and correct for batch effects during the data-processing stage.

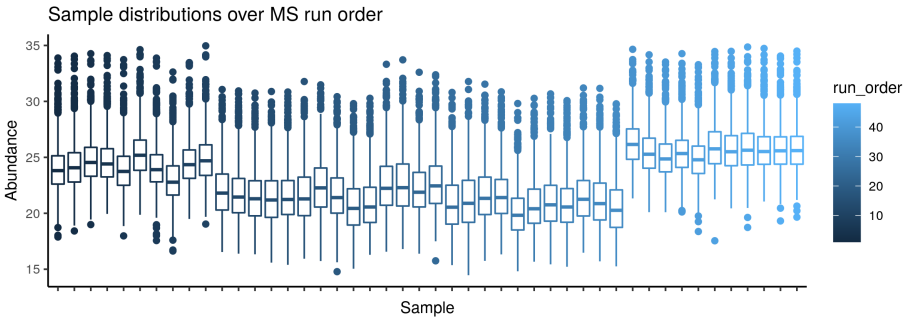
The limit for how well a batch effect can be managed during the data analysis steps is defined by the design of the experiment (discussed in Chapter 1 and illustrated in Figure 4). If a batch effect is evenly distributed across the biological groups of interest, it can be corrected for such that the sensitivity of subsequent statistical steps is improved (Gregori et al. 2012). Still, the additional technical variation cannot be entirely removed and will lead to lower sensitivity than experiments without batch effects. There is also a risk for overcorrecting a batch effect, introducing additional bias in the data. The risk for a biased correction has

been shown to be particularly high if batch effect correction is performed with imbalanced data (when the sample conditions of interest are not evenly distributed across the batches), where it can induce false positives in the subsequent statistical analysis (Nygaard, Rødland and Hovig 2016). If the batch effect is confounded, meaning that biological conditions overlap with the technical, it becomes challenging to distinguish the types of variation. For confounded experiments, results should be regarded with suspicion, although some approaches to batch effect correction have been shown to improve outcomes also in these cases (Luo et al. 2010) in the context of validation studies.

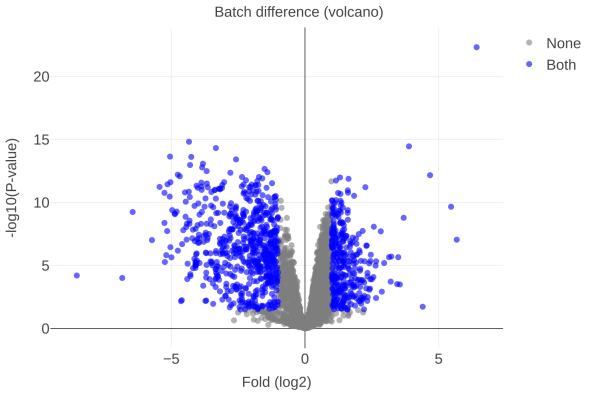
There exist various approaches to batch effect correction. Batch effects can either be corrected as part of the statistical calculations by including a known batch effect as a covariate (discussed in Chapter 1), meaning that the statistical test can attempt to model and ignore variance from that condition. This approach is available for statistical comparisons in Limma and is provided in **Paper I**. Another popular method is SVA (surrogate variable analysis) (Leek and Storey 2007), which attempts to directly identify batch effects within the data and model them as so-called "surrogate variables". These surrogate variables are then incorporated as covariates in the statistical test. In these cases, no data transformations are made - the batch effect is modelled within the statistical approach. Other methods transform the data similarly to normalization procedures such as the empirical Bayes approach Combat (Johnson, Li and Rabinovic 2007; Zhang et al. 2018), which can adjust for differences in the mean or mean and variance between batches, and has been shown to perform well in several studies (Chen et al. 2011). Finally, RUV (Remove Unwanted Variation) is another approach using features (peptides in the case of mass spectrometry) believed not to be differing between samples as control and rescales the data based on these (Gagnon-Bartsch and Speed 2012).

The identification of batch effects is commonly made using visualization tools such as principal component analysis (PCA) plots or dendrograms. Furthermore, to identify effects related to the run order in the mass spectrometer, samples can be visualized along with their order using, for instance, boxplots or bar plots to illustrate the number of missing values or total intensity, as shown in Figure 15 where two intensity shifts are present, indicating that batch effects may be present. Further, an illustration of the number of MS1 and MS2 features identified in each sequential sample can reveal both outliers and drifts in performance over time. For understanding batch effects, BatchI can identify sets of samples along run order likely belonging to a batch (Papiez et al. 2019) while BatchQC (Manimaran et al. 2016) allows direct exploration of batch effect corrections and can run both ComBat and SVA within the application. Other visualizations useful for batch explorations are provided by **Papers I-II**, such as interactive density plots that can reveal different sample distributions in separate batches, and illustrations of how individual features are distributed differently with and without batch effect correction.

In the analysis performed in **Paper III-V**, different types of batch effects were present and



(a) Sample intensity values observed along run order on the mass spectrometer - two systematic shifts in intensity can be seen, indicating potential batch effects



(b) Volcano plot illustrating comparison across batch (FDR < 0.05, |log2 fold| > 1 shown in blue)

Figure 15: Illustrations of known batch effects using OmicLoupe.

required careful consideration. Here, either the batch condition was included as a covariate in the statistical test or the statistical contrasts were organized such that they did not cross a known batch effect. These approaches are further discussed in Chapter 3.

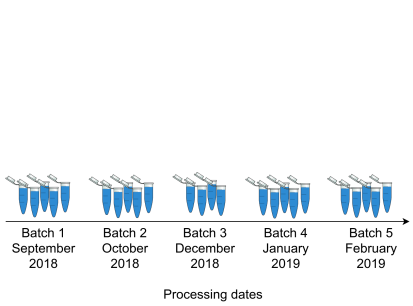
Using peptide characteristics to improve understanding of batch effects in mass spectrometry

By using structures uniquely present in the mass spectrometry-based data, potentially existing methods often designed for microarray platforms could be further improved. One example of this is including run order effects into the batch effect correction algorithms (Wang, Kuo and Tseng 2013; Kuligowski et al. 2014). During the work with batch effects

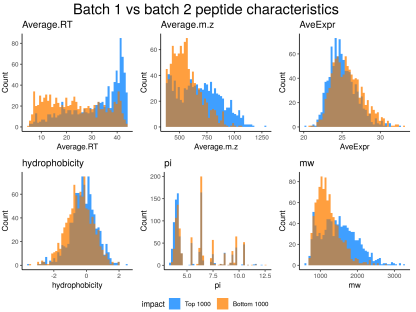
in this thesis, I hypothesized that how a type of batch effect impacts individual peptides varies with the physicochemical characteristics of the peptides. Here, some peptides may be more stable under the influence of batch effects and thus more reliable and more suitable as biomarkers. Similar attempts to understand how individual features are influenced by batch effects have been explored for probe sequences in microarrays. In these two approaches, relationships between probe sequences and batch effects were found within individual datasets, but they did not easily generalize between datasets (Hilary and Jeffrey 2012; Scherer 2009). Compared to microarray probe sequences, characteristics of peptides are more diverse, and thus the trends might be more distinct. This topic is explored in a poster (Willforss and Levander 2019) and with the help of a project student (Chi 2020). In this section, key ideas from this work are discussed.

A dataset with known batch effects was obtained by collecting sets of HeLa samples routinely used for quality control on the local mass spectrometer (of the model QExactive HF-X). Highly stable and susceptible peptides were selected to either use statistical differences (p-values and log2 fold changes - the difference between the peptide means of the compared groups) between batches or by their respective loadings for principal components found representative of the batch difference. In this second approach, a batch effect is first identified using principal component analysis, and the major component along which it is oriented is selected. Each protein contributes to this component to a different degree - this is the loading used to determine how closely linked the feature was to the batch effect. These methods provided groups of peptides classified as ”sensitive” and ”stable” that were subsequently investigated (illustrated in Figure 16). In both feature selection strategies, missing values were an issue, particularly when using principal component analysis, which only uses features with no missing values. The peptide characteristics were used as inputs to machine learning algorithms to see which were more important in separating the groups in different algorithms and to see whether the findings would generalize to other datasets. Differences in peptide characteristics and ROC-plots are shown in Figure 16) illustrating how the classifiers can separate the stable and susceptible features using cross-validation varying with batch, both without and with prior batch correction using ComBat (Johnson, Li and Rabinovic 2007). These algorithms showed an ability to separate peptides related to the size of the batch effect but did not easily generalize to other sets of samples with different batch effects. Still, it indicates that additional feature specific information could be used to improve existing methods, though considerations would need to be taken to the type of samples and type of batch effect.

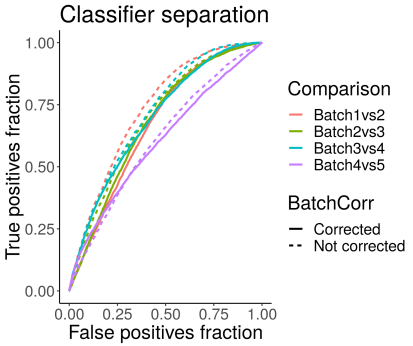
In the future, batch effect corrections will likely play a critical role in many areas, including multi-omics, where multiple types of data are integrated, and biomarker studies, where often the validation dataset will consist of a different set of samples. Existing batch effect correction algorithms could be improved by incorporating mass spectrometry specific characteristics and peptide characteristics as explored in this work. Still, as discussed by Goh *et*



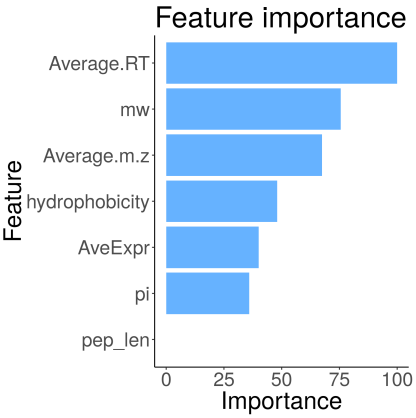
(a) Experimental setup



(b) Key differing characteristics between Batch 1 and Batch 2



(c) Classifier separation without and with prior ComBat batch correction



(d) Parameter importance as measured by the classifier when comparing Batch 1 and Batch 2

Figure 16: Highlights from the batch inspection study, identifying key peptide characteristics distinguishing the conditions using a random forest machine learning model (adapted from Willforss and Levander 2019).

al. (Goh, Wang and Wong 2017), the most critical part might be to increase the robustness of the experiments themselves to reduce the risk of batch effects occurring in the first place. In conclusion, batch effects is a persistent problem that needs to be approached from both the experimental design, robust execution of the experiments, and careful consideration during the data analysis. If handled well, their negative impact can be minimized, giving a better view of the biological variation of interest and increasing the chances of reliable findings.

Statistics in omics

To find gene products that differ between conditions of interest using statistical approaches is a common end-step for the data analysis. Even when not considering the sources of technical variation, this is a challenging task for several reasons, such as the many features (the studied molecule - proteins, transcripts, metabolites) tested, the random nature of the statistical testing itself, and the flexibility in selecting methods for pre-processing and statistical testing. To make reliable analysis decisions, the analyst needs to be clear on whether the analysis is explorative. The analyst also needs to determine whether false positives (you are tested sick, but you are healthy) or false negatives (you are tested healthy, but you are sick) are more severe. This section will discuss some key considerations when using statistical tools to find features of interest in omics datasets.

Statistical analysis often relies on the p-value, a measure of how frequently an outcome would occur by chance. This is subsequently often used as evidence of differences between groups of measurements. For instance, a p-value of 0.05 indicates that an as clear or clearer difference between measurements than what is observed would occur in 5% of measurements if there is no difference between the compared groups from which the measurements are taken (Altman and Krzywinski 2016). This becomes problematic in omics analyses where hundreds or thousands of statistical tests can be performed, leading to large numbers of false findings (features incorrectly thought to be different) if not corrected for. To illustrate the result of this, I simulated a dataset with 1000 features with no systematic difference ("negatives") and 50 features with systematic differences ("positives") in three replicates in two groups. I subsequently used a t-test to calculate p-values between the two groups and did false discovery rate (FDR) adjustments using the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995), which is widely used in omics-analysis, and the stricter Bonferroni procedure (Bonferroni 1936). Each of these corrects the data such that the number of false positives is limited to a fraction of the number of identified positives (true and false positives), meaning that if 50 features are found with FDR below 0.05, only 5% of these would on average be expected to be false findings. This stands in contrast to using regular p-values where the number of false findings is dependent on the size of the dataset, rather than the number of positives. The resulting distributions of statistical outcomes is illustrated using p-value histograms in Figure 17. In the top panel, a p-value below 0.05 was used as filtering criteria. Here, most of the regulated features were identified (94%), but almost half of the features found to be significant were false positives (random differences). On the other hand, after performing multiple hypothesis corrections (Benjamini-Hochberg or Bonferroni), the fraction of detected regulated features fell to 10% and 6%, while the precision rose to 83% and 100%. The exact outcome here depends on several parameters, such as the number of actually regulated features, replicates and effect size. If you want to further see how different parameters would impact the data, it can be interactively explored at `https://www.jakobwillforss.com/post/`

`interactive-exploration-of-pvalues-and-fdr`. This exploration demonstrates that how to adjust for multiple hypotheses depends strongly on whether false negatives or false positives are more problematic. In some cases, it has been argued that multiple hypothesis testing causes more harm than benefit due to the drastic reduction in power, which also contributes to a lack of reproducibility (Wang, Sue and Goh 2017). On the other hand, if not doing this correction, there is no control of whether identified features are purely significant by chance. Thus, whether this correction should be performed comes down to how concerning false positives are compared to false negatives. Still, the main decider of statistical power is the experimental design (as discussed in Chapter 1) and the number of replicates. If a deeper view of the measured omics is desired, designing the experiment for higher power can increase sensitivity without an increased number of false positives.

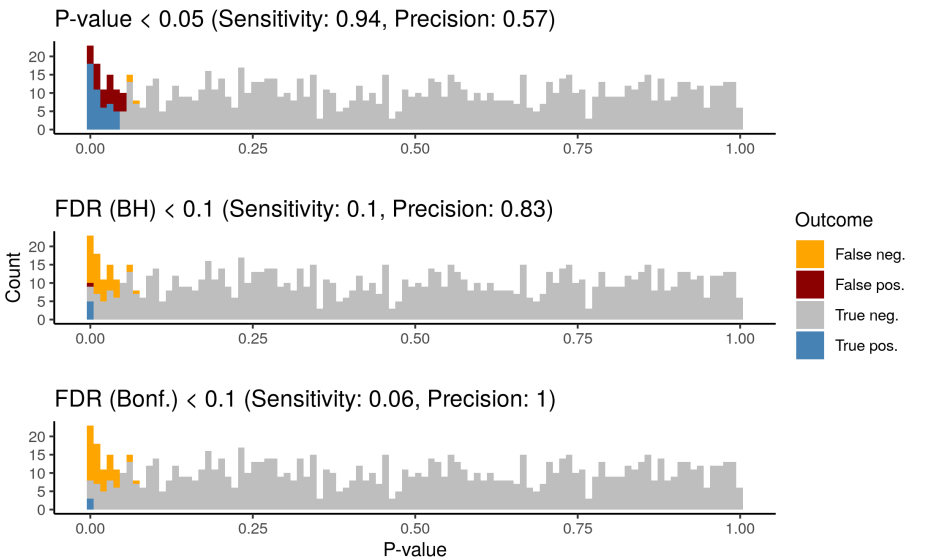


Figure 17: Impact of using p-values or adjusted p-values for feature selection. A thousand features were sampled with no systematic differences of the underlying distributions and 50 features sampled from distributions with a known difference. (BH: Benjamini-Hochberg, Bonf: Bonferroni)

A second related issue that often appears in omics-experiments is mixing explorative analyses and reporting of p-values (colloquially called "p-hacking"). Trying out different preprocessing and statistical approaches can effectively skew the obtained p-values by acting as a type of multiple hypothesis testing where many approaches are performed while only reporting the most interesting findings across all of them. This type of explorative statistics risks leading to results that are more compelling, but less robust and less likely to still be there in separate datasets. Still, again, whether this is problematic depends on if the goal is explorative to find potential trends to later be tested or if it is to report the features with a degree of certainty (Reinhart 2015). In **Paper I** a design decision was consciously made

to separate the selection of normalization procedure from the statistical testing so that the choice of method is based on performance measures for the data, not on the statistical findings. In conclusion, exploratively analysing omics data can be useful in finding interesting trends. Still, in doing so, the person analysing the data needs to carefully consider the purpose of the analysis and how the results are reported.

There are many approaches to identifying gene products that differ between conditions in omics. These include frequentist statistical methods (such as the t-test and ANOVA), Bayesian methods, and machine learning approaches (Tang et al. 2020). They can focus on detecting single features (such as single proteins or transcripts differing between groups) or combinations of features (such as gene ontology enrichments where groups of related genes are linked to biological differences). Here I focus on the detection of single features differing between groups. In frequentist statistical methods, the variation within groups of interests is compared to the variation between the groups, determining if any observed difference is unlikely to happen by chance. Here, calculations are done feature-by-feature, with each feature studied independently from the rest of the data. On the contrary, empirical Bayes methods such as Limma (Ritchie et al. 2015) or DEqMS (Zhu et al. 2020) make a preliminary estimate (so-called prior) of the feature variation using the full dataset and adjust it based on what is observed in each specific feature. When more data is present in one feature, more emphasis is put on this observed information rather than the prior information (gathered from the whole dataset). Limma has shown strong performance in several studies (Kammers et al. 2015; Ooijen et al. 2017) including **Paper I**, and is incorporated as the standard statistical approach in **Paper I**. It is used for statistical calculations in **Papers III-V**. ROTS (Reproducibility Optimized Test Statistic) (Elo et al. 2008) is another statistical method that successfully uses the multidimensional structure of the omics data. It uses a t-test for which the parameters are optimized using a resampling strategy called bootstrapping, adapting the t-test settings to produce a maximally robust ordering of features in the dataset at hand. This strategy has shown a superior or similar performance to empirical Bayes methods for proteomics in several recent studies (Pursiheimo et al. 2015; Suomi et al. 2017; Zhu et al. 2020). In conclusion, statistical methods utilizing the multidimensional structure of the omics data have been shown on several occasions to outperform classical methods such as the regular t-test. For these methods, future research will likely demonstrate their relative performance with greater certainty.

During the statistical testing, the p-value only indicates the presence of *any* difference, not whether the size of the difference is meaningful. Using the effect size of the comparisons (here called fold changes) can provide additional information, such as which features differ in a biologically meaningful way. They can also be used to reveal patterns of similarity when similar tests are performed in multiple datasets, such as in multi-omics or follow-up validation studies. For instance, if features that pass a loose p-value threshold in two separate datasets have a similar fold-trend overall, it is an indication that they are similar,

despite not being significant. **Paper II** introduces new visualizations using fold-changes to compare similarities between comparisons, which in turn were used to understand how similar repeated studies of the same biological system were in **Papers IV-V**.

A common issue in proteomics is the presence of missing values. Values can be missing for different reasons, such as the peptide ions being present in too low abundance to be observed in certain samples, or peptides being missed randomly in a sample. Each of these cases requires different optimal strategies (Lazar et al. 2016). A common strategy to handle missing values is imputation, where artificial values are inserted based on criteria estimating likely values, in many cases originally developed for microarrays and repurposed for proteomics (Lazar et al. 2016; Troyanskaya et al. 2001). An alternative strategy is to keep the values as missing through the analysis. Keeping values as missing limits the performance of the statistical tests and may lead to artificially high averages of measured values (Karpievitch, Dabney and Smith 2012). In contrast, the imputed values are generally used as true values in the subsequent statistics, which may introduce biases. The optimal imputation strategy has been found to vary between datasets and even for different features within datasets (Webb-Robertson et al. 2015; Goeminne et al. 2015). Strategies considering missing values specifically in proteomics have been presented (Webb-Robertson et al. 2010; Schwämmle et al. 2020). One recent approach uses Bayesian models to model the uncertainties of the imputed values and incorporate this in the subsequent statistical test (The and Käll 2019), allowing them to be considered as less reliable measurements. Missing values can be interesting by themselves, such as when a protein is missing in samples belonging to a certain biological condition. Here, the missing values might be due to a protein simply not being expressed at a level below the detection limit in one of the biological conditions. **Paper II** introduces a straight-forward approach to considering missing values by visualization, using an UpSet plot (Conway, Lex and Gehlenborg 2017). This provides an overview of in which conditions features are present or missing, and allows further inspection of these subsets.

When selecting features using statistical methods, it is essential to know the purpose of the analysis and how it impacts the reporting of statistical measures such as p-values. If not done carefully, this feature selection risks finding patterns that look compelling but are unlikely to generalize to other datasets. Multiple hypothesis correction tools are widely used and important to keep down the number of false positives but lead to such reduction in power that they might not always be beneficial. Statistical approaches considering the multidimensional nature of the omics data (such as Limma, implemented as the default statistical method in **Paper I**) can improve the sensitivity compared to standard methods such as the t-test. Approaches to handling missing values can provide ways to glean useful information from where they are missing, such as the UpSet-based visualization provided in **Paper II** and as done in the analysis in **Paper III**. Finally, similarly to when handling technical variation, the boundary of what can be achieved statistically in an experiment is

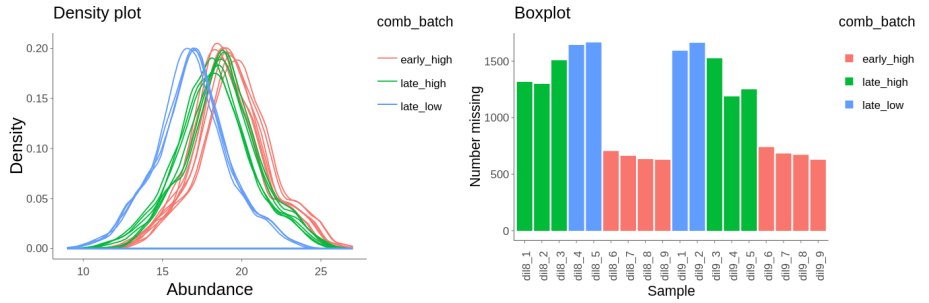
decided by the experimental design - the number of replicates and how they are organized, which needs to be carefully considered before starting the experiments.

Data visualization and analysis decisions

Data visualizations allow the human eye to look for patterns in the data to unlock insight not evident when inspecting it directly in a table. It can guide decisions about how to computationally process samples and how to perform subsequent analysis, and are often critical in spotting potential artifacts or unexpected findings. Both **Paper I** and **Paper II** rely on visualizations to guide analysis decisions, and visualizations have been crucial when navigating the datasets presented in **Papers III-V**. This section aims to highlight categories of visualizations as present in **Paper II** and some of their novel approaches to visualize multiple statistical comparisons to help understand similarities across datasets and statistical comparisons.

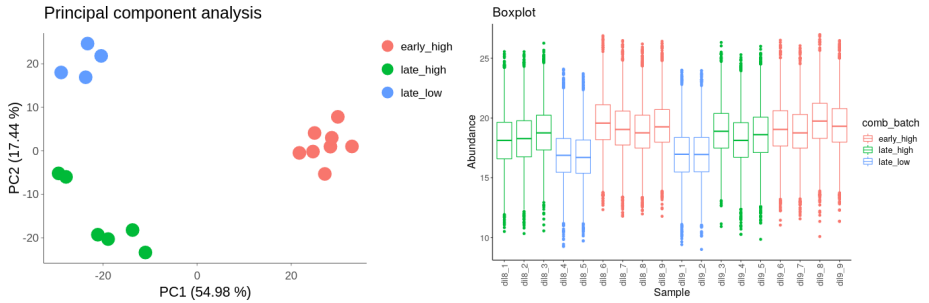
Sample-wide visualizations reveal trends on a dataset-wide level using each sample as a data point. These visualizations can guide analysis decisions such as removing outliers, changing upstream processing settings, and giving insight into how to best tackle technical variation using normalization and batch effect correction procedures, and choosing how to perform the subsequent statistical comparisons. Using sample-wide visualizations such as bar plots, box plots, or density plots (shown in Figure 18) quickly illustrates overall patterns such as the total intensity in each sample, the number of missing values, and shapes of distributions. These can often be used to rapidly spot abnormal samples with, for instance, many missing values or a different overall distribution ((a)-(b) in Figure 18). Other visualizations such as principal component analysis (Jolliffe 2002) ((c) in Figure 18) and dendrograms ((d) in Figure 18) project the multidimensional data into fewer dimensions. This projection allows the identification of patterns showing how samples are similar or different and studying whether technical or biological factors seem to carry the most impact on the overall view of the samples. For these visualizations, having access to information about the sample conditions, such as experimental conditions, is essential for the identification of related trends.

Next, individual statistical comparisons can be illustrated using statistical measures such as p-values, fold changes, or average expression. This type of visualization is often informative by indicating the strength of the underlying signal in the data and can reveal different technical artifacts. Common examples of statistical visualizations include the p-value histogram ((a) in Figure 19), the MA-plot ((b) in Figure 19), and the volcano plot ((c)-(d) in Figure 19). P-values are expected to be distributed evenly between zero and one when no effect is present, but with a spike appearing close to zero when a signal is present (as seen in (a) in Figure 19). The p-value histogram gives a sense of the comparison's strength (the spike's



(a) Density

(b) Number of missing values in different samples



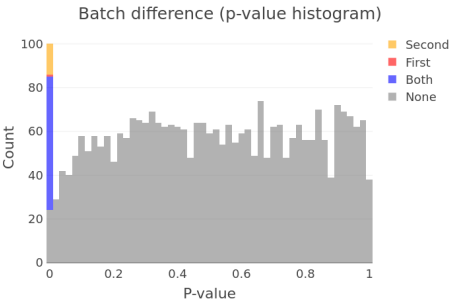
(c) PCA

(d) Boxplots

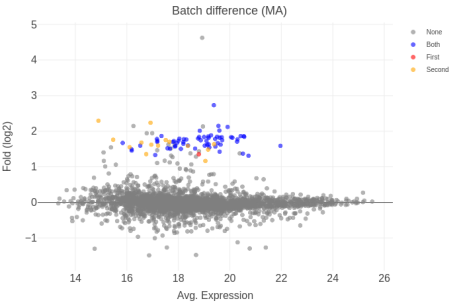
Figure 18: Sample-level illustrations of a dataset influenced by two known batch effects, illustrated by the colouring. Figures generated by OmicLoupe.

size) and whether there are underlying distortions (a non-even background distribution). The volcano and the MA-plots illustrate the fold changes in combination with either the average intensity or p-value. These visualizations can reveal trends such as an overall skew in expression direction, outliers, or other anomalies (Breheny, Stromberg and Lambert 2018; Li 2012). **Paper II** allows a direct comparison of features between two statistical comparisons, revealing the distribution of both jointly differentially expressed features and how those only found passing the threshold in one comparison are changing ((c)-(d) in Figure 19). In Figure 19, we see how a set of peptides is added at a different concentration, which is subsequently processed using two different software approaches. Most of the spiked-in features are identified by both methods (blue), while the red and yellow illustrates how those only found significant in one case are distributed. Using OmicLoupe, the user can directly interact with these figures to inspect annotations of each feature, and with one click zoom in to study the underlying data similarly to as shown in Figure 20.

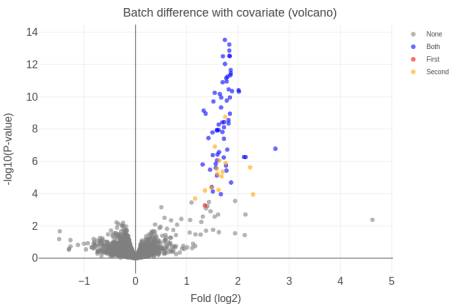
Overlap plots are useful for spotting common or diverging trends among several either statistical comparisons or features missing in certain conditions. Popular tools to do overlap



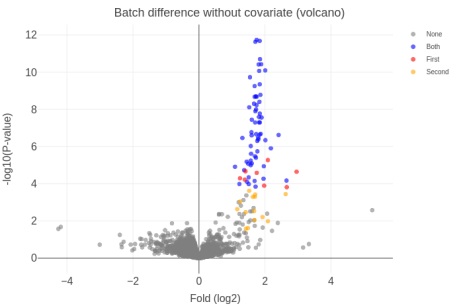
(a) P-value histogram



(b) MA-plot



(c) Volcano plot, software method 1

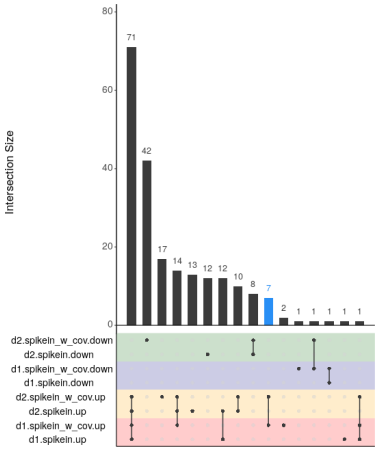


(d) Volcano plot, software method 2

Figure 19: Illustration of statistical comparisons with known spiked-in proteins and either adjusted or unadjusted for a known batch effect by specifying it as a covariate to the statistical test. Figures generated by OmicLoupe.

plots are the Venn diagram (for few comparisons) and the UpSet plot (Conway, Lex and Gehlenborg 2017) for a higher number of comparisons. OmicLoupe, the software presented in **Paper II**, interactively provides these visualizations and extends them by comparing fold directions for abundance changes. These visualizations can quickly reveal where the trends are similar (having similar fold change direction for features that pass multiple significance thresholds) or different (having many features passing significance thresholds in multiple comparisons with the reverse regulation direction). OmicLoupe further allows direct inspection of features underlying the overlaps. These overlap visualizations have proved to be a useful tool for understanding similarities between statistical comparisons during data exploration in **Papers III-V**. One example of overlap visualization using an upset-plot and illustrating where the fold-direction is same or different is illustrated in Figure 20.

The data for single features underlies all previously mentioned visualizations. Inspecting single features can reveal unexpected patterns otherwise hidden. This distribution can be illustrated using boxplots or violin plots (a type of density plot) but is for a smaller number of data points, preferably illustrated by showing the data directly. In **Paper II**, the single

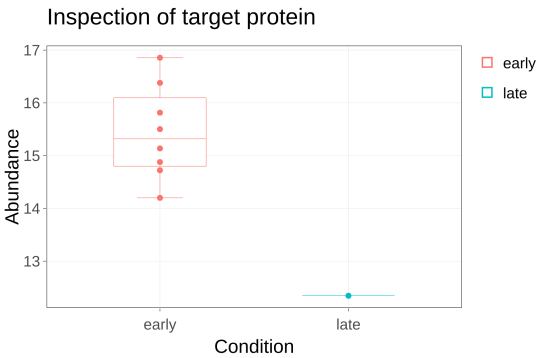


(a) Upset plot illustrating shared proteins significant across different comparisons

	comb_id	d1.Average.RT	d1.Cluster.ID	d1.Peptide.Sequence
1	C6	28.625	1418128225789	GSVMNPNDHPHGGGEGK...
2	C7	29.281	1418128214811	SATQDDAQNMCK Oxid...
3	C11	29.982	1418128176202	AHAEMQNIQR Oxidat...
4	C14	30.404	1418128212667	ILQMHANSR Oxidati...
5	C22	31.023	1418128177422	LSGTADRPR
6	C23	31.047	1418128183757	VVPAGQAGNR
7	C35	31.693	1418128182105	NVRPGEPNDPK
8	C39	31.885	1418128211584	KADESDEDEMVR Oxid...
9	C40	31.894	1418128226252	VVTTGAGNPGK
10	C42	31.99	1418128211134	VVESEDPKAQEK

Showing 1 to 10 of 474 entries

(b) Table for highlighted overlap in Upset plot



(c) Single-feature illustration over condition

Figure 20: Illustration of the workflow going from overlap plots (here - an UpSet plot) to inspecting a table with underlying features (corresponding to the blue column in the UpSet plot) to inspection of single features. The linked dots below the UpSet plot illustrates the overlaps, and the size of the bars the number of features within that overlap. Figures generated by OmicLoupe.

feature visualizations are made easily accessible from both statistical and overlap plots to allow effortless inspection of underlying data points and can be coloured on any sample condition. One example is seen in Figure 20, showing a feature found significant only with batch effect correction. As shown in (a), one data point belongs to a different batch and is a strong outlier compared to the others in the same conditions, seen in b). The impact of the batch effect is reduced after batch correction and the feature appears as significantly different.

Several of these plots are provided in part by commonly used visualization software in pro-

teomics such as Perseus (Tyanova and Cox 2018), DanteR (Polpitiya et al. 2008) and commercial platforms such as QluCore and the Proteome Discoverer. Recently, a wave of tools based in the R package Shiny (also employed in **Paper II**) has emerged. These tools often fill different niches not covered by the dominant visualization software (Shah et al. 2019; Chang et al. 2018; Nagaraj et al. 2015; Rigbolt, Vanselow and Blagoev 2011). To be useful, visualization software needs to be user-friendly and provide a needed collection of visualizations. OmicLoupe provides a set of visualizations which was frequently used during the studies in **Papers III-V**, and was developed subsequent to a student project doing initial explorations of how to best use single-feature visualizations for rapid understanding of the data (Lindh 2020), and the interactive interface developed to the data presented in **Paper III**. Now, OmicLoupe focuses specifically on comparing multiple datasets or statistical comparisons and traversing from these to the single-feature level, providing functionality not present in the other mentioned software.

In conclusion, visualization fills an essential role in proteomics analysis, is central in both **Paper I** and **Paper II**, and has been of key importance during the analyses performed in **Papers III-V**. If used well, accessible visualization tools increase the understanding of the data at hand, leading to better decisions on how to perform the analysis and thus more robust biological findings.

Building robust software for omics analysis

When performing biological data analysis, it is not uncommon to reach a situation where existing tools do not fully meet the needs at hand. This situation might require writing new code to solve the task, which sometimes is later provided as a software to a broader audience. To maximize the chances of this software being used requires paying attention to good software development practices, user-friendliness and realizing the constraints of developing software in an academic context.

Much of scientific research builds on non-commercial academic software. This ecosystem has advantages such as the prevalence of open-source software, allowing other developers to inspect and build on top of existing code. This code reuse is facilitated by package repositories such as CRAN and Bioconductor (Gentleman et al. 2004) for R and pip and BioConda (Dale et al. 2018) for Python, and for public code-storage locations such as the popular website GitHub (https://www.github.com). Without this kind of repositories, neither **Paper I** nor **Paper II** would be possible, as both extensively use existing R packages. The software presented in **Paper I** is now part of Bioconductor, and both software are openly available as R packages on GitHub, contributing back to the ecosystem. Further, as the bioinformatics field rapidly expands to include new types of data and algorithms, the rapid publishing of new software allows researchers to adopt cutting edge algorithms to analyse

their data. Still, academic software development has substantial challenges which tend to be different from those seen in traditional software engineering. The structure of software development in academia differs from commercial software development. Academic software is often maintained by a lone Ph.D. student or post-doc (Altschul et al. 2013), primarily motivated by reaching the point of publication (Mangul et al. 2019). Maintenance beyond this point may give citations and recognition where successful examples include the open-source OpenMS and the non-open-source MaxQuant/Perseus in proteomics, as well as open-source R packages such as DESeq2 and Limma in transcriptomics. The downside of maintenance and further development is that it requires prioritizing time away from other projects, which could be spent working towards additional publications. Proposed incitements for maintenance of existing software include using GitHub (a common web page to store open-source code) metrics to measure the prestige of the software (Dozmorov 2018) or implementing rigorous standards the software need to pass for publication (Mangul et al. 2019). It seems to me like neither of these address the core issue with publications being the main driver. Further, the Ph.D. students and post-docs inevitably transition to a different position, leaving after them a codebase often rarely seen by anyone else. This codebase is often built without formal training in software development, which stands in contrast to the code review, shared ownership and continuous maintenance of many commercial software. Another proposal is to hire trained software engineers (Lawlor and Walsh 2015) in academia, which would require competitive and sustained funding to attract good software engineers to take on the task.

Beyond the challenge of building and sustaining software is to build software usable by the audience. For instance, documentation is critical for software usability (Karimzadeh and Hoffman 2018; Marx 2020) and should preferably be written during software development. User-friendliness is also critical but can be difficult to achieve, requiring users that provide feedback on the software, the developer time to make the adjustments, and additional developer skills beyond the scientific and programmatically. Another long-term investment in software is the use of systematic testing, ensuring the correct output from a given software over time as additional changes are introduced (Zeeya 2010). This can be implemented as unit-tests running data for which the expected outcome is known through individual functions within the code. Another approach which may require less effort to implement are system-wide tests, where a dataset giving a known output is run through the software on every substantial update to verify that it still is identical to before the changes. The downside with system-wide tests is that they may miss parts of the program, and when errors occur, it may be more difficult to find the source. Both types of tests proved highly useful during the development of **Paper I** to catch unintended side effects by code updates. Finally, moderate knowledge in software development tools such as version control and Docker- or Singularity-containers (Kurtzer, Sochat and Bauer 2017) can help significantly in the management and deployment of the software. These tools can help ensure the reproducibility of analyses using the software. On the optimistic side, recent technological

developments such as Shiny in R and Dash in Python provide accessible ways to develop software with an interactive user interface. These tools make it possible to build simple and usable interfaces for algorithms otherwise hidden in command-line software, only accessible by technical users. This has led to a recent influx of this type of software. Still, these graphical pieces of software are more complex to maintain, increasing the challenge in implementation and maintenance.

Academic software fills a vital role in life sciences and has provided many widely used software driving research forward. Still, they are plagued by low implementation quality and often limited maintenance. This phenomenon is likely driven by the current lack of incitements to improve existing software quality and longevity, and in some cases due to limited training of those building the software. I hope the field eventually manages to address these widely recognized issues (Zeeya 2010; Goble 2014; Jiménez et al. 2017; Grüning et al. 2019). Meanwhile, technological improvements continue to provide more developer-friendly tools, new learning resources are developed making it easier to pick up the needed technical skills, and ecosystems such as Bioconductor and BioConda continue to promote the usability of software in life sciences.

Chapter 3: Discovery of proteomic biomarkers for sustainable agriculture

Proteins as biomarkers for molecular breeding

The breeding of plants and animals plays an essential role in securing the global food supply. This importance is emphasized by the current widespread changes to the climate. Breeding allows for continuous adaptation of crops and livestock for traits as disease resistance, the ability to grow in previously inaccessible regions, and increased yield (Salekdeh and Komatsu 2007). Traditional breeding uses directly observable characteristics seen at a phenotypical level to decide which individuals can breed the next generation. Conventional breeding has been successful in the past but is limited as it requires the individual to be fully grown before studied and as phenotypical traits are often influenced by a combination of genes. Here, molecular breeding can provide tools to speed up the breeding both by early measurements and by identifying which genes underlies complex traits (Jiang 2013a). One common approach to molecular breeding is to use variations in the genomic sequence and relate their positions to the genes associated with the trait of interest (Jiang 2013b; Meuwissen 2007; Desta and Ortiz 2014). Regions in the genome known to be linked to certain characteristics are called Quantitative Trait Loci (QTL). Using these markers allows tracing individual genes across generations and can be studied before the organism is fully grown, speeding up breeding (Jiang 2013a). This technique has helped advance agriculture (Langridge and Fleury 2011) as a complement to traditional breeding (Das, Paudel and Rohila 2015). Still, many challenges remain to be addressed (Collard and Mackill 2008; Jiang 2013a; Nakaya and Isobe 2012), and it has this far mainly been used to target known single or few genes linked to a trait (Collard and Mackill 2008; Wang and Chee 2010).

Expression data such as those obtained in transcriptomics and proteomics can be used to simultaneously profile the expression levels of thousands of genes, which has proved useful as an addition to genomic markers. Differential expression analysis is a common strategy to analyse this type of data (Velculescu et al. 1995). Here, the aim is to identify gene products

differing in abundance between conditions of interest. The proteins or transcripts identified as different can then be used for purposes such as to better understand the underlying biological mechanisms or for the development of biomarkers. In proteomics, this has been used extensively to profile valuable traits in plants related to factors such as growth, ripening and handling of different types of stresses (Tan, Lim and Lau 2017), and is used in **Paper III** and **Paper V** to identify proteins differing between conditions of interest. Analysis of the abundance of gene products can also be used to identify regions in the genome linked to the expressed quantities of that gene product. These regions are called expression quantitative trait loci (eQTL) and are often categorized as either being close to the location of the expressed genes (called *cis* eQTLs) or in distal parts of the genome (called *trans* eQTLs). These eQTLs provide additional information beyond the QTLs and can help link SNVs to molecular mechanisms (Gilad, Rifkin and Pritchard 2008). Both transcriptomic and proteomic expression data can be used to identify eQTLs. By studying the proteins directly, we are closer to the phenotype (Das, Paudel and Rohila 2015; Sabel, Liu and Lubman 2011), which gives a better view of what biology is behind the phenotype as the correlation between the transcriptome and proteome often is low (Nie et al. 2007; Maier, Güell and Serrano 2009), and due to that proteins are further modified with PTMs. Thus the transcriptome will not capture the full variation present in the proteome, and proteomics may be used to identify molecular relationships not easily identified using only genomics and transcriptomics (Das, Paudel and Rohila 2015; Langridge and Fleury 2011; Diz, Martínez-Fernández and Rolán-Alvarez 2012; Su et al. 2019). Attempts to incorporate the proteome expression in marker discovery have previously helped identify complex QTLs linked to valuable traits (Damerval et al. 1994; De Vienne et al. 1999; Gunnaiah et al. 2012; Eldakak et al. 2013; Consoli et al. 2002; Amiour et al. 2003; Rodziewicz et al. 2019) and revealed mechanistic understanding underlying these traits. This trend will likely continue as the increasing presence of reference genomes and technique developments in proteomics is making it easier to carry out this type of studies.

A difficulty when working with plants is the complexity of their genomes, with plants such as oat being hexaploid having six copies of each gene. This makes the finding of robust QTLs more challenging (Wu and Hu 2012). The use of direct measurements of proteins as markers could circumvent this and has long been discussed for use in biomarker discovery in clinical settings (Rifai, Gillette and Carr 2006; Whiteaker et al. 2011). More recently, proof-of-concept approaches for using direct protein measurements to predict agricultural traits have been demonstrated and used for predicting resistance to the oomycete *Phytophthora infestans* in potato (Chawade et al. 2016), and for predicting resistance to *Ascochyta* blight in pea (Castillejo et al. 2020). These studies were carried out using the proteomic technique Single Reaction Monitoring (SRM) in the first case, and shotgun-DDA combined with DIA in the second. SRM, also known as Multiple Reaction Monitoring (MRM) (Wolf-Yadlin et al. 2007), measure specific previously known peptides in the mass spectrometer and have proven useful due to a relative simplicity and high accuracy. Still, protein expres-

sion levels are generally measured in relative levels, comparing the difference in abundance between groups of samples. Attempts to quantify absolute abundances of protein levels (AQUA) are on their way and may, over time, remedy this issue (Gerber et al. 2003), further increasing the potential of using protein abundances in molecular breeding.

Molecular breeding is a changing field, with proteomics showing an increasing promise. Proteomics provides an explorative technology that can identify proteins linked to traits of interest, which could subsequently be used to improve on existing gene linkage maps or directly studied as markers. It has demonstrated its utility in several studies, and as the techniques continue to be developed, it will likely be further used, improving our ability to shape our food.

Investigating *Fusarium* head blight infection in oat

Oat (*Avena sativa*) is a widely important crop with high nutrient contents (Gorash et al. 2017) and many demonstrated health benefits (Martínez-Villaluenga and Peñas 2017) such as reduction of blood cholesterol (EFSA 2010) and high levels of beta-glucans, which have shown benefits both for industry and human health (Ibrahim and Selezneva 2017; Gorash et al. 2017; Biel, Bobko and Maciorowski 2009; Daou and Zhang 2012). Fusarium Head Blight (FHB) is a fungal disease both harming the health of humans and livestock by emitting a toxin called deoxynivalenol (DON) (Escrivá, Font and Manyes 2015; Alshannaq and Yu 2017) and causing widespread economic costs (Martinelli et al. 2014; Tekauz et al. 2004). Resistance breeding has been argued to be one of the most promising strategies to tackle diseases in plants (Brown 2015), reducing the disease pressure and the need to spray fungicides. It has been successfully employed for other diseases such as crown rust in oat (Lin et al. 2014), and it has been proposed as a strategy to control *Fusarium* (Bjørnstad and Skinnes 2008). QTLs related to DON resistance have been identified (He et al. 2013), indicating the presence of genes related to the resistance. On a proteomic level, there have only been few studies in oat to date (Chang et al. 2011; Bai et al. 2016; Chen et al. 2016; Rajnincová, Gálová and Chňapek 2019; Bai et al. 2017; Zhao et al. 2019), likely in part due to the previous lack of published reference genome. At the point of publishing, the study presented in **Paper III** had the deepest proteomic coverage to date in oat. Recently the first reference genomes in oat was published for a diploid oat variety (Maughan et al. 2019) and the first full sequencing of a hexaploid oat variety was made available online by a commercial company (PepsiCo 2020). These advances will reduce the barrier to perform proteomic studies in oat in the future.

The aim of **Paper III** was to characterize the molecular response of oats to *Fusarium* head blight. We confirmed the differences in disease response between the commercial oat variety Belinda and the partially resistant variety Argamak, and carried out a proteogenomic study

of its response during infection.

Analysis decisions

The starting material for the analysis is transcriptomics data from the two oat varieties Belinda (a commercial variety not resistant to *Fusarium* species) and Argamak (a Russian non-commercial variety shown to have partial resistance to *Fusarium* species), and proteomics measurements of infected and non-infected varieties at different time points. This setup is illustrated in Figure 21.

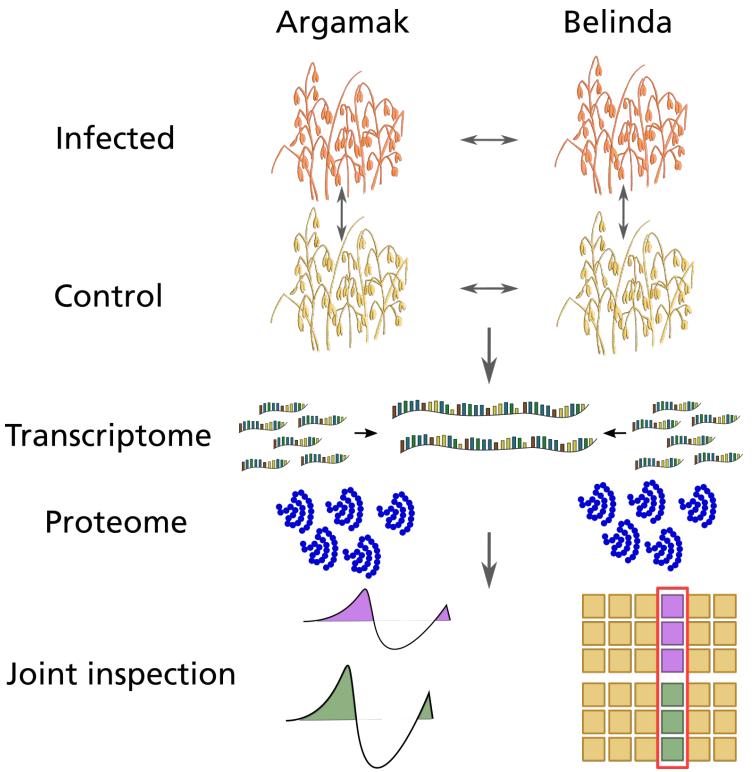


Figure 21: Experimental setup for oat study (part of figure adapted from Paper III).

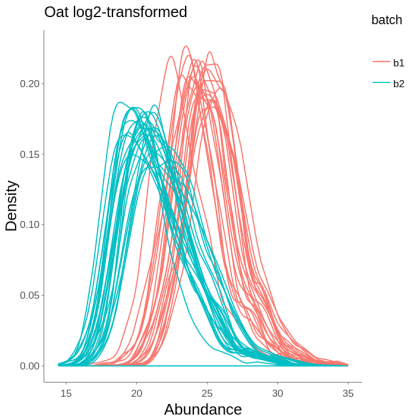
Due to the lack of a reference genome sequence, the sequenced transcriptome was assembled into a reference through a process called *de novo* assembly, where the transcriptome is sequenced and built into a reference representing the actively transcribed parts of the genome. This provides the opportunity to distinguish variety-specific sequence variations. Still, it gives comparably more complex reference, with many transcripts related to

single genes, and sometimes causes redundant transcripts from the assembly process. Due to the complexity of the dataset, a customized R Shiny interface was developed to allow further inspection of the generated data. This interface was published together with the dataset and involves several visualizations and analyses, such as gene ontology enrichment and screening for sequence variations in the assembled sequences from the two varieties. Some of these visualizations were later incorporated in OmicLoupe (presented in **Paper II**).

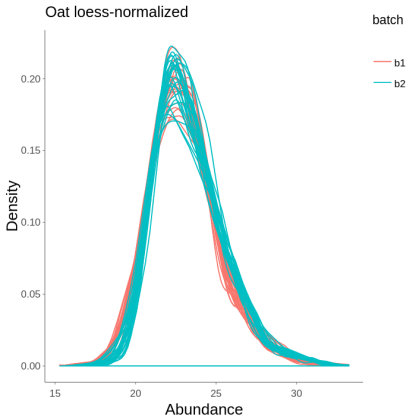
NormalyzerDE was used to perform the initial screening of the dataset and identified cyclic Loess normalization as well-performing. During the exploration of the dataset, two separate batch effects were identified, illustrated in Figure 22. The first, the most dramatic one, accounted for the majority of the variation in the PCA analysis ((c) in Figure 22) and was likely caused by variations in the sampling handling protocol. The impact of this batch effect was deemed too large to feasibly correct for using batch correction strategies. When inspecting the patterns within the groups of the samples, the samples belonging to one group were deemed less reliable based on the number of missing values. It was decided to focus on the higher-quality set of samples. Furthermore, a smaller batch effect was identified related to the run order in the mass spectrometer, where a drop in the number of identified MS2 spectra was seen. This second batch effect was much weaker, but was confounded with the infection state which prevented the use of batch effect corrections as the technical variation was inseparable from the biological. During the data analysis it was found that full sets of samples for doing comparisons between Argamak and Belinda at four days after infection were intact within these sets of samples, and thus not influenced by any known batch effects (shown in Figure 22 (d)). Further, the number of missing values was compared between the two groups of samples with no systematic differences found. Based on this, it was decided to do an explorative comparison to identify peptides only present during infection in each variety.

One sample was lost and not present in the final obtained data reducing one of the statistical comparisons to three versus two samples. For the statistical comparison, Limma was used, which is less susceptible to differences in variation caused by few replicates (further described in Chapter 2), but the lack of replicates will still limit the sensitivity of the experiment. Target candidates were further assessed using the Shiny interface to identify putative mutation sites, which could potentially be involved in the differences of these proteins, as illustrated in Figure 23, showing a sequence variation underlying one of the proteins found differentially expressed between the varieties during infection.

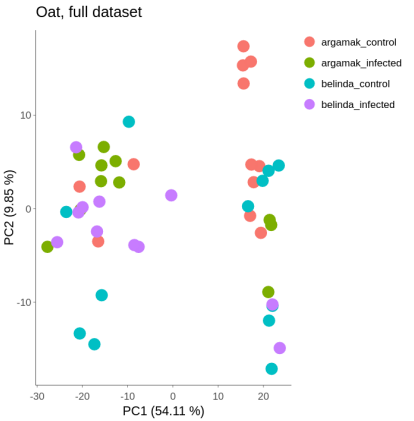
In the end, sets of proteins found as differentially expressed between Argamak and Belinda during infection and non-infection were identified. Further, the qualitative analysis identified proteins uniquely present during infection in both Argamak and Belinda. These results were used for further enrichment analyses and explored for protein-specific underlying mutations, as shown in Figure 23.



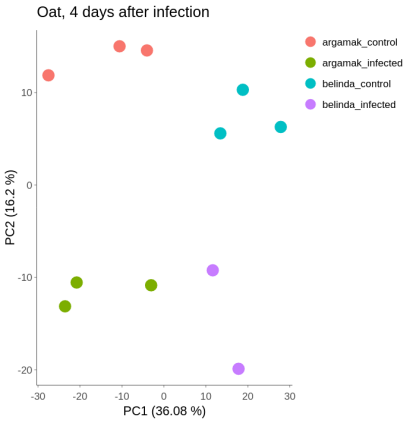
(a) Distribution of samples in batches without normalization



(b) Distribution of samples in batches with normalization



(c) Illustration of the major batch effect



(d) Separation of samples within 4 days after infection

Figure 22: Illustrations of samples and batch effects in the oat dataset (illustrated using OmicLoupe).

Key findings

In conclusion, several analysis decisions had to be made throughout the analysis of this dataset in order to reliably tackle the presence of batch effects. Visualizations were crucial to identify these, which otherwise might have gone unnoticed. At a physiological level, the partial resistance in Argamak was confirmed by measurement of DON content, indicating a slower disease progression when compared to Belinda. Electron microscopy images indicated a difference in wax production between the two varieties. For the proteogenomic

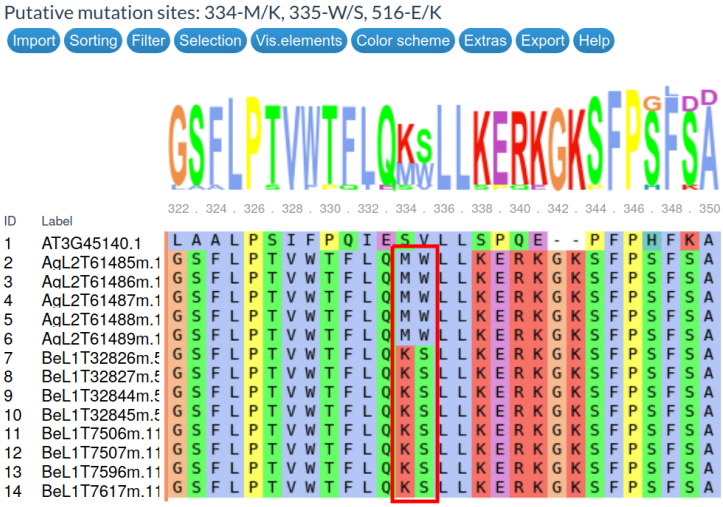


Figure 23: Interactive exploration of variety sequences. Two adjacent amino acids were found different between the two varieties for a protein homologous to lipoxygenase, one of the proteins differentially expressed between the two infected varieties. (Adapted from Paper III, using the R Shiny interface developed for the dataset).

analysis, several proteins linked to the differing disease response were identified by statistical comparisons between the two oat varieties and by qualitative analysis of peptides in different infection states. These are explorative findings that could be further investigated in future studies. Finally, this provides the deepest proteomics dataset to date in oat, a valuable molecular resource for further research both within oat in general and during response to *Fusarium* infection specifically. These findings could help the breeding of oat varieties with a higher resistance towards *Fusarium* head blight, thus contributing towards a more sustainable agriculture. For further reading, see **Paper III**.

Finding robust markers for bull fertility in seminal plasma

Bull fertility is a critical trait in breeding, with unsuccessful insemination attempts being costly for breeding facilities, and simultaneously slowing down the breeding of desired traits (Butler et al. 2020). Many factors are known to influence fertilization rate in cattle related to the viability of the sperm (Butler et al. 2020), the fertility of the bulls themselves, and to the freezability of sperm (Rickard et al. 2015; Leahy et al. 2020). The protein composition of seminal plasma, the surrounding liquid with which sperm is ejaculated has been shown to influence the sperms ability to fertilize (Robertson 2007; Rickard et al. 2014). Furthermore, many other factors such as the season of the year are known to influence the

fertility (Stott 1961). Estimating the fertility of bulls by directly measuring the success rate of inseminations is slow and expensive, with bulls having to reach a mature age and be used in enough inseminations before obtaining a reliable estimate of their fertility (Utt 2016). It would thus be valuable to have measures to detect lowly performing bulls at an early stage (Braundmeier and Miller 2001). Fertility as a trait is complex and involves many factors. Genomic studies have identified SNVs (single nucleotide variation, differences in the genome sequence) thought to be related to it (Abdollahi-Arpanahi, Morota and Peñagaricano 2017), but could likely further benefit from the additional information present in the proteomics. In recent years, the first studies comprehensively profiling the bull seminal plasma proteome have been presented. The proteome of spermatozoa and of the seminal plasma (Druart and Graaf 2018) have been investigated, and different aspects of the role of membrane proteins during the fertilization have been studied (Leahy et al. 2020). Further studies have investigated what proteins are transferred from the seminal plasma to the spermatozoa (Pini et al. 2016) and the role of freezability on the ability of sperm to fertilize (Gomes et al. 2020).

In this study, we extend on the knowledge about the seminal plasma proteome by following a set of bulls with varying fertility over three separate seasons to identify proteins robustly correlated with fertility. The identified set of proteins is built into a predictive signature and assessed in an independent cohort (as illustrated in Figure 24). Here, the aim is to find a molecular basis for identifying bulls with a low fertility rate at an early stage which would save large resources for the breeding facilities.

Analysis decisions

The data used in this analysis consists of three sets of proteomic measurements across three seasons from 20 bull individuals with varying fertility were collected as double ejaculates, followed by a set of proteomic samples from a separate set of 17 bulls. The target was to identify proteins robustly correlated with fertility, particularly considering variation from both season and resamplings. Further, the first set of samples were carried out in duplicates to assess the technical variation, and four samples were rerun together with the second batch to investigate the extent of which the mass spectrometry influenced the outcome.

NormalyzerDE was used for initial outlier detection and for assessing the normalization techniques, deciding on cyclic Loess for the first dataset, and staying with it in the subsequent datasets to not introduce additional differences between the samples. Upon inspection using sample-level visualizations, two types of outliers were identified. All samples taken from one specific bull appeared consistently different from the others across all three seasonal measurements. This was confirmed with the breeding station, who knew from before that this bull was different, and thus confirmed it as a biological outlier. Beyond that, one sample was found exceptionally different in both sample-level plots and density

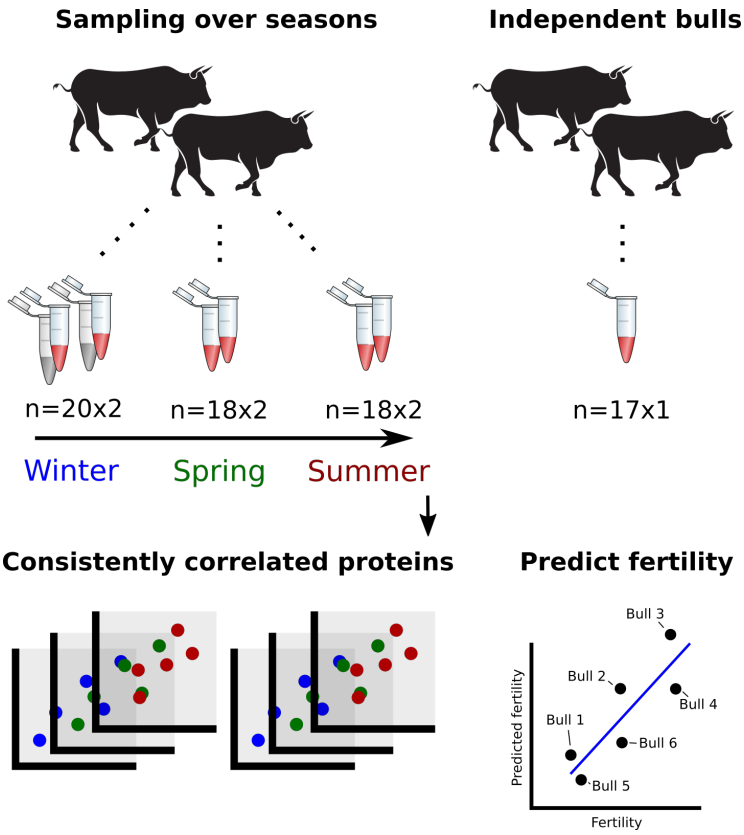
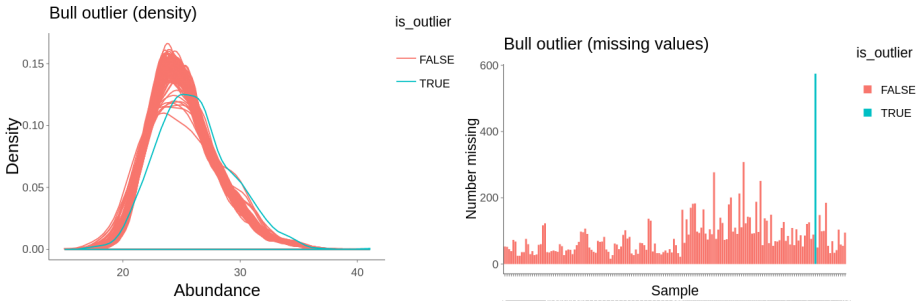


Figure 24: Experimental setup for bull study (adapted from Paper IV).

curves, as illustrated in green in Figure 25, having a high number of missing values and a distorted density profile. This sample was omitted from further analysis. OmicLoupe was applied to assess similarities between fertility-related differences in the different sets of bulls, showing a high similarity when comparing how proteins correlated with fertility in the three seasonal samplings, and a low similarity when comparing this correlation with how the proteins correlated in the independent set of bulls, further discussed below.

Originally, the bulls were divided into groups based on their estimated fertilities classified as 'HIGH' and 'LOW'. This resulted in the identification of proteins with different abundances in the groups, but upon further consideration it was decided to change it to correlation between bull fertility and the outcome as this better captures the continuous nature of the fertility and avoids the need of using an arbitrary classification cut-off. As each set of samples constitutes both a biological batch (due to seasonal variation and other biological effects) and a technical batch (due to being sampled at different timepoints), it



(a) Illustration of outlier using sample density curves (b) Illustration of outlier by illustrating the number of missing values per sample

Figure 25: Outlier detection using OmicLoupe.

was decided to primarily perform statistical tests within these batches, and then compare the resulting lists. Variations over season was briefly explored using a repeated sample ANOVA, but as the season is confounded with sampling effects this data was difficult to draw conclusions from and was not further investigated. When assessing the statistical measurement, it was considered how to best handle the duplicate ejaculates from each bull within each time point. It was decided to merge these prior to statistical calculations, as they could not be considered independent samples (Reinhart 2015) coming from the same individual. Finally, two groups of proteins correlated with fertility were identified - one with Pearson correlations with consistently low p-values ($p < 0.1$) across all seasons (9 protein groups), and secondarily for proteins with low p-values across two seasons (34 protein groups). Based on these, we explored different machine learning models to predict fertility, selecting a linear regression model based on three proteins due to its simplicity and relatively strong performance (illustrated in Figure 26). The best performing model was selected based on adjusted r^2 which penalizes the addition of additional predictive variables, balancing the predictive ability with the complexity of the model.

An independent cohort was collected which allowed testing of the developed predictive algorithms and comparison to previously observed correlations. Disappointingly, how the proteins correlated with fertility in this independent set of bulls showed an overall low similarity with the correlations found in the original sets of samples, including for the predictive model. This could in part be explained due to the narrow fertility range in the obtained independent set of bulls reducing its reliability (42-51 with one lower sample) compared to the seasonal samplings (35-60), and would require further investigations in future proteomics datasets. The exception was for one protein of particular interest (a lipase) which had shown a strong and clear correlation across all seasons. For this protein, all four shared underlying peptides showed a similar trend.

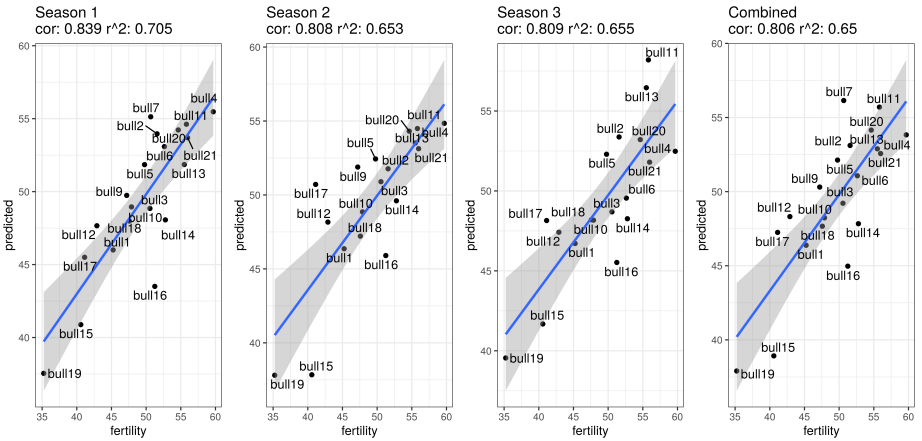


Figure 26: Predictions based on a signature built from three proteins, applied for seasons individually and for the median values across all three seasons (adapted from Paper IV).

Key findings

This study (**Paper IV**) led to wide profiling of the proteome in the seminal plasma of bulls over multiple seasons. Sets of proteins highly correlated to fertility were identified, some previously identified in the literature with similar trends and some novel findings. An independent dataset was generated, providing a chance to cross-check the findings, and although not successfully verifying the predictive signature, it showed similar trends for one of the most promising candidates. Still, further validations would be needed to establish which of these proteins are linked to the fertilization rate. Overall, this study acts as a foundation for further fertility research in bull, in particular in seminal plasma, and contributes towards reducing losses due to poor fertility in breeding. If successfully applied in practice, this could increase the efficiency of the breeding, allowing the same breeding to be performed using a smaller set of bulls, thus reducing its environmental impact (Scholtz et al. 2013) and the costs for the breeding facilities.

Identifying proteins linked to Nordic growth conditions

Potato is one of the most consumed crops in the world, providing a large part of both the energy intake and nutrient intake worldwide (Zaheer and Akhtar 2016; Camire, Kubow and Donnelly 2009). The changing climate causes new challenges for food security both through differences in climate and alternations of disease patterns (Lobell et al. 2008; Thornton et al. 2011; Dempewolf et al. 2014; Hijmans 2003). Global warming is expected to negatively impact the potato production, but this could be partially offset by adopting

strategies for where and when the crops are grown (Hijmans 2003). One potential strategy to adapt to the warmer climate is to shift the growing areas north (Haverkort and Verhagen 2008). To efficiently utilize these farmlands, the farmers need to adapt to the relatively higher number of sun hours and a shorter growing season. Using varieties better adapted for these conditions could play an important role in enabling this (Hellin et al. 2012; Varshney et al. 2011). Despite being a globally important crop, the current proteomic knowledge of potato as studied in the field is limited, and further omic-studies will play an important role in establishing a multi-omic view of potato in the field (Alexandersson et al. 2014). Here, the impact of growing different potato varieties at different latitudes in Sweden was studied with the aim of better understanding what influences the yield in relation to the differences in growth conditions while providing a deep proteomic profiling of potato as grown in the field.

Analysis decisions

This dataset consisted of proteome measurements taken from potato leaf samples collected in field trials during the years 2016, 2018 and 2019. For the first field trial, samples from 17 different varieties were collected, primarily in Borgeby (representing Southern Sweden), with some varieties including Desiree also sampled in Umeå (representing Northern Sweden). Out of these, 13 varieties were used in the final analysis. Further, in 2016, additional RNA-seq and metabolomics were collected for a smaller set of varieties giving a complementary view to the proteomics. For the subsequent years, Desiree was sampled at both locations. The experimental setup is illustrated in Figure 27.

NormalyzerDE was used for the initial screening of outliers and the identification of well-performing normalization methods. Cyclic Loess was again found to perform well and kept for the subsequent years analyses to not introduce additional variation by using different normalization methods. During the mass spectrometry analysis for the year 2016, the chromatographic column in the mass spectrometer was changed during acquisition of the sample set. This was later found to lead to an observable batch effect using a PCA plot (illustrated in Figure 28). The run order of the samples were randomized to balance the varieties, but not for the location, which led to imbalance across the batch in these comparisons. This was compensated for by rerunning a set of samples and incorporating the effect from the column change as a covariate in subsequent statistical tests. Similarly to in the bull study, a cross year batch-effect is present consisting of both the experimental variation and differences caused by the different samplings, which makes it difficult to directly compare samples taken during the different years. Instead, contrasts between Umeå and Borgeby were performed within each year, and the resulting lists of proteins compared, focusing on protein groups found differentially expressed across all years in Desiree. Furthermore, a comparison was made within Borgeby samples 2016 between groups of potatoes which

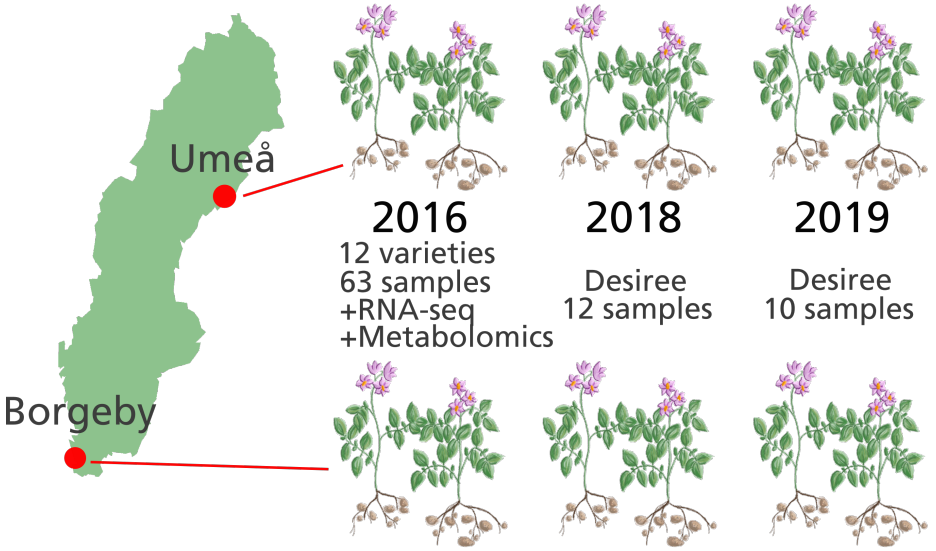


Figure 27: Experimental setup for the potato study (adapted from Paper V).

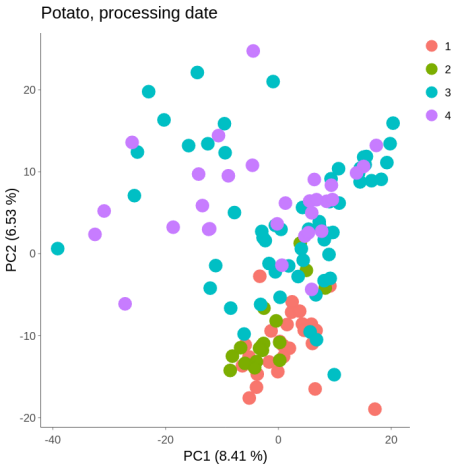


Figure 28: Principal component plot illustrating the impact of run order on the data generated from 2016 samples, with samples belonging to groups 1-2 performed before the column swap and samples belonging to groups 3-4 performed after the column swap.

showed a comparably higher and lower yield respectively in Umeå compared to Borgeby during that year, giving differentially expressed protein groups potentially linked to the relative geographic performance.

Outlier samples were identified using principal component analysis and density plots and did in some cases lead to these samples being reprocessed on the mass spectrometer. Using OmicLoupe to explore further trends identified that samples taken during one month in 2018 had a strongly different expression profile. This could potentially be related to the extreme weather conditions during this summer. It was decided to focus on the samples from the second month during this year which showed a higher overall similarity in expression patterns, as illustrated in Figure 29, showing the distributions of p-values when comparing year 2016 and 2019. In both cases, a clear trend is present (a spike around zero) and an even distribution of p-values show no apparent anomalies. Notably, in blue are features that are significant (here: $p < 0.05$) in both cases and differentially expressed in the same direction, while in green are those found differentially expressed in the opposite direction indicating a similarity in the trend even though many features are only significant in one comparison. Beyond the proteomics data, also transcriptomics and metabolomics data were available for

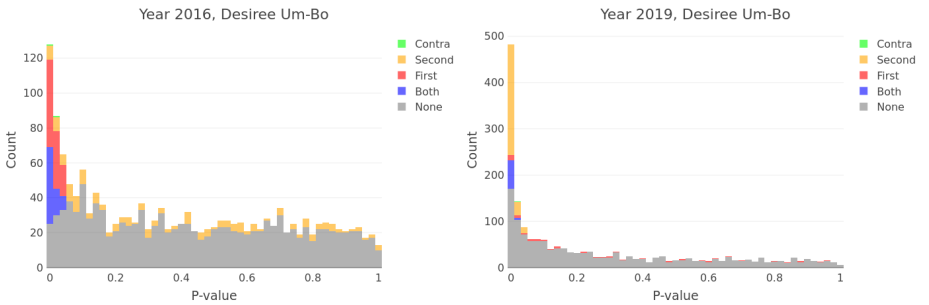


Figure 29: Comparison of expression patterns, 2016/2019 comparing Desiree between Umeå (Um) and Borgeby (Bo) using OmicLoupe. The left histogram shows the distribution of p-values for Desiree when comparing Umeå to Borgeby 2016, while the right histogram shows the corresponding comparison during the year 2019. The colours indicate in which datasets the proteins pass the significance threshold. Green (contra) show proteins which are differentially expressed in both with reverse fold-direction between the two datasets.

the 2016 samples, providing a comprehensive multi-omics view. Here, patterns found on the proteomic level showed overall similar patterns on the transcriptomics level, as identified using OmicLoupe.

Key findings

This study provides a unique multi-year view of the molecular profiles of field-grown potato varieties at different latitudes in Sweden. Here, we identified consistently differentially expressed proteins with similar differences in abundance across three years, with similar trends also seen at the transcriptomic level. We further identified proteins differing between varieties with measured comparably higher and lower yields in northern compared to southern Sweden. Further verification could reveal whether these changes are specifically related to

the differences in hours of light. After validation, these observations could be used to help select potato varieties that optimally can handle differing climates and thus better utilize these field areas and thus increase our ability to adapt to a changing climate. The dataset in itself represents one of the first proteomic profilings of potato grown in the field which may prove valuable in the light of future integrative omics-studies, and can help to understand the molecular variation that may appear in the field. Overall, this study provides a basis for further research into strategies for selecting potato varieties better adapted for the growing conditions in northern Sweden. For further reading, see **Paper V**.

Chapter 4: Concluding words

In this work I have explored the process of carrying out proteomics studies using label-free proteomics for biomarker discovery. Throughout the work, I have focused on how to navigate unwanted variation in the data and how make appropriate choices of software and methods. Many of these insights are generally applicable in omics beyond proteomics. This focus has led to the development of two pieces of software for improved computational analysis and analysis decisions in omics. Furthermore, these pieces of software have been applied across three separate studies, each studying agriculturally important traits in different organisms. An important aspect through both the software development and the applied studies has been to optimally handle technical limitations to maximize the potential of the datasets as explorative sources of biological understanding and biomarkers. These challenges can be encountered in all types of omics-data. It is my hope that the presented software and the conclusions drawn throughout this thesis will be of utility for other researchers who find themselves in a similar situation where careful analysis decisions need to be made to make the most out of the data at hand.

In **Paper I**, we introduced the software NormalyzerDE, a now well-used software available as a web application and as a Bioconductor R package. NormalyzerDE provides a performance screening of normalization techniques, introduces a normalization approach to consider the retention time-dependent biases such as those caused by electrospray ionization variations in mass spectrometry, and provides a tool for performing and visualizing the downstream statistical analysis. The output from NormalyzerDE is directly compatible with OmicLoupe, the software presented in **Paper II**. NormalyzerDE was used for the initial outlier detection and for informing the normalization selection strategy in **Papers III-V**, and to smoothen the downstream statistical analysis. In **Paper II**, an interactive visualization software called OmicLoupe was introduced, aiming to make diagnostic visualizations maximally available, and it was extensively used across the three applied studies. OmicLoupe was used to better understand the unwanted and wanted variations present in the studies, and in particular to understand to which extent biological trends were shared across multiple time points. Novel visualization techniques were developed for this purpose, some of which are demonstrated in Chapter 3. The software development has greatly

benefited from the fact that the software have been actively used both by me and others while being developed. This has provided continuous feedback which has helped to focus the development on the aspects which are most important for the user.

I have been fortunate to work with engaged and knowledgeable collaborators from whom I have learned a great deal throughout the projects. These studies are not single-person endeavours, and cross-discipline communication has been critical in carrying out these studies. The collaborations have given me valuable insight into the full proteomic workflow applied in agricultural studies, in two cases (**Paper III-IV**) leading me to participate all the way into the biological interpretations. **Paper III** explored sources of differences in resistance to the fungal pathogen *Fusarium graminearum* using a proteogenomic approach, confirmed the differing resistance between the varieties, and identified candidate proteins potentially involved in the disease response. Furthermore, using a custom-developed and now publicly accessible interface, mutations underlying some of these proteins were identified. This work contributes towards the breeding of commercial oat varieties with a stronger resistance to *Fusarium* species. In **Paper IV** we studied how proteins in the bull seminal plasma related to bull fertility varies across three seasons and samplings. Proteins with stable correlations were identified and used to build a predictive signature of bull fertility, which could be further validated in future studies. If validated, this would provide markers to detect bulls with low fertility with the potential to reduce the losses of materials while retaining the speed of developing new traits of interest. In **Paper V**, potato field trials were carried out over three summers. In this study we identified a set of proteins which were found consistently differing between northern and southern Sweden across three seasons. Here, we also identified a set of proteins differing between groups of varieties with differing yield at the two locations. The result from this study could be used to better understand climate adaptability in plants, and could ideally lead to the selection of crops better able to utilize the growth conditions in northern Sweden. Overall, these studies show some of the difficulties and opportunities in using proteomics for the further development of molecular breeding. With more studies coming out and the techniques steadily developing, I believe that proteomics will play an increasingly important role in the identification of biomarkers and to provide a deeper understanding of molecular biology underlying important traits. These will be employed for a wide range of applications, including the refinement of molecular breeding techniques, giving us the tools to shape our food more efficiently to increase the sustainability of our agriculture.

Outlook

Complex omics-studies are continuously being published at a high rate. Furthermore, new approaches such as single-cell technologies are becoming established, further increasing the challenges of the data processing. Here, I will take the opportunity to outline some

thoughts on future developments within the area of data processing in omics.

A key to improved statistical methods is considering what structures are present in the data at hand. As discussed, statistical procedures considering the multidimensionality of the data (Ritchie et al. 2015; Zhu et al. 2020; Pursiheimo et al. 2015) can outperform those that singly consider features, such as the commonly used t-test. Many of the existing methods used in proteomics are originally developed for microarrays. These methods could potentially be improved by considering technical variation linked to sample preparation effects such as run order or the performance of the mass spectrometer. Other examples would be to consider the prevalence and patterns of missing values and the unique behaviour of peptides with different physicochemical properties. Some of these characteristics have been successfully used in approaches to analyse proteomics data (Käll et al. 2007; Zhu et al. 2020; Gessulat et al. 2019), and can likely be further used to improve existing data analysis methods. A valuable resource for this purpose is the growing amount of high-quality public data (Perez-Riverol et al. 2018). Here, the practical utility has often been limited by the lack of sample-level information. Recent steps have been taken to address this (Perez-Riverol 2020), which if successful would greatly improve their utility.

A recurring challenge in handling technical variation is the difficulty of assessing whether the applied adjustments are made correctly, as adjustments always risk reducing the biological variation or introduce new erroneous signal. Software such as NormalyzerDE (**Paper I**) and NOREVA (Yang et al. 2020) are helpful through using visualizations to inspect the overall trends under different normalization methods, but can be challenging to interpret and provide measures on a sample-wide level. Again, using the unique structures in mass spectrometry data and the growing amount of data available could help identify specific peptides more prone to be influenced by certain types of bias, as explored in this work (discussed in Chapter 2). A more comprehensive profiling could give tools to provide confidence in whether the correction procedures are doing the right thing on both a sample- and gene product-level, thus leading to more robust findings, potentially increasing reproducibility.

Visualizations play a crucial role in understanding omics datasets. Here, in OmicLoupe, I have explored the idea of extending widely used visualizations by incorporating cross-dataset information. This idea could be extended by considering further aspects of the data, new visualizations, and other ways to integrate information across datasets. The ideal goal would be to help users consistently and more efficiently come to optimal conclusions on how to approach their data and provide tools to spot patterns that otherwise would have gone unnoticed. This could lead to new and more accurate results and a reduced cost and effort of the data interpretation, but requires that these visualizations are presented such that they are accessible and understandable.

The mass spectrometry technology and data analysis approaches are continuously develop-

ing, leading to new opportunities and challenges. Single-cell proteomics is maturing (Marx 2019), and with it comes a host of new problems to address. The data will be noisy and will require careful handling of unwanted variation. Tools will likely initially be repurposed from single-cell transcriptomics, opening for the possibility to enhance these if the unique aspects of the single-cell proteomics can be considered. In turn, these tools would together with the single-cell technologies have the potential to unlock an even more fine-grained understanding of biological systems.

In conclusion, many opportunities lie ahead in proteomics, both in the development of new software and the application of these to increase the robustness and utility of proteomics analyses, in agriculture and elsewhere.

Final words

In particular, this work has highlighted the need for understanding each of the involved steps in the complete omics workflow - from experimental design to carrying out experiments to data interpretations and, finally, to drawing biological conclusions. A coherent understanding of all these steps gives the best foundation for the data analysis. Well-designed software further gives the ability to navigate among limitations and opportunities in the data, revealing patterns otherwise not visible, and helps make reliable analysis decisions. Part of the issue could be related to the quote from Feynman: "The first principle is that you must not fool yourself - and you are the easiest one to fool." When inspecting the many patterns emerging from the bioinformatic analyses, even with the best intentions, it is easy to get lost in the analysis and to go for what is more compelling rather than what is robust, leading to findings that will fail to reproduce in other datasets. A solid understanding of the data and the employed statistical tools is critical to stick with what is likely to be accurate.

Molecular biomarker studies are difficult but a challenge worth taking on as they can beautifully contribute to solving among the biggest challenges facing us, from areas such as personalized medicine to shaping our food for a more sustainable agriculture. I believe strong cross-discipline collaborations, foundational understanding of the full biomarker workflow and sharp, accessible and well-documented software are important pieces in the puzzle to get us there.

Populärvetenskaplig sammanfattning

Allt levande är byggt från de byggstenar vi kallar celler. Dessa celler består i sin tur av olika typer av molekyler vilka vi kan mäta för att förutsäga deras egenskaper. Dessa molekyler kallas för biomarkörer, och kan användas för att accelerera forskning inom både jordbruk och medicin. Dagens jordbruk möter stora utmaningar i att både producera tillräckligt mycket mat till världens befolkning, och för att samtidigt anpassa sig till ett klimat i förändring. Biomarkörer har här en viktig roll i att skynda på avel av växter och djur genom att snabbare hjälpa oss att förstå vilka individer som har de egenskaper man vill ha, och kan på så sätt hjälpa jordbruket att möta dess utmaningar.

I det här arbetet mäter vi protein - den molekyl som utför större delen av arbetet i cellerna. Protein har många olika funktioner, till exempel att bygga strukturer, omvandla solljus till socker i växter och försvara celler mot angrepp av främmande organismer.

Arbetet består av två huvudspår. I den ena delen studerar vi biomarkörer i tre olika jordbruksprojekt. I det första jordbruksprojektet studerar vi hur två olika havresorter reagerar på angrepp från svamp, där den ena havresorten har ett mer effektivt försvar och den andra har ett sämre försvar, men ger en bättre skörd. Genom att studera skillnaderna bidrar vi till att utveckla havresorter som både kan försvara sig bättre mot svampangrepp och ge en bra skörd. I det andra projektet studerar vi hos tjurar hur protein i sädesvätskan påverkar deras fertilitet. Det är känt att protein i sädesvätskan påverkar spermans förmåga att befrukta, men kunskapen om hur det fungerar är fortfarande begränsad. Här identifierar vi protein som är relaterade till befruktningsförmågan, vilket kan bidra till att bättre kunna förutse tjurar med låg fertilitet vilken kan bespara stora resurser och underlätta aveln av andra viktiga egenskaper. Slutligen studerar vi hur olika potatissorter reagerar när de växer i norra och södra Sverige, där vissa sorter bättre kan utnyttja de annorlunda förhållandena i norra Sverige med längre dagar och kortare somrar. Detta bidrar till att förstå hur vi bättre kan använda jordbruksarealerna i norra Sverige.

För att mäta mängden av olika protein i celler använder man maskiner som kallas masspektrometrar, vilka kan mäta molekylers vikt med stor noggrannhet. För att mäta protein så delar man dem först i små bitar - peptider - som man skickar in i masspektrometern. Peptiderna skickas via en vätska genom vad som kallas en elektronspray - ett tunt munstycke som skickar ut en dimma av små droppar som sedan tillförs laddningar av en stark elektrisk spänning. Vätskan hos dessa små droppar dunstar snabbt bort, och kvar blir elektriskt laddade peptider. Laddade molekyler accelereras av elektriska fält och hur snabbt de accelereras beror på deras vikt och hur stark laddning de har. Detta används inne i masspektrometern för att mäta molekylernas vikt med stor noggrannhet. Peptiderna bryts sedan ned i små bitar genom att krockas med en gas under högt tryck. Slutligen mäts även dessa peptidbitar. Därmed har vi noggranna mätningar av vikten hos de ursprungliga peptiderna, och

mätningar av deras fragment. Dessa fragment kan ses som peptidernas fingeravtryck –något som unikt identifierar dem.

Mätningarna skickas sedan till en dator där en lång resa börjar för att pussla ihop en bild av hur mycket av olika proteiner som ursprungligen fanns i cellerna man mätte. Här räknar man först ut hur mycket som fanns av de olika peptiderna, och använder sedan deras fragment (deras fingeravtryck") för att jämföra mot en stor samling kända peptider och därmed avgöra deras identiteter. Sista steget är att använda olika datorprogram för att pussla ihop peptiderna till en bild av hur mycket av olika protein som fanns i det ursprungliga materialet. Dessa mätningar kan vi använda för att hitta biomarkörer.

Datorprogrammen man använder för att analysera protein är ofta svåra att använda och uppdateras ständigt med nya analysmetoder. Den andra delen av arbetet består av att utveckla två datorprogram som gör det enklare att hitta rätt metoder för att analysera proteindata. Det första programmet används för att illustrera protein-datan med olika typer av visualiseringar, vilket bland annat underlättar jämförelser när man upprepar ett experiment för att försäkra sig om att det man sett i ett första försök fortfarande finns där. Varje steg i mätningarna från experiment till mätning i masspektrometern tillför en viss osäkerhet i resultatet, och det finns en risk att detta ger en felaktig bild av den ursprungliga mängden protein. Det andra datorprogrammet hjälper användaren att välja den metod som bäst minskar mängden osäkerhet i protein-datan. Dessa datorprogram har båda använts i ovan nämnda biomarkörstudier för att minska osäkerheten i analysen och för att ge en bättre förståelse av datan.

Sammanfattningsvis ger detta arbete tillgång till nya datorprogram som kan användas för både studie av protein och andra molekyler - i jordbruk, eller andra biologiska områden som till exempel medicin. Dessa verktyg har sedan tillämpats i tre olika jordbruksstudier för att så bra som möjligt använda protein-datan, och för att hitta biomarkörer som kan användas för att snabba på utvecklingen av ett mer hållbart jordbruk.

科普摘要 (Popular science summary in Chinese)

所有生物都是从我们称为细胞的结构中构建的。这些细胞又由不同类型的分子组成，我们可以通过测量这些分子来预测细胞的不同特性。这些分子即被称为生物标记并广泛用于加速农业和医学领域的研究。今天的农业在努力为世界人口生产足够的粮食，同时适应气候变化带来的重大挑战。在农业研究中，生物标记物在促进动植物育种中起着重要作用，从而帮助农业应对这些挑战。

在我的研究中，通过测量蛋白质有助于生物标记物的发现和应用。绝大多数的细胞功能是通过蛋白质实现的，例如构建细胞骨架，参与光合作用，帮助防御外来生物的侵袭。

我的工作主要包括两个方面。在第一部分中，我们研究了三个不同农业项目中的生物标记。在第一个农业项目中，通过对比两个燕麦品种对真菌侵袭的反应，我们发现野生燕麦品种具有更强的抵抗力并确定了相应的生物标记物，而通过选择具有这种生物标记物的燕麦品种人们可以优化育种过程。在第二个项目中，我们研究了公牛精液中的蛋白质如何影响其生育能力。众所周知，精液中的蛋白质会影响精子的受精能力，通过发现与之相关的蛋白质，人们能更好地预测公牛的生育能力并节省大量资源。最后，我们研究了不同的马铃薯品种在瑞典北部和南部的生长情况。结果显示不同的生长条件影响马铃薯的产量，通过确定相应的生物标记物有助于人们选择在特定生长条件下产量更高的品种。为了测量细胞中不同蛋白质的含量，人们使用了被称为质谱仪的机器，它可以非常精确地测量分子的含量。首先蛋白质被分解为肽链，然后肽链逐一通过所谓的电子喷雾器被送入质谱仪。电子喷雾器是一种可以产生雾状液滴的细喷嘴，喷射出的液滴在外界电场的作用下附上电荷。随后这些带电的液滴迅速蒸发只留下带电的肽链，带电分子可以被电场加速，加速的速度取决于其重量和带电程度。它在质谱仪内部用于高精度测量分子。肽链在高压下与气体碰撞并分解成更小的片段。带电分子可以被电场加速，加速的速度取决于其重量和带电程度，质谱仪正是利用这个原理可以高精度测量分子。通过对这些小片段的测量我们可以还原原始肽链的含量和表达水平。

随后我们在计算机上处理这些原始数据，人们可以利用肽链的含量估算被测样品中存在多少种不同的蛋白质，然后通过对比已知的肽链表达进一步确定这些蛋白质的种类。这样我们便找到了可用作生物标记物的蛋白质。

我的研究的第二部分主要通过开发计算机程序帮助人们能更快更准确地分析蛋白质实验结果。其中第一个程序将原始数据通过不同类型的图表进行展示。例如在进行多次实验时人们可以通过对比图表更容易确定实验的可重复性，即第一次实验的结果在后续实验中依然可见。在科学研究中，测量数据的每一个步骤都会给结果增加一些不确定性，而累计的不确定性可能导致人们无法计算出正确的原始蛋白质含量。我开发的第二个计算机程序旨在减少数据分析中的不确定性，从而帮助使用者选择最佳分析方法。这两个计算机程序都已用于上述

生物标记研究中，在减少数据的不确定性的同时使研究者更好地理解数据。

总而言之，我的研究使得开发的计算机程序可以用于蛋白质和其他分子的研究，例如在生物和医学邻域。此外这些程序已用于各项农业研究中，研究者通过程序应用可以更好地理解数据，并找到可以帮助我们加快发展更可持续农业的生物标记物。

Acknowledgements

I need to start with a disclaimer - There are many more people who have been important parts of this journey. Even if your name is not here, this acknowledgment is to you too.

As I think most people who have gone through a PhD know, it can be a long and sometimes soul-searching journey. I am happy to say that my journey has been in most part, an enjoyable experience. This is to a large extent due to helpful supervision, interesting collaborations and a great environment at the Department of Immunotechnology.

In particular, I would like to thank my main supervisor, **Fredrik**, with whom I have had a great number of discussions throughout the years. You have always been helpful, while never hesitating to strike down on my half-baked ideas and sometimes poor grammar. I am thankful for that you, during this time, have given me space to pursue my own ideas.

My co-supervisors **Aakash** and **Erik**. You have given me guidance on many topics, from understanding what is important in biology to how to navigate in the world after the PhD. This guidance has been important to help me grow as a researcher and to help me navigate beyond my PhD.

To my collaborators at Alnarp and Uppsala. Working together with you has been one of the most satisfying parts of this work. I would like to give a particular mention to **Svante** who have sweated by my side towards the end, helping me make sure all the manuscript have come into place. **Patrice**, who guided my work in the bull project, with whom I have had many interesting discussions where I have learned a great deal about bull fertility. Also, **Svetlana** who did a great job helping me finish the oat project.

I have encountered many guiding figures over the years, who have shaped my path into research. **Urban** who guided my very first confused steps into bioinformatics and the web lab. **Björn** whose engagement has propelled me and so many others into the world of bioinformatics and open source. **Lukas**, my mentor, an inspiring figure who provided useful advice I still carry with me.

Big thanks to all who helped improving this thesis by giving feedback on different parts! In alphabetical order: **Aakash**, **Daniel**, **Danne**, **Deborah**, **Erik**, **Fredrik**, **Joana**, **Line**, **Magdalena**, **mom**, **Tim**, **Valentina** and **Xuan**. Very much appreciated. (I am also thankful to those who beyond this list expressed an interest in helping out!) A particular acknowledgement goes to **Zuzanna** who took on the significant challenge of drawing the cover art and did so spectacularly.

For the students I have helped supervising: **Line**, **Joel**, **Aaron**, **Shuyi**, **Deborah**, **Victor** and **Xu**. Being part of your journeys have been among the greatest learning during my PhD. I am happy to have shared these with you, and I wish you all the best ahead.

During these years, I have worked in a dedicated and kind environment at the Department of Immunotechnology. **Sara** and **Mats** - you have done a great job keeping the ship on course, A particular mention to **Cornelia** who has managed to keep the ship together over the years. Furthermore, I have been fortunate to work with a friendly bunch of PhD students and bioinformaticians. It has been great fun to be around you all. A special thought goes to my fellow PhD student **Sergio**, whom I wish everything goes well for ahead.

To the statistics club members **Erik**, **Line**, **Martin**, **Aaron**, **Xuan**, **Shuyi**, **Neli** and everyone else who have joined these discussions over the years. These discussions have made me start to really appreciate statistics and have made a deep impact on this work. I think you can recognize our discussions in some of the chapters.

Thank you, my old friends, from Ljungby and Lund, whom I sometimes have not contacted enough. We have spent much time together at different times, and I look forward to spending more time with you in the future. I appreciate your occasional attempts to understand what I am ”actually doing” and for despite my vigorous protests summarizing my work by naming me ”the potato guy”.

谢谢我在中国的家人。我非常高兴和你们相识。我希望我的汉语越来越好，这样我们就能更好地交流，更好地了解彼此。

Tack till min **familj**, vars stöd är så ständigt närvarande att det är lätt att ta det för givet. Ni vet att jag inte är en person av stora gester - men ni ska veta att jag uppskattar er mycket.

Xuan, we learn and we grow together. We take on the world and its hammers together. Through discussions and quiet study times, with you, every day is a joy. I cannot wait to take on whatever challenges lies ahead side by side with you.

Bibliography

Abdollahi-Arpanahi, Rostam, Gota Morota and Francisco Peñagaricano. 2017. "Predicting bull fertility using genomic data and biological information". *Journal of Dairy Science* 100 (12): 9656–9666. doi:10.3168/jds.2017-13288.

Alexandersson, Erik, Dan Jacobson, Melané A. Vivier, Wolfram Weckwerth and Erik Andreasson. 2014. "Field-omics-understanding large-scale molecular data from field crops". *Frontiers in Plant Science* 5:286. doi:10.3389/fpls.2014.00286.

Alshannaq, Ahmad, and Jae Hyuk Yu. 2017. "Occurrence, toxicity, and analysis of major mycotoxins in food". *International Journal of Environmental Research and Public Health* 14:632–651. doi:10.3390/ijerph14060632.

Altman, Naomi, and Martin Krzywinski. 2016. "Points of significance: P values and the search for significance". *Nature Methods* 14 (1): 3–4. doi:10.1038/nmeth.4120.

Altschul, Stephen, Barry Demchak, Richard Durbin, Robert Gentleman, Martin Krzywinski et al. 2013. "The anatomy of successful computational biology software". *Nature Biotechnology* 31 (10): 894–897. doi:10.1038/nbt.2721.

Amiour, N, M Merlino, P Leroy and G Branlard. 2003. "Chromosome mapping and identification of amphiphilic proteins of hexaploid wheat kernels." *Theoretical and applied genetics* 108 (1): 62–72. doi:10.1007/s00122-003-1411-0.

Andersen, Claus Lindbjerg, Jens Ledet Jensen, Torben Falck Ørntoft, J Ledet-Jensen, T F Orntoft et al. 2004. "Normalization of Real-Time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets". *Cancer Research* 64 (15): 5245–5250. doi:10.1158/0008-5472.CAN-04-0496.

Bai, Jianhui, Jinghui Liu, Weihong Jiao, Rula Sa, Na Zhang et al. 2016. "Proteomic analysis of salt-responsive proteins in oat roots (Avena sativa L.)" *Journal of the science of food and agriculture* 96 (11): 3867–3875. doi:10.1002/jsfa.7583.

Bai, Jianhui, Yan Qin, Jinghui Liu, Yuqing Wang, Rula Sa et al. 2017. "Proteomic response of oat leaves to long-term salinity stress". *Environmental Science and Pollution Research* 24 (4): 3387–3399. doi:10.1007/s11356-016-8092-0.

Ballman, Karla V., Diane E. Grill, Ann L. Oberg and Terry M. Therneau. 2004. "Faster cyclic loess: Normalizing RNA arrays via linear models". *Bioinformatics* 20 (16): 2778–2786. doi:10.1093/bioinformatics/bth327.

Barton, S.J., and J.C. Whittaker. 2009. "Review of factors that influence the abundance of ions produced in a tandem mass spectrometer and statistical methods for discovering these factors". *Mass Spectrometry Reviews* 28:177–187. doi:10.1002/mas.20188.

Bell, Alexander W., Eric W. Deutsch, Catherine E. Au, Robert E. Kearney, Ron Beavis et al. 2009. "A HUPO test sample study reveals common problems in mass spectrometry-based proteomics". *Nature Methods* 6 (6): 423–430. doi:10.1038/nmeth.1333.

Benjamini, Y., and Y. Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". *Journal of the Royal Statistical Society* 57 (1): 289–300.

Biel, Wioletta, Kazimierz Bobko and Robert Maciorowski. 2009. "Chemical composition and nutritive value of husked and naked oats grain". *Journal of Cereal Science* 49 (3): 413–418. doi:10.1016/j.jcs.2009.01.009.

Bjørnstad, Asmund, and Helge Skinnes. 2008. "Resistance to fusarium infection in oats (Avena sativa L.)" *Cereal Research Communications* 36:57–62. doi:10.1556/CRC.36.2008.Suppl.B.9.

Blainey, Paul, Martin Krzywinski and Naomi Altman. 2014. "Replication". *Nature Methods* 11 (9): 879–880. doi:10.1038/nmeth.3091.

Bolstad, B. M., R. A. Irizarry, M. Åstrand and T. P. Speed. 2003. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias". *Bioinformatics* 19 (2): 185–193. doi:10.1093/bioinformatics/19.2.185.

Bonferroni, C. 1936. "Teoria statistica delle classi e calcolo delle probabilita". *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* 8:3–62.

Braundmeier, A. G., and D. J. Miller. 2001. "The search is on: Finding accurate molecular markers of male fertility". *Journal of Dairy Science* 84 (9): 1915–1925. doi:10.3168/jds.S0022-0302(01)74633-4.

Breheny, Patrick, Arnold Stromberg and Joshua Lambert. 2018. "P-Value histograms: Inference and diagnostics". *High-Throughput* 7 (3): 1–13. doi:10.3390/HT7030023.

Brown, James K.M. 2015. "Durable Resistance of Crops to Disease: A Darwinian Perspective". *Annual Review of Phytopathology* 53:513–539. doi:10.1146/annurev-phyto-102313-045914.

Burger, Bram, Marc Vaudel and Harald Barsnes. 2020. "Importance of Block Randomization When Designing Proteomics Experiments". *Journal of proteome research*. doi:10.1021/acs.jproteome.0c00536.

Burkhart, Julia Maria, Cornelia Schumbrutzki, Stefanie Wortelkamp, Albert Sickmann and René Peiman Zahedi. 2012. "Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics". *Journal of Proteomics* 75:1454–1462. doi:10.1016/j.jprot.2011.11.016.

Bustin, Stephen A. 2014. "The reproducibility of biomedical research: Sleepers awake!" *Biomolecular Detection and Quantification* 2:35–42. doi:10.1016/j.bdq.2015.01.002.

Butler, Madison L, Jennifer M Bormann, Robert L Weaber, David M Grieger and Megan M Rolf. 2020. "Selection for bull fertility: a review". *Translational Animal Science* 4 (1): 423–441. doi:10.1093/tas/txz174.

Callister, Stephen J., Richard C. Barry, Joshua N. Adkins, Ethan T. Johnson, Wei Jun Qian et al. 2006. "Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics". *Journal of Proteome Research* 5 (2): 277–286. doi:10.1021/pr0503001.

Camire, Mary Ellen, Stan Kubow and Danielle J. Donnelly. 2009. "Potatoes and human health". *Critical Reviews in Food Science and Nutrition* 49 (10): 823–840. doi:10.1080/10408390903041996.

Castillejo, Mariá Ángeles, Sara Fondevilla-Aparicio, Carlos Fuentes-Almagro and Diego Rubiales. 2020. "Quantitative Analysis of Target Peptides Related to Resistance against Ascochyta Blight (Peyronellaea pinodes) in Pea". *Journal of Proteome Research* 19 (3): 1000–1012. doi:10.1021/acs.jproteome.9b00365.

Chang, Cheng, Kaikun Xu, Chaoping Guo, Jinxia Wang, Qi Yan et al. 2018. "PANDAview: An easy-to-use tool for statistical analysis and vis- ualization of quantitative proteomics data". *Bioinformatics* 34 (20): 3594–3596. doi:10.1093/bioinformatics/bty408.

Chang, Yu Wei, Inteaz Alli, Yasuo Konishi and Edmund Ziomek. 2011. "Characterization of protein fractions from chickpea (Cicer arietinum L.) and oat (Avena sativa L.) seeds using proteomic techniques". *Food Research International* 44 (9): 3094–3104. doi:10.1016/j.foodres.2011.08.001.

Chawade, Aakash, Erik Alexandersson, Therese Bengtsson, Erik Andreasson and Fredrik Levander. 2016. "Targeted Proteomics Approach for Precision Plant Breeding". *Journal of Proteome Research* 15 (2): 638–646. doi:10.1021/acs.jproteome.5b01061.

Chawade, Aakash, Erik Alexandersson and Fredrik Levander. 2014. "Normalyzer: A Tool for Rapid Evaluation of Normalisation Methods for Omics Data Sets". *Journal of Proteome Research* 13 (6): 3114–3120. doi:10.1021/pr401264n.

Chawade, Aakash, Marianne Sandin, Johan Teleman, Johan Malmström, Fredrik Levander et al. 2015. "Data processing has major impact on the outcome of quantitative label-free LC-MS analysis". *Journal of Proteome Research* 14 (2): 676–687. doi:10.1021/pr500665j.

Chen, Chao, Kay Grennan, Judith Badner, Dandan Zhang, Elliot Gershon et al. 2011. "Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods". *PLoS ONE* 6 (2): e17238. doi:10.1371/journal.pone.0017238.

Chen, Lingling, Quanzhu Chen, Lingqi Kong, Fangshan Xia, Huifang Yan et al. 2016. "Proteomic and Physiological Analysis of the Response of Oat (Avena sativa) Seeds to Heat Stress under Different Moisture Conditions". *Frontiers in Plant Science* 7:896. doi:10.3389/fpls.2016.00896.

Chi, Xu. 2020. *Identifying batch susceptible peptides in proteomics using machine learning*. Tech. rep. Lund University.

Chibon, Frederic. 2013. "Cancer gene expression signatures-The rise and fall?" *European Journal of Cancer* 49 (8): 2000–2009. doi:10.1016/j.ejca.2013.02.021.

Choi, Meena, Zeynep F. Eren-Dogu, Christopher Colangelo, John Cottrell, Michael R. Hoopmann et al. 2017. "ABRF Proteome Informatics Research Group (iPRG) 2015 Study: Detection of Differentially Abundant Proteins in Label-Free Quantitative LC-MS/MS Experiments". *Journal of Proteome Research* 16 (2): 945–957. doi:10.1021/acs.jproteome.6b00881.

Churchill, Gary A. 2002. "Fundamentals of experimental design for cdna microarrays". *Nature Genetics* 32:490–495. doi:10.1038/ng1031.

Collard, Bertrand C.Y., and David J. Mackill. 2008. "Marker-assisted selection: An approach for precision plant breeding in the twenty-first century". *Philosophical Transactions of the Royal Society* 363 (1491): 557–572. doi:10.1098/rstb.2007.2170.

Consoli, L., A. Lefèvre, M. Zivy, D. De Vienne and C. Damerval. 2002. "QTL analysis of proteome and transcriptome variations for dissecting the genetic architecture of complex traits in maize". *Plant Molecular Biology* 48 (5-6): 575–581. doi:10.1023/A:1014840810203.

Conway, Jake R., Alexander Lex and Nils Gehlenborg. 2017. "UpSetR: An R package for the visualization of intersecting sets and their properties". *Bioinformatics* 33 (18): 2938–2940. doi:10.1093/bioinformatics/btx364.

Cook, Tyler, Yinfa Ma and Sanjeewa Gamagedara. 2020. "Evaluation of statistical techniques to normalize mass spectrometry-based urinary metabolomics data". *Journal of Pharmaceutical and Biomedical Analysis* 177:112854. doi:10.1016/j.jpba.2019.112854.

Cox, Jürgen, Marco Y. Hein, Christian A. Luber, Igor Paron, Nagarjuna Nagaraj et al. 2014. "Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ". *Molecular and Cellular Proteomics* 13 (9): 2513–2526. doi:10.1074/mcp.M113.031591.

Cox, Jürgen, and Matthias Mann. 2008. "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification". *Nature Biotechnology* 26 (12): 1367–1372. doi:10.1038/nbt.1511.

Craig, Robertson, and Ronald C. Beavis. 2004. "TANDEM: Matching proteins with tandem mass spectra". *Bioinformatics* 20 (9): 1466–1467. doi:10.1093/bioinformatics/bth092.

Cuklina, Jelena, Patrick G.A. Pedrioli and Ruedi Aebersold. 2020. "Review of Batch Effects Prevention, Diagnostics, and Correction Approaches". In *Mass Spectrometry Data Analysis in Proteomics*, ed. by Matthiesen Rune. Springer. ISBN: 978-1-4939-9744-2. doi:10.1007/978-1-4939-9744-2_16.

Dale, Ryan, Björn Grüning, Andreas Sjödin, Brad A. Chapman, Jillian Rowe et al. 2018. "Bioconda: Sustainable and comprehensive software distribution for the life sciences". *Nature Methods* 15 (7): 475–476. doi:10.1038/s41592-018-0046-7.

Damerval, C., A. Maurice, J. M. Josse and D. De Vienne. 1994. "Quantitative trait loci underlying gene product variation: A novel perspective for analyzing regulation of genome expression". *Genetics* 137 (1): 289–301.

Daou, Cheickna, and Hui Zhang. 2012. "Oat Beta-Glucan: Its Role in Health Promotion and Prevention of Diseases". *Comprehensive Reviews in Food Science and Food Safety* 11 (4): 355–365. doi:10.1111/j.1541-4337.2012.00189.x.

Das, Aayudh, Bimal Paudel and Jai S. Rohila. 2015. "Potentials of Proteomics in Crop Breeding". In *Advances in Plant Breeding Strategies: Breeding, Biotechnology and Molecular Tools*, ed. by Jameel M. Al-Khayri, Shri Mohan Jain and Dennis V. Johnson, 513–538. Springer. doi:10.1007/978-3-319-22521-0.

De Vienne, Dominique, Agnès Leonardi, Catherine Damerval and Michel Zivy. 1999. "Genetics of proteome variation for QTL characterization: Application to drought-stress responses in maize". *Journal of Experimental Botany* 50 (332): 303–309. doi:10.1093/jxb/50.332.303.

Dempewolf, Hannes, Ruth J. Eastwood, Luigi Guarino, Colin K. Khoury, Jonas V. Müller et al. 2014. "Adapting Agriculture to Climate Change: A Global Initiative to Collect, Conserve, and Use Crop Wild Relatives". *Agroecology and Sustainable Food Systems* 38 (4): 369–377. doi:10.1080/21683565.2013.870629.

Desta, Zeratsion Abera, and Rodomiro Ortiz. 2014. "Genomic selection: Genome-wide prediction in plant improvement". *Trends in Plant Science* 19 (9): 592–601. doi:10.1016/j.tplants.2014.05.006.

Dittrich, Julia, Susen Becker, Max Hecht and Uta Ceglarek. 2015. "Sample preparation strategies for targeted proteomics via proteotypic peptides in human blood using liquid chromatography tandem mass spectrometry". *Proteomics Clinical Applications* 9:5–16. doi:10.1002/prca.201400121.

Diz, Angel P., Mónica Martínez-Fernández and Emilio Rolán-Alvarez. 2012. "Proteomics in evolutionary ecology: Linking the genotype with the phenotype". *Molecular Ecology* 21 (5): 1060–1080. doi:10.1111/j.1365-294X.2011.05426.x.

Dobbin, K., J. H. Shih and R. Simon. 2003. "Statistical design of reverse dye microarrays". *Bioinformatics* 19 (7): 803–810. doi:10.1093/bioinformatics/btg076.

Dozmorov, Mikhail G. 2018. "GitHub statistics as a measure of the impact of open-source bioinformatics software". *Frontiers in Bioengineering and Biotechnology* 6:1–4. doi:10.3389/fbioe.2018.00198.

Druart, Xavier, and Simon de Graaf. 2018. "Seminal plasma proteomes and sperm fertility". *Animal Reproduction Science* 194:33–40. doi:10.1016/j.anireprosci.2018.04.061.

Druart, Xavier, Jessica P. Rickard, Guillaume Tsikis and Simon P. de Graaf. 2019. "Seminal plasma proteins as markers of sperm fertility". *Theriogenology* 137:30–35. doi:10.1016/j.theriogenology.2019.05.034.

EFSA. 2010. "Scientific Opinion on the substantiation of a health claim related to oat beta-glucan and lowering blood cholesterol and reduced risk of (coronary) heart disease pursuant to Article 14 of Regulation (EC) No 1924/20061". *EFSA Journal* 8 (12): 1885–2000. doi:10.2903/j.efsa.2010.1885..

Eldakak, Moustafa, Sanaa I.M. Milad, Ali I. Nawar and Jai S. Rohila. 2013. "Proteomics: A biotechnology tool for crop improvement". *Frontiers in Plant Science* 4:35. doi:10.3389/fpls.2013.00035.

Elo, Laura L., Sanna Filén, Riitta Lahesmaa and Tero Aittokallio. 2008. "Reproducibility-optimized test statistic for ranking genes in microarray studies". *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 5 (3): 423–431. doi:10.1109/tcbb.2007.1078.

Eng, Jimmy K., Ashley L. McCormack and John R. Yates III. 1994. "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database". *Journal of the American Society for Mass Spectrometry* 5 (11): 976–989. doi:10.1016/1044-0305(94)80016-2.

Escrivá, L., G. Font and L. Manyes. 2015. "In vivo toxicity studies of fusarium mycotoxins in the last decade: A review". *Food and Chemical Toxicology* 78:185–206. doi:10.1016/j.fct.2015.02.005.

Fenn, John B, Matrhias Mann, Chin K A I Meng, Shek Fu Wong and Craig M Whitehouse. 1989. "Electrospray Ionization for Mass Spectrometry of Large Biomolecules". *Science* 246 (4926): 64–71. doi:10.1126/science.2675315.

Fu, Qin, Michael P. Kowalski, Mitra Mastali, Sarah J. Parker, Kimia Sobhani et al. 2018. "Highly Reproducible Automated Proteomics Sample Preparation Workflow for Quantitative Mass Spectrometry". *Journal of Proteome Research* 17:420–428. doi:10.1021/acs.jproteome.7b00623.

Gagnon-Bartsch, Johann A., and Terence P. Speed. 2012. "Using control genes to correct for unwanted variation in microarray data". *Biostatistics* 13 (3): 539–552. doi:10.1093/biostatistics/kxr034.

Gentleman, Robert C, Vincent J Carey, Douglas M Bates, Benjamin M Bolstad, Marcel Dettling et al. 2004. "Bioconductor: open software development for computational biology and bioinformatics." *Genome biology* 5 (10): 1–16. doi:10.1186/gb-2004-5-10-r80.

Gerber, Scott A., John Rush, Olaf Stemman, Marc W. Kirschner and Steven P. Gygi. 2003. "Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS". *PNAS* 100 (12): 6940–6945. doi:10.1073/pnas.0832254100.

Gessulat, Siegfried, Tobias Schmidt, Daniel Paul Zolg, Patroklos Samaras, Karsten Schnatbaum et al. 2019. "Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning". *Nature Methods* 16 (6): 509–518. doi:10.1038/s41592-019-0426-7.

Gilad, Yoav, and Orna Mizrahi-man. 2015. "A reanalysis of mouse ENCODE comparative gene expression data". *F1000Research* 4 (121): 1–38. doi:10.12688/f1000research.6536.1.

Gilad, Yoav, Scott A. Rifkin and Jonathan K. Pritchard. 2008. "Revealing the architecture of gene regulation: the promise of eQTL studies". *Trends in Genetics* 24 (8): 408–415. doi:10.1016/j.tig.2008.06.001.

Gillet, Ludovic C, Pedro Navarro, Stephen Tate, Hannes Röst, Nathalie Selevsek et al. 2012. "Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis". *Molecular and Cellular Proteomics* 11 (6): O111.016717. doi:10.1074/mcp.O111.016717.

Goble, Carole. 2014. "Better software, better research". *IEEE Internet Computing* 18 (5): 4–8. doi:10.1109/MIC.2014.88. https://www.software.ac.uk/resources/publications/better-software-better-research.

Goeminne, Ludger J.E., Andrea Argentini, Lennart Martens and Lieven Clement. 2015. "Summarization vs peptide-based models in label-free quantitative proteomics: Performance, pitfalls, and data analysis guidelines". *Journal of Proteome Research* 14 (6): 2457–2465. doi:10.1021/pr501223t.

Goh, Wilson Wen Bin, Wei Wang and Limsoon Wong. 2017. "Why Batch Effects Matter in Omics Data, and How to Avoid Them". *Trends in Biotechnology* 35 (6): 498–507. doi:10.1016/j.tibtech.2017.02.012.

Gomes, Fabio P, Robin Park, Arabela G Viana, Carolina Fernandez Costa, Einko Topper et al. 2020. "Protein signatures of seminal plasma from bulls with contrasting frozen-thawed sperm viability". *Scientific Reports* 10:14661. doi:10.1038/s41598-020-71015-9.

Gorash, A., R. Armonienė, J. Mitchell Fetch, Liatukas and V. Danytė. 2017. "Aspects in oat breeding: nutrition quality, nakedness and disease resistance, challenges and perspectives". *Annals of Applied Biology* 171 (3): 281–302. doi:10.1111/aab.12375.

Gregori, Josep, Laura Villarreal, Olga Méndez, Alex Sánchez, José Baselga et al. 2012. "Batch effects correction improves the sensitivity of significance tests in spectral counting-based comparative discovery proteomics". *Journal of Proteomics* 75 (13): 3938–3951. doi:10.1016/j.jprot.2012.05.005.

Grüning, Björn A., Samuel Lampa, Marc Vaudel and Daniel Blankenberg. 2019. "Software engineering for scientific big data analysis". *GigaScience* 8 (5): 1–6. doi:10.1093/gigascience/giz054.

Gunnaiah, Raghavendra, Ajjamada C. Kushalappa, Raj Duggavathi, Stephen Fox and Daryl J. Somers. 2012. "Integrated metabolo-proteomic approach to decipher the mechanisms by which wheat qtl (Fhb1) contributes to resistance against Fusarium graminearum". *PLoS ONE* 7 (7): e40695. doi:10.1371/journal.pone.0040695.

Gygi, S. P., B. Rist, A. Gerber, F. Turecek, M.H. Gelb et al. 1999. "Access Quantitative analysis of complex protein mixtures using isotope-coded affinity tags Nature Biotechnology". *Nature* 17:994–999. doi:10.1038/13690.

Häkkinen, Jari, Gregory Vincic, Olle Månsson, Kristofer Wårell and Fredrik Levander. 2009. "The Proteios Software Environment: An extensible multiuser platform for management and analysis of proteomics data". *Journal of Proteome Research* 8 (6): 3037–3043. doi:10.1021/pr900189c.

Haverkort, A. J., and A. Verhagen. 2008. "Climate change and its repercussions for the potato supply chain". *Potato Research* 51:223–237. doi:10.1007/s11540-008-9107-0.

He, Xinyao, Helge Skinnes, Rebekah E. Oliver, Eric W. Jackson and Åsmund Bjørnstad. 2013. "Linkage mapping and identification of QTL affecting deoxynivalenol (DON) content (Fusarium resistance) in oats (Avena sativa L.)." *Theoretical and applied genetics* 126 (10): 2655–2670. doi:10.1007/s00122-013-2163-0.

Hellin, Jon, Bekele Shiferaw, Jill E Cairns, M P Reynolds, Ivan Ortiz-monasterio et al. 2012. "Climate Change and Food Security in the Developing World: Potential of Maize and Wheat Research to Expand Options for Adaptation and Mitigation". *Journal of Development and Agricultural Economics* 4 (12): 311–321. doi:10.5897/JDAE11.112.

Hicks, Stephanie C., and Rafael A. Irizarry. 2015. "quantro: A data-driven approach to guide the choice of an appropriate normalization method". *Genome Biology* 16 (1): 1–8. doi:10.1186/s13059-015-0679-0.

Hicks, Stephanie C., Kwame Okrah, Joseph N. Paulson, John Quackenbush, Rafael A. Irizarry et al. 2018. "Smooth quantile normalization". *Biostatistics* 19 (2): 185–198. doi:10.1093/biostatistics/kxx028.

Hijmans, Robert J. 2003. "The effect of climate change on global potato production". *American journal of potato research* 80 (4): 271–280. doi:10.1016/S0308-521X(02)00081-1.

Hilary, S. Parker, and T. Leek Jeffrey. 2012. "The practical effect of batch on genomic prediction Hilary". *Statistical Applications in Genetics and Molecular Biology* 11 (3): 1–20. doi:10.1038/jid.2014.371.

Holloway, Beth, and Bailin Li. 2010. "Expression QTLs: Applications for crop improvement". *Molecular Breeding* 26 (3): 381–391. doi:10.1007/s11032-010-9396-2.

Hu, Jianhua, Kevin R. Coombes, Jeffrey S. Morris and Keith A. Baggerly. 2005a. "The importance of experimental design in proteomic mass spectrometry experiments: Some cautionary tales". *Briefings in Functional Genomics and Proteomics* 3 (4): 322–331. doi:10.1093/bfgp/3.4.322.

Hu, Qizhi, Robert J. Noll, Hongyan Li, Alexander Makarov, Mark Hardman et al. 2005b. "The Orbitrap: A new mass spectrometer". *Journal of Mass Spectrometry* 40:430–443. doi:10.1002/jms.856.

Huang, Ting, Jingjing Wang, Weichuan Yu and Zengyou He. 2012. "Protein inference: a review." *Briefings in bioinformatics* 13 (5): 586–614. doi:10.1093/bib/bbs004. http://www.ncbi.nlm.nih.gov/pubmed/22373723.

Huber, W., A. von Heydebreck, H. Sultmann, A. Poustka and M. Vingron. 2002. "Variance stabilization applied to microarray data calibration and to the quantification of differential expression". *Bioinformatics* 18 (Suppl 1): S96–S104. doi:10.1093/bioinformatics/18.suppl_1.S96.

Humblot, P., G. Decoux and T. Dhorne. 1991. "Effects of the Sire and District of A1 on Cow Fertility". *Reproduction in Domestic Animals* 26:225–234. doi:10.1111/j.1439-0531.1991.tb01533.x.

Ibrahim, M. N.G., and I. S. Selezneva. 2017. "β-glucan extract from oat bran and its industrial importance". *AIP Conference Proceedings* 1886:020100. doi:10.1063/1.5002997.

Irizarry, Rafael A., Daniel Warren, Forrest Spencer, Irene F. Kim, Shyam Biswal et al. 2005. "Multiple-laboratory comparison of microarray platforms". *Nature Methods* 2 (5): 345–349. doi:10.1038/nmeth756.

Jiang, Guo-liang. 2013a. "Advances in Crop Science and Technology Plant Marker-Assisted Breeding and Conventional Breeding: Challenges and Perspectives". *Advances in Crop Science and Technology* 1 (3): 1–2. doi:10.4172/2329-8863.

Jiang, Guo-Liang. 2013b. "Molecular Markers and Marker-Assisted Breeding in Plants". In *Plant Breeding from Laboratories to Fields*, ed. by Sven Bode Andersen. IntechOpen. doi:10.5772/52583.

Jiménez, Rafael C., Mateusz Kuzak, Monther Alhamdoosh, Michelle Barker, Bérénice Batut et al. 2017. "Four simple recommendations to encourage best practices in research software". *F1000Research* 6:876. doi:10.12688/f1000research.11407.1.

Johnson, W. Evan, Cheng Li and Ariel Rabinovic. 2007. "Adjusting batch effects in microarray expression data using empirical Bayes methods". *Biostatistics* 8 (1): 118–127. doi:10.1093/biostatistics/kxj037.

Jolliffe, I.T. 2002. *Principal Components Analysis*, 374–377. Springer. ISBN: 0-387-95442-2. doi:10.1016/B978-0-08-044894-7.01358-0.

Käll, Lukas, Jesse D. Canterbury, Jason Weston, William Stafford Noble and Michael J. MacCoss. 2007. "Semi-supervised learning for peptide identification from shotgun proteomics datasets". *Nature Methods* 4 (11): 923–925. doi:10.1038/nmeth1113.

Kammers, Kai, Robert N. Cole, Calvin Tiengwe and Ingo Ruczinski. 2015. "Detecting significant changes in protein abundance". *EuPA Open Proteomics* 7:11–19. doi:10.1016/j.euprot.2015.02.002.

Karimzadeh, Mehran, and Michael M. Hoffman. 2018. "Top considerations for creating bioinformatics software documentation". *Briefings in bioinformatics* 19 (4): 693–699. doi:10.1093/bib/bbw134.

Karpievitch, Yuliya V, Alan R Dabney and Richard D Smith. 2012. "Normalization and missing value imputation for label-free LC-MS analysis". *BMC Bioinformatics* 13 (Suppl 16): S5. doi:10.1186/1471-2105-13-S16-S5.

Karpievitch, Yuliya V., Sonja B. Nikolic, Richard Wilson, James E. Sharman and Lindsay M. Edwards. 2014. "Metabolomics data normalization with EigenMS". *PLoS ONE* 9 (12): 1–10. doi:10.1371/journal.pone.0116221.

Karpievitch, Yuliya V., Thomas Taverner, Joshua N. Adkins, Stephen J. Callister, Gordon A. Anderson et al. 2009. "Normalization of peak intensities in bottom-up MS-based proteomics using singular value decomposition". *Bioinformatics* 25 (19): 2573–2580. doi:10.1093/bioinformatics/btp426.

Kim, Sangtae, and Pavel A. Pevzner. 2014. "MS-GF+ makes progress towards a universal database search tool for proteomics". *Nature Communications* 5:5277. doi:10.1038/ncomms6277.

Krüger, Thomas, Thomas Lehmann and Heidrun Rhode. 2013. "Effect of quality characteristics of single sample preparation steps in the precision and coverage of proteomic studies-A review". *Analytica Chimica Acta* 776:1–10. doi:10.1016/j.aca.2013.01.020.

Kulak, Nils A., Garwin Pichler, Igor Paron, Nagarjuna Nagaraj and Matthias Mann. 2014. "Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells". *Nature Methods* 11 (3): 319–324. doi:10.1038/nmeth.2834.

Kuligowski, J., D. Pérez-Guaita, I. Lliso, J. Escobar, Z. León et al. 2014. "Detection of batch effects in liquid chromatography-mass spectrometry metabolomic data using guided principal component analysis". *Talanta* 130:442–448. doi:10.1016/j.talanta.2014.07.031.

Kultima, Kim, Anna Nilsson, Birger Scholz, Uwe L. Rossbach, Maria Fälth et al. 2009. "Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides." *Molecular And Cellular proteomics* 8 (10): 2285–2295. doi:10.1074/mcp.M800514-MCP200.

Kurtzer, Gregory M., Vanessa Sochat and Michael W. Bauer. 2017. "Singularity: Scientific containers for mobility of compute". *PLoS ONE* 12 (5): 1–20. doi:10.1371/journal.pone.0177459.

Langridge, Peter, and Delphine Fleury. 2011. "Making the most of 'omics' for crop breeding". *Trends in Biotechnology* 29 (1): 33–40. doi:10.1016/j.tibtech.2010.09.006.

Lawlor, Brendan, and Paul Walsh. 2015. "Engineering bioinformatics: Building reliability, performance and productivity into bioinformatics software". *Bioengineered* 6 (4): 193–203. doi:10.1080/21655979.2015.1050162.

Lazar, Cosmin, Laurent Gatto, Myriam Ferro, Christophe Bruley and Thomas Burger. 2016. "Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies". *Journal of Proteome Research* 15 (4): 1116–1125. doi:10.1021/acs.jproteome.5b00981.

Leahy, Tamara, Jessica P. Rickard, Taylor Pini, Bart M. Gadella, Simon P Graaf et al. 2020. "Quantitative Proteomic Analysis of Seminal Plasma, Sperm Membrane Proteins and Seminal Extracellular Vesicles Suggests Vesicular Mechanisms Aid in the Removal and Addition of Proteins to the Ram Sperm Membrane". *Proteomics* 1900289 (12): 1–15. doi:10.1002/pmic.201900289.

Leek, Jeffrey T., Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead et al. 2010. "Tackling the widespread and critical impact of batch effects in high-throughput data". *Nature Reviews Genetics* 11 (10): 733–739. doi:10.1038/nrg2825.

Leek, Jeffrey T., and John D. Storey. 2007. "Capturing heterogeneity in gene expression studies by surrogate variable analysis". *PLoS Genetics* 3 (9): e161. doi:10.1371/journal.pgen.0030161.

Li, Bo, Jing Tang, Qingxia Yang, Shuang Li, Xuejiao Cui et al. 2017. "NOREVA: Normalization and evaluation of MS-based metabolomics data". *Nucleic Acids Research* 45:W162–W170. doi:10.1093/nar/gkx449.

Li, Wentian. 2012. "Volcano plots in analyzing differential expressions with mRNA microarrays". *Journal of Bioinformatics and Computational Biology* 10 (6): 1–24. doi:10.1142/S0219720012310038.

Lin, Yang, Belaghihalli N. Gnanesh, James Chong, Gang Chen, Aaron D. Beattie et al. 2014. "A major quantitative trait locus conferring adult plant partial resistance to crown rust in oat". *BMC Plant Biology* 14 (1): 1–11. doi:10.1186/s12870-014-0250-2.

Lindh, Victor. 2020. *NIB - A visualization tool for feature-level multi-omic data based on global metrics*. Tech. rep. http://lup.lub.lu.se/student-papers/record/9018460.

Liu, Yansheng, Yang Mi, Torsten Mueller, Saskia Kreibich, Evan G. Williams et al. 2019. "Multi-omic measurements of heterogeneity in HeLa cells across laboratories". *Nature Biotechnology* 37 (3): 314–322. doi:10.1038/s41587-019-0037-y.

Lobell, David B., Marshall B. Burke, Claudia Tebaldi, Michael D. Mastrandrea, Walter P. Falcon et al. 2008. "Prioritizing Climate Change Adaptation Needs for Food Security in 2030". *Science* 319:607–610. doi:10.1126/science.1152339.

Luo, J., M. Schumacher, A. Scherer, D. Sanoudou, D. Megherbi et al. 2010. "A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data". *The Pharmacogenomics Journal* 10 (4): 278–291. doi:10.1038/tpj.2010.57.

Lyutvinskiy, Yaroslav, Hongqian Yang, Dorothea Rutishauser and Roman A. Zubarev. 2013. "In Silico Instrumental Response Correction Improves Precision of Label-free Proteomics and Accuracy of Proteomics-based Predictive Models". *Molecular and Cellular Proteomics* 12 (8): 2324–2331. doi:10.1074/mcp.O112.023804.

Ma, Nyuk Ling, Zaidah Rahmat and Su Shiung Lam. 2013. "A review of the 'omics' approach to biomarkers of oxidative stress in Oryza sativa". *International Journal of Molecular Sciences* 14:7515–7541. doi:10.3390/ijms14047515.

Maier, Tobias, Marc Güell and Luis Serrano. 2009. "Correlation of mRNA and protein in complex biological samples". *FEBS Letters* 583 (24): 3966–3973. doi:10.1016/j.febslet.2009.10.036.

Mangul, Serghei, Lana S. Martin, Eleazar Eskin and Ran Blekhman. 2019. "Improving the usability and archival stability of bioinformatics software". *Genome Biology* 20:47. doi:10.1186/s13059-019-1649-8.

Manimaran, Solaiappan, Heather Marie Selby, Kwame Okrah, Claire Ruberman, Jeffrey T. Leek et al. 2016. "BatchQC: interactive software for evaluating sample and batch effects in genomic data". *Bioinformatics* 32 (24): 3836–3838. doi:10.1093/bioinformatics/btw538.

Martinelli, José Antônio, Márcia Soares Chaves, Felipe André Sganzerla Graichen, Luiz Carlos Federizzi and Luiz Felipe Dresch. 2014. "Impact of Fusarium Head Blight in Reducing the Weight of Oat Grains". *Journal of Agricultural Science* 6 (5): 188–198. doi:10.5539/jas.v6n5p188.

Martínez-Villaluenga, Cristina, and Elena Peñas. 2017. "Health benefits of oat: current evidence and molecular mechanisms". *Current Opinion in Food Science* 14:26–31. doi:10.1016/j.cofs.2017.01.004.

Marx, Vivien. 2019. "A dream of single-cell proteomics". *Nature Methods* 16 (9): 809–812. doi:10.1038/s41592-019-0540-6.

– . 2020. "When computational pipelines go clank". *Nature Methods* 17:659–662. doi:10.1038/s41592-020-0886-9.

Maughan, Peter J., Rebekah Lee, Rachel Walstead, Robert J. Vickerstaff, Melissa C. Fogarty et al. 2019. "Genomic insights from the first chromosome-scale assemblies of oat (Avena spp.) diploid species". *BMC Biology* 17 (1): 92. doi:10.1186/s12915-019-0712-y.

McShane, L. M. 2017. "In Pursuit of Greater Reproducibility and Credibility of Early Clinical Biomarker Research". *Clinical and Translational Science* 10 (2): 58–60. doi:10.1111/cts.12449.

Mertens, Bart J. A. 2017. "Transformation, Normalization, and Batch Effect in the Analysis ofMass Spectrometry Data for Omics Studies". In *Statistical Analysis of Proteomics, Metabolomics, and Lipidomics Data Using Mass Spectrometry*, ed. by Susmita Datta and Bart J. A. Mertens, 1–21. Springer. ISBN: 978-3-319-45809-0. doi:10.1007/978-3-319-45809-0.

Meuwissen, Theo. 2007. "Genomic selection : Marker assisted selection on a genome wide scale". *Journal of Animal Breeding and Genetics* 124 (6): 321–322. doi:10.1111/j.1439-0388.2007.00708.x.

Møller, Line Kofod. 2017. *A novel pipeline for protein level quantification using peptide mass spectrometry data*. Tech. rep. Lund University. http://lup.lub.lu.se/student-papers/record/8929504.

Nadeem, Muhammad Azhar, Muhammad Amjad Nawaz, Muhammad Qasim Shahid, Yıldız Doğan, Gonul Comertpay et al. 2018. "DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing". *Biotechnology and Biotechnological Equipment* 32 (2): 261–285. doi:10.1080/13102818.2017.1400401.

Nagaraj, Shivashankar H., Nicola Waddell, Anil K. Madugundu, Scott Wood, Alun Jones et al. 2015. "PGTools: A software suite for proteogenomic data analysis and visualization". *Journal of Proteome Research* 14 (5): 2255–2266. doi:10.1021/acs.jproteome.5b00029.

Nakaya, Akihiro, and Sachiko N. Isobe. 2012. "Will genomic selection be a practical method for plant breeding?" *Annals of Botany* 110 (6): 1303–1316. doi:10.1093/aob/mcs109.

Navarro, Pedro, Jörg Kuharev, Ludovic C. Gillet, Oliver M. Bernhardt, Hannes L. Röst et al. 2016. "A multicenter study benchmarks software tools for label-free proteome quantification". *Nature Biotechnology* 34 (11): 1130–1136. doi:10.1038/nbt.3685.A.

Nesvizhskii, Alexey I., and Ruedi Aebersold. 2005. "Interpretation of shotgun proteomic data: The protein inference problem". *Molecular and Cellular Proteomics* 4 (10): 1419–1440. doi:10.1074/mcp.R500012-MCP200.

Nie, Lei, Gang Wu, David E. Culley, Johannes C.M. Scholten and Weiwen Zhang. 2007. "Integrative analysis of transcriptomic and proteomic data: Challenges, solutions and applications". *Critical Reviews in Biotechnology* 27 (2): 63–75. doi:10.1080/07388550701334212.

Nygaard, Vegard, Einar Andreas Rødland and Eivind Hovig. 2016. "Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses". *Biostatistics* 17 (1): 29–39. doi:10.1093/biostatistics/kxv027.

Oberg, Ann L., and Olga Vitek. 2009. "Statistical design of quantitative mass spectrometry-based proteomic experiments". *Journal of Proteome Research* 8:2144–2156. doi:10.1021/pr8010099.

Ong, Shao En, Blagoy Blagoev, Irina Kratchmarova, Dan Bach Kristensen, Hanno Steen et al. 2002. "Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics." *Molecular and Cellular Proteomics* 1 (5): 376–386. doi:10.1074/mcp.M200025-MCP200.

Ooijen, Michiel P. van, Victor L. Jong, Marinus J.C. Eijkemans, Albert J.R. Heck, Arno C. Andeweg et al. 2017. "Identification of differentially expressed peptides in high-throughput proteomics data". *Briefings in Bioinformatics*, no. June: 1–11. doi:10.1093/bib/bbx031.

Papiez, Anna, Michal Marczyk, Joanna Polanska and Andrzej Polanski. 2019. "BatchI: Batch effect Identification in high-throughput screening data using a dynamic programming algorithm". *Bioinformatics* 35 (11): 1885–1892. doi:10.1093/bioinformatics/bty900.

PepsiCo. 2020. *PepsiCo OT3098 Hexaploid Oat Genome Assembly and Annotation Release in collaboration with GrainGenes*. Visited on 23/06/2020. https://wheat.pw.usda.gov/GG3/node/922.

Perez-Riverol, Yasset. 2020. "Toward a Sample Metadata Standard in Public Proteomics Repositories". *Journal of proteome research* 19 (10): 3906–3909. doi:10.1021/acs.jproteome.0c00376.

Perez-Riverol, Yasset, Attila Csordas, Jingwen Bai, Manuel Bernal-Llinares, Suresh Hewapathirana et al. 2018. "The PRIDE database and related tools and resources in 2019: improving support for quantification data". *Nucleic Acids Research* 47 (D1): D442–D450. doi:10.1093/nar/gky1106.

Perkins, David N, Darryl J C Pappin, David M Creasy and John S Cottrell. 1999. "Probability-based protein identification by searching sequence databases using mass spectrometry data Proteomics and 2-DE". *Electrophoresis* 20 (18): 3551–3567.

Picott, Paula, Ruedi Aebersold and Bruno Domont. 2007. "The implications of proteolytic background for shotgun proteomics". *Molecular and Cellular Proteomics* 6:1589–1598. doi:10.1074/mcp.M700029-MCP200.

Piehowski, Paul D, Vladislav A Petyuk, Daniel J Orton, Fang Xie, Manuel Ramirez et al. 2013. "Sources of Technical Variability in Quantitative LC-MS Proteomics: Human Brain Tissue Sample Analysis". *Journal of Proteome Research* 12 (5): 2128–2137. doi:10.1021/pr301146m.

Pini, Taylor, Tamara Leahy, Clement Soleilhavoup, Guillaume Tsikis, Valerie Labas et al. 2016. "Proteomic Investigation of Ram Spermatozoa and the Proteins Conferred by Seminal Plasma". *Journal of Proteome Research* 15 (10): 3700–3711. doi:10.1021/acs.jproteome.6b00530.

Polpitiya, Ashoka D., Wei Jun Qian, Navdeep Jaitly, Vladislav A. Petyuk, Joshua N. Adkins et al. 2008. "DAnTE: A statistical tool for quantitative analysis of -omics data". *Bioinformatics* 24 (13): 1556–1558. doi:10.1093/bioinformatics/btn217.

Pursiheimo, Anna, Anni P. Vehmas, Saira Afzal, Tomi Suomi, Thaman Chand et al. 2015. "Optimization of Statistical Methods Impact on Quantitative Proteomics Data". *Journal of Proteome Research* 14 (10): 4118–4126. doi:10.1021/acs.jproteome.5b00183.

Purvine, Samuel, Jason Thomas Eppel, Eugene C. Yi and David R. Goodlett. 2003. "Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer". *Proteomics* 3:847–850. doi:10.1002/pmic.200300362.

Rajnincová, Dana, Zdenka Gálová and Milan Chňapek. 2019. "Comparison of selected wheat, oat and buckwheat genotypes on proteomic level". *Journal of Central European Agriculture* 20 (3): 891–899. doi:10.5513/JCEA01/20.3.2293.

Ransohoff, David F. 2005. "Bias as a threat to the validity of cancer molecular-marker research". *Nature Reviews Cancer* 5 (2): 142–149. doi:10.1038/nrc1550.

Reinhart, Alex. 2015. *Statistics done wrong - The Woefully Complete Guide*. San Francisco: No Starch Press. ISBN: 9781593276201.

Rickard, J. P., T. Leahy, C. Soleilhavoup, G. Tsikis, V. Labas et al. 2015. "The identification of proteomic markers of sperm freezing resilience in ram seminal plasma". *Journal of Proteomics* 126:303–311. doi:10.1016/j.jprot.2015.05.017.

Rickard, J. P., T. Pini, C. Soleilhavoup, J. Cognie, R. Bathgate et al. 2014. "Seminal plasma aids the survival and cervical transit of epididymal ram spermatozoa". *Reproduction* 148 (5): 469–478. doi:10.1530/REP-14-0285.

Rifai, Nader, Michael A. Gillette and Steven A. Carr. 2006. "Protein biomarker discovery and validation: The long and uncertain path to clinical utility". *Nature Biotechnology* 24 (8): 971–983. doi:10.1038/nbt1235.

Rigbolt, Kristoffer T. G., Jens T. Vanselow and Blagoy Blagoev. 2011. "GProX, a User-Friendly Platform for Bioinformatics Analysis and Visualization of Quantitative Proteomics Data". *Molecular and Cellular Proteomics* 10 (8): O110.007450. doi:10.1074/mcp.O110.007450.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law et al. 2015. "limma powers differential expression analyses for RNA-sequencing and microarray studies". *Nucleic acids research* 43 (7): e47. doi:10.1093/nar/gkv007.

Robertson, S. A. 2007. "Seminal fluid signaling in the female reproductive tract: Lessons from rodents and pigs". *Journal of animal science* 85:36–44. doi:10.2527/jas.2006-578.

Rodziewicz, Paweł, Klaudia Chmielewska, Aneta Sawikowska, Łukasz Marczak, Magdalena Łuczak et al. 2019. "Identification of drought responsive proteins and related proteomic QTLs in barley". *Journal of Experimental Botany* 70 (10): 2823–2837. doi:10.1093/jxb/erz075.

Röst, Hannes L., George Rosenberger, Pedro Navarro, Ludovic Gillet, Saša M. Miladinoviä et al. 2014. "OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data". *Nature Biotechnology* 32 (3): 219–223. doi:10.1038/nbt.2841.

Röst, Hannes L., Timo Sachsenberg, Stephan Aiche, Chris Bielow, Hendrik Weisser et al. 2016. "OpenMS: A flexible open-source software platform for mass spectrometry data analysis". *Nature Methods* 13 (9): 741–748. doi:10.1038/nmeth.3959.

Rourke, Matthew B O, Stephanie E L Town, Penelope V Dalla, Fiona Bicknell, Naomi Koh Belic et al. 2019. "What is Normalization? The Strategies Employed in Top-Down and Bottom-Up Proteome Analysis Workflow". *Proteomes* 7 (29): 1–19. doi:10.3390/proteomes7030029.

Ruane, Alex C., John Antle, Joshua Elliott, Christian Folberth, Gerrit Hoogenboom et al. 2018. "Biophysical and economic implications for agriculture of +1.5 and +2.0C global warming using AgMIP Coordinated Global and Regional Assessments". *Climate Research* 76 (1): 17–39. doi:10.3354/cr01520.

Sabel, Michael S., Yashu Liu and David M. Lubman. 2011. "Proteomics in Melanoma Biomarker Discovery: Great Potential, Many Obstacles". *International Journal of Proteomics* 2011:1–8. doi:10.1155/2011/181890.

Salekdeh, Ghasem Hosseini, and Setsuko Komatsu. 2007. "Crop proteomics: Aim at sustainable agriculture of tomorrow". *Proteomics* 7 (16): 2976–2996. doi:10.1002/pmic.200700181.

Sandin, Marianne, Ashfaq Ali, Karin Hansson, Olle Månsson and Erik Andreasson. 2013. "An Adaptive Alignment Algorithm for Quality-controlled Label-free LC-MS". *Mol Cell Proteomics* 12 (5): 1407–1420. doi:10.1074/mcp.O112.021907.

Sandin, Marianne, Aakash Chawade and Fredrik Levander. 2015. "Is label-free LC-MS/MS ready for biomarker discovery?" *Proteomics - Clinical Applications* 9:289–294. doi:10.1002/prca.201400202.

Scherer, Andreas. 2009. *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*, ed. by Andreas Scherer, 1–260. John Wiley / Sons. ISBN: 9780470685983. doi:10.1002/9780470685983.

– . 2017. "Reproducibility in biomarker research and clinical development: A global challenge". *Biomarkers in Medicine* 11 (4): 309–312. doi:10.2217/bmm-2017-0024.

Scholtz, M. M., A. Maiwashe, F. W.C. Neser, A. Theunissen, W. J. Olivier et al. 2013. "Livestock breeding for sustainability to mitigate global warming, with the emphasis on developing countries". *South African Journal of Animal Sciences* 43 (3): 269–281. doi:10.4314/sajas.v43i3.4.

Schurch, Nicholas J, Pietá Schofield, Marek Gierliński, Christian Cole, Alexander Sherstnev et al. 2016. "How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?" *RNA* 22:839–851. doi:10.1261/rna.053959.115.

Schwämmle, Veit, Christina E. Hagensen, Adelina Rogowska-Wrzesinska and Ole N. Jensen. 2020. "PolySTest: Robust statistical testing of proteomics data with missing values improves detection of biologically relevant features". *Molecular and Cellular Proteomics* 19 (8): 1396–1408. doi:10.1074/mcp.ra119.001777.

Scott, Aaron. 2019. "GhostMS: An error-controlled machine learning approach to efficient alignment and quantification of multi-sample experiments in Mass Spectrometry-based Proteomics". PhD thesis, Lund University.

Searle, Brian C., Lindsay K. Pino, Jarrett D. Egertson, Ying S. Ting, Robert T. Lawrence et al. 2018. "Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry". *Nature Communications* 9:5128. doi:10.1038/s41467-018-07454-w.

Shah, Anup D, Robert J A Goode, Cheng Huang, David R Powell and Ralf B Schittenhelm. 2019. "LFQ-Analyst: An easy-to-use interactive web-platform to analyze and visualize label-free proteomics data preprocessed with MaxQuant." *Journal of proteome research* 19:204–211. doi:10.1021/acs.jproteome.9b00496.

Shteynberg, D., A. I. Nesvizhskii, R. L. Moritz and E. W. Deutsch. 2013. "Combining Results of Multiple Search Engines in Proteomics". *Molecular and Cellular Proteomics* 12 (9): 2383–2393. doi:10.1074/mcp.R113.027797.

Simon, Richard, Michael D. Radmacher and Kevin Dobbin. 2002. "Design of studies using DNA microarrays". *Genetic Epidemiology* 23:21–36. doi:10.1002/gepi.202.

Simpson, Richard.J. 2003. *Proteins and Proteomics A laboratory manual*. Ed. by Richard.J. Simpson. Cold Spring Harbor Laboratory Press. ISBN: 0879695544.

Smith, Lloyd M., and Neil L. Kelleher. 2013. "Proteoform: A single term describing protein complexity". *Nature Methods* 10 (3): 186–187. doi:10.1038/nmeth.2369.

Smith, Rob, Dan Ventura and John T. Prince. 2013. "LC-MS alignment in theory and practice: A comprehensive algorithmic review". *Briefings in Bioinformatics* 16 (1): 104–117. doi:10.1093/bib/bbt080.

Stott, G. H. 1961. "Female and Breed Associated with Seasonal Fertility Variation in Dairy Cattle". *Journal of Dairy Science* 44 (9): 1698–1704. doi:10.3168/jds.S0022-0302(61)89942-6.

Su, Jiangshuo, Jiafu Jiang, Fei Zhang, Ye Liu, Lian Ding et al. 2019. "Current achievements and future prospects in the genetic breeding of chrysanthemum: a review". *Horticulture Research* 6:109. doi:10.1038/s41438-019-0193-8.

Suomi, Tomi, Fatemeh Seyednasrollah, Maria K. Jaakkola, Thomas Faux and Laura L. Elo. 2017. "ROTS: An R package for reproducibility-optimized statistical testing". *PLoS Computational Biology* 13 (5): 1–10. doi:10.1371/journal.pcbi.1005562.

Suresh, K. 2011. "An overview of randomization techniques: An unbiased assessment of outcome in clinical research". *Journal of Human Reproductive Sciences* 4 (1): 8–11. doi:10.4103/0974-1208.82352.

Tan, Boon Chin, Yin Sze Lim and Su Ee Lau. 2017. "Proteomics in commercial crops: An overview". *Journal of Proteomics* 169:176–188. doi:10.1016/j.jprot.2017.05.018.

Tang, Jing, Yunxia Wang, Yongchao Luo, Jianbo Fu, Yang Zhang et al. 2020. "Computational advances of tumor marker selection and sample classification in cancer proteomics". *Computational and Structural Biotechnology Journal* 18 (2020): 2012–2025. doi:10.1016/j.csbj.2020.07.009.

Tekauz, A., B. McCallum, N. Ames and J. Mitchell Fetch. 2004. "Fusarium head blight of oat —current status in western Canada". *Canadian Journal of Plant Pathology* 26 (4): 473–479. doi:10.1080/07060660409507167.

Teleman, Johan, Aakash Chawade, Marianne Sandin, Fredrik Levander and Johan Malmström. 2016. "Dinosaur: A Refined Open-Source Peptide MS Feature Detector". *Journal of Proteome Research* 15 (7): 2143–2151. doi:10.1021/acs.jproteome.6b00016.

Teleman, Johan, Hannes L. Röst, George Rosenberger, Uwe Schmitt, Lars Malmström et al. 2015. "DIANA-algorithmic improvements for analysis of data-independent acquisition MS data". *Bioinformatics* 31 (4): 555–562. doi:10.1093/bioinformatics/btu686.

The, Matthew, and Lukas Käll. 2019. "Integrated identification and quantification error probabilities for shotgun proteomics". *Molecular and Cellular Proteomics* 18 (3): 561–570. doi:10.1074/mcp.RA118.001018.

Thompson, Andrew, Jürgen Schäfer, Karsten Kuhn, Stefan Kienle, Josef Schwarz et al. 2003. "Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS". *Analytical Chemistry* 75 (8): 1895–1904. doi:10.1021/ac0262560.

Thornton, Philip K., Peter G. Jones, Polly J. Ericksen and Andrew J. Challinor. 2011. "Agriculture and food systems in sub-Saharan Africa in a 4°C+ world". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 369 (1934): 117–136. doi:10.1098/rsta.2010.0246.

Troyanskaya, Olga, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie et al. 2001. "Missing value estimation methods for DNA microarrays". *Bioinformatics* 17 (6): 520–525. doi:10.1093/bioinformatics/17.6.520.

Tsou, Chih Chiang, Dmitry Avtonomov, Brett Larsen, Monika Tucholska, Hyungwon Choi et al. 2015. "DIA-Umpire: Comprehensive computational framework for data-independent acquisition proteomics". *Nature Methods* 12 (3): 258–264. doi:10.1038/nmeth.3255.

Tyanova, Stefka, and Juergen Cox. 2018. "Perseus: A bioinformatics platform for integrative analysis of proteomics data in cancer research". In *Cancer Systems Biology*, ed. by Louise von Stechow, 1711:133–148. Springer. ISBN: 9781493967568. doi:10.1007/978-1-4939-7493-1_7.

Tyanova, Stefka, Tikira Temu and Juergen Cox. 2016. "The MaxQuant computational platform for mass spectrometry-based shotgun proteomics". *Nature Protocols* 11 (12): 2301–2319. doi:10.1038/nprot.2016.136.

Utt, Matthew D. 2016. "Prediction of bull fertility". *Animal Reproduction Science* 169:37–44. doi:10.1016/j.anireprosci.2015.12.011.

Välikangas, Tommi, Tomi Suomi and Laura L. Elo. 2017. "A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation". *Briefings in Bioinformatics* 19 (6): 1344–1355. doi:10.1093/bib/bbx054.

– . 2018. "A systematic evaluation of normalization methods in quantitative label-free proteomics". *Briefings in Bioinformatics* 19 (1): 1–11. doi:10.1093/bib/bbw095.

Van Riper, Susan K., Ebbing P. De Jong, Leeann Higgins, John V. Carlis and Timothy J. Griffin. 2014. "Improved intensity-based label-free quantification via proximity-based intensity normalization (PIN)". *Journal of Proteome Research* 13:1281–1292. doi:10.1021/pr400866r.

Varshney, Rajeev K., Kailash C. Bansal, Pramod K. Aggarwal, Swapan K. Datta and Peter Q. Craufurd. 2011. "Agricultural biotechnology for crop improvement in a variable climate: Hope or hype?" *Trends in Plant Science* 16 (7): 363–371. doi:10.1016/j.tplants.2011.03.004.

Velculescu, Victor E., Lin Zhang, Bert Vogelstein and Kenneth W. Kinzler. 1995. "Serial Analysis of Gene Expression". *Science* 270 (5235): 484–487. doi:10.1126/science.270.5235.484.

Venet, David, Jacques E. Dumont and Vincent Detours. 2011. "Most random gene expression signatures are significantly associated with breast cancer outcome". *PLoS Computational Biology* 7 (10): e1002240. doi:10.1371/journal.pcbi.1002240.

Walach, Jan, Peter Filzmoser and Karel Hron. 2018. "Data Normalization and Scaling: Consequences for the Analysis in Omics Sciences". In *Data Analysis for Omic Sciences: Methods and Applications*, 1st ed., ed. by Joaquim Jaumot, Carmen Bedia and Romà Tauler, 82:165–196. Elsevier. ISBN: 9780444640444. doi:10.1016/bs.coac.2018.06.004.

Wang, Baohua, and Peng W Chee. 2010. "Application of advanced backcross quantitative trait locus (QTL) analysis in crop improvement". *Journal of Plant Breeding and Crop Science* 2 (8): 221–232. doi:10.1007/BF00223376.

Wang, Pei, Hua Tang, Heidi Zhang, Jeffrey Whiteaker and Amanda G Paulovich. 2006. "Normalization Regarding Non-Random Missing Values in High-Throughput Mass Spectrometry Data". *Pacific Symposium on Biocomputing*: 315–326.

Wang, San Yuan, Ching Hua Kuo and Yufeng J. Tseng. 2013. "Batch normalizer: A fast total abundance regression calibration method to simultaneously adjust batch and injection order effects in liquid chromatography/time-of-flight mass spectrometry-based metabolomics data and comparison with current calibration met". *Analytical Chemistry* 85 (2): 1037–1046. doi:10.1021/ac302877x.

Wang, Wei, Andrew C.H. Sue and Wilson W.B. Goh. 2017. "Feature selection in clinical proteomics: with great power comes great reproducibility". *Drug Discovery Today* 22 (6): 912–918. doi:10.1016/j.drudis.2016.12.006.

Wang, Wei Qing, Ole Nørregaard Jensen, Ian Max Møller, Kim H. Hebelstrup and Adelina Rogowska-Wrzesinska. 2018. "Evaluation of sample preparation methods for mass spectrometry-based proteomic analysis of barley leaves". *Plant Methods* 14:72. doi:10.1186/s13007-018-0341-4.

Webb-Robertson, Bobbie Jo M., Melissa M. Matzke, Jon M. Jacobs, Joel G. Pounds and Katrina M. Waters. 2011. "A statistical selection strategy for normalization procedures in LC-MS proteomics experiments through dataset-dependent ranking of normalization scaling factors". *Proteomics* 11 (24): 4736–4741. doi:10.1002/pmic.201100078.

Webb-Robertson, Bobbie Jo M., Lee Ann McCue, Katrina M. Waters, Melissa M. Matzke, Jon M. Jacobs et al. 2010. "Combined statistical analyses of peptide intensities and peptide occurrences improves identification of significant peptides from MS-based proteomics data". *Journal of Proteome Research* 9 (11): 5748–5756. doi:10.1021/pr1005247.

Webb-Robertson, Bobbie Jo M., Holli K. Wiberg, Melissa M. Matzke, Joseph N. Brown, Jing Wang et al. 2015. "Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics". *Journal of Proteome Research* 14 (5): 1993–2001. doi:10.1021/pr501138h.

Whiteaker, Jeffrey R., Chenwei Lin, Jacob Kennedy, Liming Hou, Mary Trute et al. 2011. "A targeted proteomics-based pipeline for verification of biomarkers in plasma". *Nature Biotechnology* 29 (7): 625–634. doi:10.1038/nbt.1900.

Willforss, Jakob, and Fredrik Levander. 2019. "Evaluation of peptide quantification susceptibility for technical bias from batch effects and batch compensation". In *Poster presented at the Proteomic Forum 2019, Potsdam, Germany*.

Wolf-Yadlin, Alejandro, Sampsa Hautaniemi, Douglas A. Lauffenburger and Forest M. White. 2007. "Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks". *PNAS* 104 (14): 5860–5865. doi:10.1073/pnas.0608638104.

Wu, Felicia, John D. Groopman and James J. Pestka. 2014. "Public Health Impacts of Foodborne Mycotoxins". *Annual Review of Food Science and Technology* 5:351–372. doi:10.1146/annurev-food-030713-092431.

Wu, Xiao Lin, and Zhi Liang Hu. 2012. "Meta-analysis of QTL mapping experiments". In *Quantitative Trait Loci (QTL)*, ed. by John M. Walker, 871:145–171. Springer Protocols. ISBN: 9781617797842. doi:10.1007/978-1-61779-785-9_8.

Xie, Chongqing, and Shizhong Xu. 1998. "Efficiency of multistage marker-assisted selection in the improvement of multiple quantitative traits". *Heredity* 80:489–498. doi:10.1046/j.1365-2540.1998.00308.x.

Yang, Qingxia, Jiajun Hong, Yi Li, Weiwei Xue, Song Li et al. 2019. "A novel bioinformatics approach to identify the consistently well-performing normalization strategy for current metabolomic studies". *Briefings in Bioinformatics*: bbz137. doi:10.1093/bib/bbz137.

Yang, Qingxia, Yunxia Wang, Ying Zhang, Fengcheng Li, Weiqi Xia et al. 2020. "NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data". *Nucleic acids research* 48:W436–W448. doi:10.1093/nar/gkaa258.

Yang, Yee Hwa, and Terry Speed. 2002. "Design issues for cDNA microarray experiments". *Nature Reviews Genetics* 3:579–588. doi:10.1038/nrg863.

Yost, R. A., and C. G. Enke. 1978. "Selected Ion Fragmentation with a Tandem Quadrupole Mass Spectrometer". *Journal of the American Chemical Society* 100 (7): 2274–2275. doi:10.1021/ja00475a072.

Zaheer, Khalid, and M. Humayoun Akhtar. 2016. “Potato Production, Usage, and Nutrition—A Review”. *Critical Reviews in Food Science and Nutrition* 56 (5): 711–721. doi:10.1080/10408398.2012.724479.

Zeeya, Merali. 2010. “...Error ...why scientific programming does not compute”. *Nature* 467:775–777. doi:10.1038/467775a.

Zhang, Bo, Lukas Käll and Roman A. Zubarev. 2016. “DeMix-Q: Quantification-centered dataprocessing workflow”. *Molecular and Cellular Proteomics* 15 (4): 1467–1478. doi:10.1074/mcp.O115.055475.

Zhang, Yuqing, David F. Jenkins, Solaiappan Manimaran and W. Evan Johnson. 2018. “Alternative empirical Bayes models for adjusting for batch effects in genomic studies”. *BMC Bioinformatics* 19 (1): 262. doi:10.1186/s12859-018-2263-6.

Zhao, Zhou, Jinghui Liu, Ruizong Jia, Sarina Bao, Haixia et al. 2019. “Physiological and TMT-based proteomic analysis of oat early seedlings in response to alkali stress”. *Journal of Proteomics* 193:10–26. doi:10.1016/j.jprot.2018.12.018.

Zhu, Yafeng, Lukas M. Orre, Yan Zhou Tran, Georgios Mermelekas, Henrik J. Johansson et al. 2020. “DEqMS: a method for accurate variance estimation in differential protein expression analysis”. *Molecular and Cellular Proteomics* 19 (6): mcp.TIR119.001646. doi:10.1074/mcp.tir119.001646.