FACULTY OF SCIENCES

FACULTY OF MEDICINE AND HEALTH SCIENCES

STATISTICAL METHODS FOR DIFFERENTIAL PROTEOMICS AT PEPTIDE AND PROTEIN LEVEL

Ir. Ludger Goeminne
Student number: 00802186

Supervisors: Prof. Dr. Ir. Lieven Clement, Prof. Dr. Kris Gevaert

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Doctor of Statistical Data Analysis

Academic year: 2018 - 2019

GHENT UNIVERSITY

VIB science meets life

AGENTSCHAP INNOVEREN & ONDERNEMEN

BIG N2N Bioinformatics Institute Ghent from nucleotides to networks

The doctoral candidate, Ludger Goeminne, and the thesis supervisors, prof. Lieven Clement and prof. Kris Gevaert, guarantee, by signing this doctoral thesis that the work has been done by the doctoral candidate under the direction of the supervisors and, as far as our knowledge reaches, in the performance of this work, the rights of other authors to be cited (when their results or publications have been used) have been respected.

The research described in this thesis was conducted at the Department of Applied Mathematics, Computer Science and Statistics, Faculty of Sciences and at the Department of Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, both at Ghent University, Belgium.

Ghent, May 2019

The doctoral candidate

Ludger Goeminne

The thesis supervisors

Lieven Clement

Kris Gevaert

De promovendus, Ludger Goeminne, en de scriptiebegeleiders, prof. Lieven Clement en prof. Kris Gevaert, garanderen door het ondertekenen van dit proefschrift dat het werk is gedaan door de promovendus onder leiding van de begeleiders en, voor zover onze kennis reikt, bij de uitvoering van dit werk, de rechten van andere auteurs om geciteerd te worden (wanneer hun resultaten of publicaties zijn gebruikt) gerespecteerd werden.

Het onderzoek beschreven in dit proefschrift werd uitgevoerd aan de Vakgroep Toegepaste Wiskunde, Informatica en Statistiek, Faculteit Wetenschappen en aan de Vakgroep Biomoleculaire Geneeskunde, Faculteit Geneeskunde en Gezondheidswetenschappen, beiden aan de Universiteit Gent, België.

Gent, mei 2019

De promovendus

Ludger Goeminne

De scriptiebegeleiders

Lieven Clement

Kris Gevaert

Thesis supervisors

Prof. dr. ir. Lieven Clement, Department of Applied Mathematics, Computer Science and Statistics, Faculty of Sciences, Ghent University

Prof. dr. Kris Gevaert, Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University

Other members of the doctoral examination committee

Chair: prof. dr. Jan De Neve, Department of Data Analysis, Faculty of Psychology and Educational Sciences, Ghent University

Secretary: prof. dr. ir. Olivier Thas, Department of Data analysis and mathematical modelling, Faculty of Bioscience Engineering, Ghent University

Prof. dr. ir. Dirk Valkenborg, Department of Mathematics and Statistics, Faculty of Sciences, Hasselt University

Prof. dr. Laurent Gatto, de Duve Institute, Faculty of Pharmacy and Biomedical Sciences, Université Catholique de Louvain

Dr. ir. Gerben Menschaert, Department of Data analysis and mathematical modelling, Faculty of Bioscience Engineering, Ghent University

Dr. Maarten Dhaenens, Department of Pharmaceutics, Faculty of Pharmaceutical Sciences, Ghent University

Carl T. Bergstrom @CT_Bergstrom · Feb 19

My PhD advisor told me to put a ten dollar bill between the pages of my thesis in the university library.

"So I can check to see if anyone read it?", I asked.

"No, of course no one will read it," he replied, "but when you come back into town you'll always have money for lunch."

“An expert is a person who has made all the mistakes that can be made in a very narrow field.” — Niels Bohr

“Perfection is not when there is no more to add, but no more to take away.” — Antoine De Saint-Exupéry

“Anyone who lives within their means suffers from a lack of imagination.” — Oscar Wilde

FOREWORD – WOORD VOORAF

5 years ago, in September 2013, I was ready to start my journey as a PhD fellow. The 5 years before, I had studied Bioscience Engineering: cell and gene biotechnology and I had completed my master's thesis in the Cell Systems and Cellular Imaging (CSI) group of the fantastic – then dr., now prof. – Winnok De Vos, where I mimicked "Hutchinson-Gilford progeria", a rapid aging disease, in healthy cell cultures.

At that point, I had a decent knowledge about cellular biology and biochemistry, but I knew nothing about mass-spectrometry-based proteomics (the only thing I had learned about protein identification was one slide in our Biochemistry course about Edman degradation, a technique developed in 1950). Via courses like "Experimental design" and "Statistical genomics", I had some basic statistical knowledge (e.g. about linear regression), but I had never heard about things like "generalized linear models", "shrinkage estimation" or "Bayesian statistics".

The 5 years of my PhD have been a very long journey. Indeed, at the start of my PhD, I was affiliated with these institutions (IWT since January 2015):

Now, after my five-year journey, all my affiliations have changed their logos:

I feel proud of what I have achieved: I got thoroughly trained in both biostatistics and proteomics to the point that I developed my own software package "MSqRob" for quantification of label-free proteomics data and even gave a few trainings and an invited presentation for

researchers who are interested in using MSqRob. This is the nicest feeling: to know that I have done something useful and that MSqRob will still be used after I leave Ghent. Indeed, on May 15th, I will start a postdoctoral research in the lab of Johan Auwerx in Lausanne, Switzerland on a topic in which I have a quasi-lifelong interest: aging and longevity.

I could have never gotten to the point where I am now without the help of many people. I immediately would like to apologize to you if I would have forgotten to mention your name here. If you have contributed to this work in one way or another, directly or indirectly, please do know that I am very grateful to you.

First of all, I would like to express my gratitude to all members of the examination committee (prof. Jan De Neve, prof. Olivier This, prof. Laurent Gatto, prof. Dirk Valkenborg, dr. Maarten Dhaenens and dr. Gerben Menschaert) for thoroughly reading this thesis and for their valuable comments. They didn't make it easy for me, but their input has truly improved the quality of this document.

A special word of thanks goes of course my two amazing promotors: prof. Lieven Clement and prof. Kris Gevaert, who have always been very supportive. I would like to express my sincere gratitude to have been able to work with both of you.

Lieven is a true genius. While many biologists tend to use suboptimal methods to analyze their data and many computer scientists do not always grasp the peculiarities of a biological problem, Lieven has the exceptional ability to translate almost every biological problem, no matter its complexity, into a suited and solid statistical solution. I noticed many times that when Lieven enthusiastically shares his ideas during informal conversations at congresses, other scientists, even the most experienced ones, are impressed, much more than Lieven himself sometimes realizes. Indeed, as a true West-Fleming, he is modesty itself. Lieven's door is always open and hardly ever did I walk out of his office without having at least three possible solutions for a problem that seemed insurmountable to me at first. He is the kind of person that always makes time for his PhD students, sometimes with meetings in the Starbucks or even, e.g. during the visit of Terry Speed, in his own home in Bruges. Lieven, I am really happy to have had the chance to work for you and I sincerely hope that we can keep in touch for many years to come!

Kris is an inspiring scientist: he is always on top of the latest developments in the proteomics field and he too is always approachable. Kris gives his researchers the freedom and trust they need to perform their research. He also has an exceptional talent for writing and proofreading and a very sharp eye for both smaller and larger mistakes in a text. His contributions and changes bring the quality of a text to a higher level. Moreover, with a simple question he can uncover fundamental weaknesses in a manuscript or research proposal, enabling me to address these issues before the text is externally evaluated This is, in my opinion, one of the reasons why the review process of all my research articles went so smooth.

I also need to thank prof. Lennart Martens, who is a co-author on several of my manuscripts. Besides having invaluable expertise in protein bio-informatics, he has an excellent feeling for what is missing in the field and his strong network has helped us tremendously in getting accepted in the protein bio-informatics community. Moreover, he is one of the best speakers I know, and he has an inspiring vision on how to lead a modern research lab.

Another big thanks goes to prof. Sven Eyckerman, who allowed me to investigate the proteomics signatures of BTF-3 (Nac-β) knockdowns in WI38 and BJ cells. I have learnt a lot by performing these experiments. Sven is also the first person to point me towards the very interesting aging-related work of prof. Johan Auwerx at EPFL, where I will do my postdoc starting on May 15th.

I also want to thank prof. Klaas Vandepoele from VIB's Center for Plant Systems Biology for the nice collaboration during the first year of my PhD.

As a shared PhD student between Lieven and Kris, I divided my working time between two working environments: Sterre S9 and Rommelaere. At campus Sterre, I need to thank the following people:

I want to thank all my present and former colleagues from the StatOmics group (Adriaan, Koen, Caroline, Jeroen, Gust, Elke (many thanks for proofreading my introduction, it really improved the quality of my thesis!), Lisa, Gwendolien and Bart Jacobs) for the amazing work environment and the nice collaborations while teaching statistics to undergraduate students. Thanks to Adriaan in particular for introducing me to Krav Maga.

I'm also very grateful for the many fun office mates I have had over the years (Karel, Nele, Mushthofa, Bart Van Gasse, Sarah, Domien, Helena, Chris, Marko, Oliver Urs Lenz and Jeroen). A big thanks goes to Karel, my "godfather" at Ghent University, who introduced me to everyone and everything in the TWIST department. I also want to thank Bart Van Gasse and Sarah, as well as Christophe Ley, Benoît and Marnix for our participations in the yearly PRIME quiz.

Further thanks go to all other members of the statistics group, in particular my "Ghent University godchild" Fatemeh, Vahe (also thanks for being such a nice roommate and for proofreading part of my introduction!), Holger (also thanks for being such a nice roommate!), Camila, Paula, Beatriz, Oliver Dukes, Sjouke, Machteld, Slađja, Charlotte, Thẳng, Christophe Ley, Chloe, Arnout, David, Hans, Stijn and Els for all the nice work- and non-work-related moments. Also thanks to Kasper, with who I followed a few Mastat courses, for the nice lunches together and your inspiring ideas.

I would also like to thank the people from the TWIST secretariat: Herbert (thanks for all the IT help!), Wouter, Hilde, Katia and Ann, as well as all other TWIST members, especially those who frequently visit the coffee breaks: Robbert, Chamberlain, Hilmar, Kelly, Roy, Herman, Dieter, Bart Mesuere, Nico, Bart Van Rompaye, Charlotte, Michaël, Niels, Pieter, Veronique (thanks for driving me home from the VIB seminar!), Christophe Scholliers, Veerle, Michèle, Carmen, Annick, Tineke, Peter Dawyndt, Joris, Kris Coolsaet, Gunnar, Willy, Guido and Hans. Special thanks to Catherine for organizing the TWIST weekends and the after-work badminton. Also special thanks to Felix (for all your help with Git!) and Toon (for the help with Python!).

At campus Rommelaere, I want to thank the following people:

Special thanks to the MS operators (An, Evi, Delphi, Jarne, Pieter-Jan, Jonathan) for being the driving forces for the weekly Friday restaurant lunches, most of whom have also been my office mates. Thanks to Evi for introducing me to protein MS on my first day at Rommelaere and to all of you for answering my many MS-related questions and helping me out with MS data.

Big thanks to my friends and colleagues Maša, Ursula, Daria and to Igor and Ignacio for all the fun moments and the nice trip to Slovenia.

Also, many thanks to my Rommelaere office mates Giel, Eva (also for introducing me to the virus work), Noortje (also for explaining me how to make freeze stocks), Delphine (also for explaining me all the other protocols in the lab), Louis, Montse, Maarten, Sara, Esperanza, Michela, Elise, Arun, Jasmine, Sriram and Dai. Big thanks also to the other people from the proteomics group, many of whom also frequently joined for the Friday lunches and who also participated in after-work events (BBQs, watching the World Cup, etc.): Petra, Sven, Francis, Fabien, Lia, Rupert, Heidi, Sebastian, Annelies, Kevin, Bart Ruttens, Hans, Elisabeth, Jeff, Kim and Alan (also for explaining MaxQuant to me).

Further thanks go to the people from prof. Lennart Marten's CompOmics group, where I have also frequently stayed and who also invited me to their events: Andrea (for the nice collaboration on the integration of moFF with MSqRob), Pieter-Jan, Paola, Surya, Silvia, Niels and Ralf (both especially for the help with the coffee machine), Robbin, Pathmanaban, Tim, Demet, Şule, Elien (also thanks for proofreading my Marie Curie proposal!), Davy (also for the help with the PRIDE Archive API and Visual Basic programming) and Sven Degroeve.

I also want to thank some people from prof. Mo Lamkanfi's group (thank you for sending me an example of a former Marie Curie application, this has been tremendously helpful for me!): group: Anna, Daniel, Mike, Kevin, Pedro, Nathalia, Magda and Krzysztof. Thanks for the nice parties at countless VIB receptions and events. Many of you were also NeoDocs together with me. It has been a very nice experience to organize the NeoDoc party (also thanks to Slava and Hannes for that!).

I also want to thank all members of the VIBes organizing committee: Marleen, Halina, Cecilia, Jolien, Evi, Jessica, Yessica, Marlies, Hermien, Rocco, Aleksandra, Yannick, Iryna, Maria and Nandita. It has been a huge effort to organize this conference all by ourselves, but it was a lot of fun and I am very proud of what we achieved together! Also special thanks to dr. Michelle Linterman for joining us to the afterparty and for her kind invitation to Cambridge and to prof. Ruedi Aebersold for referring me to prof. Johan Auwerx.

Further thanks to prof. Nico Callewaert for his career advice, to my former thesis supervisor Winnok De Vos for his amazing letter of recommendation to prof. Johan Auwerx and to Peter Van den Hemel for all his IT-related help when I was at Rommelaere.

From here one, the foreword continues in Dutch, as the remainder aims at thanking my friends and family.

Verder wil ik ook graag Pieter bedanken om zo een toffe huisgenoot (en nadien overbuurman) te zijn en voor jouw advies en voor alle toffe momenten in de Goosedrive Condo en de feestjes in Gent! Om dezelfde redenen ook bedankt aan mijn voormalige huisgenoten Line, Serena, Holger, Vahe en Astrid!

Ik ben ook mijn vrienden uit de bio-ingenieursgroep zeer dankbaar dat we jaren na het afstuderen nog altijd zo een hechte groep zijn! De BBQ's, samen uitgaan, de weekends, enz.:

bedankt voor al die toffe momenten: Line, Margo, Yael, Michelle, Jeroen, Jorden, Niels, Andreia, Thomas, Kilian, Kevin en Harm!

Ook heel erg bedankt aan al mijn vrienden uit Ternat voor al de toffe evenementen en feestjes die we samen gedaan hebben: Joren, Guido, Maarten (bedankt om mijn thesis toch niet na te lezen :p), Steven, Laetitia, Elout, Eowyn, Vincent, Matthias, Kristien, Karen Schets, Kelly, Benjamin, Tom, Wander, Karen Raets.

Uiteraard ook bedankt aan mijn moeder en vader en aan Sigert, Sigien, Diederich, Boursin en Chaumes. Bedankt voor alles! Ik wil ook graag al mijn familieleden bedanken dat we zo een hechte groep zijn en voor de toffe momenten samen. In het bijzonder dank aan Pieter (zie eerder) en aan Ruth en Jarich en mijn andere mede-badmintonners!

Als allerlaatste wil ik in het bijzonder mijn allerliefste Zühal Duran bedanken voor al onze mooie momenten samen. Ik ben je ook bijzonder dankbaar voor al jouw geduld en jouw steun terwijl ik uren en uren aan het schrijven was. Dat we nog veel mooie momenten mogen beleven samen in Lausanne en daarna! Seni çok çok çok seviyorum, tatlı civciv!

Vanwege de grote hoeveelheid mensen die mij hebben geholpen, ben ik mogelijks in dit dankwoord nog een aantal mensen vergeten te vermelden. Ik wil mij bij hen op voorhand verontschuldigen: als u op de één of andere manier, direct of indirect, hebt bijgedragen aan het tot stand komen van deze thesis, weet dat ik u zeer erkentelijk ben.

Deze thesis is opdragen aan mijn grootva, Longinus Verdoodt en mijn bompa, Jean Goeminne, die dit moment helaas niet meer kunnen meemaken. Jullie hebben mij mee gemaakt tot wie ik ben en ik ben er trots op zulke fantastische mensen te hebben gekend!

Deze thesis is ook opgedragen aan Jelle Jacob, de allereerste vriend die ik ooit had en die recent veel en veel te vroeg van ons is weggerukt door een verkeersongeval. Woorden kunnen niet beschrijven hoe veel ik doorheen mijn leven aan jou gehad heb. Ik zal jouw enthousiasme nooit vergeten. Rust zacht, vriend.

SUMMARY

Proteins are very diverse biomolecules that facilitate nearly all cellular processes of life. They interact with each other in complex networks in which disruption of a single protein can severely impact an organism. Therefore, quantitative information of a proteome (i.e. the entire set of proteins present in an organism) is extremely important to gain insights in the functioning of an organism in both healthy and diseased states.

Mass spectrometry (MS)-based proteomics is the method of choice for the high-throughput identification and quantification of thousands of proteins in a single analysis. When deep proteome coverage on many samples is needed, the analysis is often performed without any stable isotope labels, label-free. Here, proteins are extracted and digested into peptides that are subsequently loaded onto a reverse-phase high-performance liquid chromatography column (HPLC) coupled to a mass spectrometer, by which they are separated, ionized and have their MS spectra recorded. The intensity peaks in these MS spectra are proxies for peptide abundance. Subsequently, (some) peptides are targeted for fragmentation and the resulting MS^2 spectra enable their identification. As a result, label-free proteomics data are hierarchical: the data are at the peptide ion level, while inference typically happens at the protein level. Important to note is that signal intensities are strongly peptide-dependent as some peptides ionize more efficiently than others. Furthermore, missing values are very common and a large fraction of this missingness is not at random. Indeed, intensities of low-abundant and poorly ionizing peptides are more likely to go missing and competition for ionization makes missingness also context-dependent.

Many *ad hoc* data analysis workflows for differential protein quantification do not handle label-free proteomics data in a statistically rigorous way, which leads to suboptimal ranking of differentially abundant proteins. Consequently, many biologically relevant proteins remain unnoticed and valuable resources are wasted by needlessly trying to validate false positive hits.

In chapter 8, we demonstrate the necessity of properly taking peptide-specific effects into account in differential protein quantification analyses. Peptide-based models, which naturally account for these effects, perform better than methods that summarize peptide intensities to the protein level prior to the statistical analysis. We further illustrate that controlling the false discovery rate becomes problematic when highly-abundant proteins are differentially abundant due to suppression of the intensity of the background proteome. Finally, we show that missing values should be handled with care as imputing these under wrong assumptions leads to worse results compared to not imputing missing values at all.

Most peptide-based models suffer from overfitting, unstable estimations of residual variances and a disproportionate impact of outlying peptide intensities. To address these issues, I developed the versatile R package MSqRob, which adds three modular improvements to existing peptide-based models: ridge regression stabilizes fold change estimates, empirical Bayes variance estimations stabilize the variances of the test statistics and M-estimation with Huber weights reduces the impact of outliers. MSqRob's algorithm has been described in detail in section 9.1 and it not only improves the fold change estimates in terms of precision and accuracy, but also the protein ranking, leading to a better discrimination between true and false positives. MSqRob is freely available on GitHub (https://github.com/statOmics/MSqRob) and has a user-friendly graphical interface that is made in "Shiny", an R package developed by RStudio that allows smooth integration of the R programming language with an HTML interface.

In section 9.2, I pinpoint important aspects of both experimental design and data analysis. Furthermore, I provide a step-by-step guide on how to use the MSqRob graphical user interface for both simple as well as more complex experimental designs. I also provide well-documented scripts to run analyses in bash mode, enabling the integration of MSqRob in automated pipelines on cluster environments.

In my latest, unpublished work (chapter 10), I focus on the missing value problem. Indeed, missingness in label-free proteomics is a mix of missingness completely at random and missingness not at random. However, the exact contributions of both mechanisms are unknown and dataset-specific, and imputing under the wrong assumptions is detrimental for the downstream protein quantifications. Therefore, I developed a hurdle model that combines the power of MSqRob with the complementary information that is available in peptide counts without having to rely on undeterminable assumptions. This enables MSqRob to quantify proteins that are completely missing in one condition in a statistically rigorous manner. Moreover, it opens new possibilities to detect the sudden appearance of post-translationally modified peptides in addition to traditional protein fold change estimation.

With the development of MSqRob, I have made an important contribution to enabling experimenters to get the most out of their proteomics data. And, even though MSqRob is already one of the most versatile differential proteomics quantification tools, there are ample opportunities to broaden MSqRob's scope, both towards new types of (prote)omics data and towards more complicated experimental designs.

SAMENVATTING

Eiwitten (ook “proteïnen” genoemd) zijn heel diverse biomoleculen die bijna alle cellulaire processen van het leven faciliteren. Ze interageren met elkaar in complexe netwerken, waarbij verstoring van één enkel eiwit een ernstige impact kan hebben op het organisme. Daarom is kwantitatieve informatie over een proteoom (het geheel van eiwitten dat in een organisme aanwezig is) van groot belang om inzicht te krijgen in het functioneren van een organisme, zowel in zieke als in gezonde toestand.

Massaspectrometrie (MS)-gebaseerde proteomics is de voorkeursmethode voor de identificatie en kwantificatie van duizenden eiwitten in één enkele analyse. Wanneer een diepe proteoomdekking op veel stalen vereist is, wordt de analyse vaak uitgevoerd zonder gebruik te maken van stabiele isotopen, zogenaamde labelvrije proteomics. Hierbij worden eiwitten geëxtraheerd en gekliefd in peptiden die vervolgens geladen worden op een *reverse-phase high-performance liquid*-chromatografiekolom (HPLC) gekoppeld aan een massaspectrometer (MS), waardoor ze worden gescheiden, geïoniseerd en hun MS-spectra worden geregistreerd. De intensiteitspieken in deze MS spectra zijn een soort surrogaat voor peptide-abundantie. Vervolgens worden (sommige) peptiden geselecteerd voor fragmentatie en de resulterende MS^2-spectra laten toe om deze peptiden te identificeren. Labelvrije proteomics-data zijn bijgevolg hiërarchisch: de data bevinden zich op het niveau van de peptide-ionen, terwijl inferentie meestal op het eiwitniveau gebeurt. Belangrijk om op te merken is dat signaalintensiteiten sterk peptide-afhankelijk zijn omdat sommige peptiden efficiënter ioniseren dan andere. Ontbrekende waarden komen dus veelvuldig voor en zijn een groot deel niet willekeurig. Inderdaad, intensiteiten voor laag abundante en slecht ioniserende peptiden zullen gemakkelijker ontbreken en competitie voor ionisatie maakt zulke ontbrekende waarden ook contextafhankelijk.

Veel *ad hoc* data-analyseworkflows voor differentiële eiwitkwantificering verwerken labelvrije proteomics-data niet op een statistisch rigoureuze manier, wat leidt tot een suboptimale rangschikking van differentieel abundante eiwitten. Bijgevolg blijven vele biologisch relevante eiwitten onopgemerkt en wordt waardevolle tijd en materiaal verspild door het onnodig proberen valideren van valse positieve hits.

In hoofdstuk 8 tonen we de noodzaak van het correct rekening houden met peptide-specifieke effecten in differentiële eiwit-kwantificatieanalyses. Peptide-gebaseerde modellen, die deze effecten op natuurlijke wijze in rekening brengen, presteren beter dan methoden die, voorafgaand aan de statistische analyse, de peptide-intensiteiten samenvoegen naar het eiwitniveau. We illustreren dat het controleren van de *false discovery rate* problematisch wordt wanneer zeer abundante eiwitten differentieel abundant zijn vanwege de suppressie van de intensiteiten van het ongewijzigde deel van het proteoom. Ten slotte laten we ook zien dat er voorzichtig moet worden omgegaan met ontbrekende waarden omdat imputeren onder verkeerde aannames leidt tot slechtere resultaten in vergelijking met niet imputeren.

De meeste peptide-gebaseerde modellen lijden onder *overfitting*, onstabiele schattingen van de residuele varianties en een buitenproportionele invloed van de uitschietende peptide-intensiteiten. Om deze problemen aan te pakken, ontwikkelde ik het veelzijdige R-pakket MSqRob, dat drie modulaire verbeteringen toevoegt aan bestaande peptide-gebaseerde modellen: ridge-regressie stabiliseert de schattingen voor de differentiële-abundantieratio’s, empirical Bayes variantieschattingen stabiliseren de varianties van de teststatistieken en *M-estimation* met Huber-gewichten vermindert de impact van uitschieters. Het algoritme van MSqRob is in detail beschreven in sectie 9.1. MSqRob verbetert niet alleen de geschatte

abundantieratio's in termen van precisie en nauwkeurigheid, maar verbetert ook de rangschikking van de eiwitten, wat leidt tot een beter onderscheid tussen echte en valse positieven. MSqRob is gratis beschikbaar op GitHub (https://github.com/statOmics/MSqRob) en heeft een gebruiksvriendelijke grafische interface gemaakt in "Shiny", een R-pakket ontwikkeld door RStudio dat een soepele integratie toelaat van de R-programmeertaal met een html-interface.

In sectie 9.2 identificeer ik belangrijke aspecten van zowel experimenteel ontwerp als van de data-analyse. Verder geef ik een stapsgewijze handleiding over het gebruik van MSqRob's grafische interface voor zowel eenvoudige als meer complexe experimentele ontwerpen. Ik lever ook goed gedocumenteerde scripts om analyses uit te voeren in bash-modus, waardoor de integratie van MSqRob in geautomatiseerde analyses in clusteromgevingen mogelijk wordt.

In mijn laatste, niet-gepubliceerde werk (hoofdstuk 10) focus ik op het probleem van de ontbrekende waarden. Ontbrekende waarden in labelvrije proteomics zijn immers een verzameling van volledig willekeurige ontbrekende waarden en ontbrekende waarden die niet willekeurig zijn. De exacte contributies van beiden zijn onbekend en dataset-specifiek, en imputeren onder verkeerde aannames is nadelig voor de eiwitkwantificaties in de daaropvolgende differentiële analyse. Daarom ontwikkelde ik een *hurdle*-model dat de kracht van MSqRob combineert met de complementaire informatie die beschikbaar is uit het aantal peptiden zonder te steunen op aannames die niet kunnen worden nagegaan. Hierdoor kan MSqRob toch eiwitten die volledig ontbreken in één conditie kwantificeren op een statistisch rigoureuze manier. Bovendien opent dit nieuwe mogelijkheden om het plotselinge opkomen van posttranslationeel gemodificeerde peptiden te detecteren, bovenop de gewone schatting van de eiwit-fold changes.

Met de ontwikkeling van MSqRob heb ik een belangrijke bijdrage geleverd om onderzoekers in staat te stellen het maximale uit hun waardevolle proteomics-data te halen. En hoewel MSqRob al een van de meest veelzijdige differentiële proteomics kwantificeringsinstrumenten is, zijn er voldoende mogelijkheden om het bereik van MSqRob uit te breiden, zowel naar nieuwe soorten (prote)omics-gegevens als naar meer gecompliceerde experimentele ontwerpen.

ABBREVIATIONS

AUC	area under the curve
CID	collision-induced dissociation
CPTAC	Clinical Proteomic Technology Assessment for Cancer Network
DA	differential abundance
DE	differential expression
DDA	data-dependent acquisition
DIA	data-independent acquisition
ESI	electrospray ionization
ETD	electron-transfer dissociation
FC	fold change
FDR	false discovery rate
FN	false negatives
FP	false positives
GFP	green fluorescent protein
HCD	higher-energy collisional dissociation
HILIC	hydrophilic interaction liquid chromatography
HPLC	high-performance liquid chromatography
IMAC	immobilized metal affinity chromatography
IQR	interquartile range
iTRAQ	isobaric tag for relative and absolute quantitation
LC	liquid chromatography
LFQ	label-free quantification
kNN	k-nearest neighbors
KO	knock-out
MALDI	matrix-assisted laser desorption
MCAR	missingness completely at random
MCMC	Markov Chain Monte Carlo
MDS	multidimensional scaling
MNAR	missingness not at random
MS	mass spectrometry

MOAC	metal-oxide affinity chromatography
NETD	negative electron-transfer dissociation
OR	odds ratio
pAUC	partial area under the curve
PPV	positive predictive values
PSM	peptide-to-spectrum match
QRILC	quantile regression imputation of left censored data
ROC	receiver operating curve
rpAUC	relative partial area under the curve
RP-HPLC	reverse-phase high-performance liquid chromatography
RR	robust ridge
SAX	strong anion exchange
SCX	strong cation exchange
SILAC	stable isotope labeling of amino acids in cell culture
SWATH-MS	sequential windowed acquisition of all theoretical fragment ion mass spectra
TMT	tandem mass tags
TN	true negatives
TOF	time-of-flight
TP	true positives
UPS	Universal Proteomics Standard
UPS1	Universal Proteomics Standard 1
WT	wild type

SHORT TABLE OF CONTENTS

LONG TABLE OF CONTENTS

PART I: INTRODUCTION

During the five years of my PhD, I thoroughly investigated different statistical approaches to quantify proteins in label-free mass spectrometry (MS)-based shotgun proteomics. Furthermore, I developed MSqRob, an R package with graphical user interface for the statistically sound analysis of label-free proteomics data. Since I have worked on the interface of protein biology and statistics, it is important to understand both the biological and the statistical aspects of my work.

Hence, in order to place my work in its proper context, this introduction is divided into four chapters. The first chapter aims to give an overview of the biology of proteins and the wide variety of applications of present-day mass spectrometry-based proteomics. In the second chapter, I will describe the technical context of bottom-up quantitative proteomics: the different quantification strategies and the specific peculiarities of the label-free proteomics workflow. In chapter three, I will explain how the spectra are processed into interpretable data. Finally, chapter four will give an overview of how this data can be used to quantify proteins.

1. BIOLOGICAL CONTEXT

Chapter 1 mainly aims at introducing proteomics to data analysists who are new to the field. In this chapter, I will first cover the very basics of protein biology (section 1.1). Then, I will give an overview of a generalized mass spectrometry-based proteomics workflow and the relationship of proteomics to other omics (section 1.2), followed by a more profound review of the possibilities of mass spectrometry-based proteomics in present-day life sciences research (section 1.3).

1.1. Proteins as the central effectors of life

Before discussing the need for proteomics, I will first introduce the biology of proteins. Proteins are an extremely diverse class of biomolecules that are essential for nearly all functions of life. In this section, I will give an overview of the molecular structure of proteins and how their structures are linked to the essential roles proteins play in health and disease.

1.1.1. The molecular structure and origin of proteins

Proteins are composed of amino acids and the general chemical structure of an amino acid and a protein is given in Fig. 1.1.

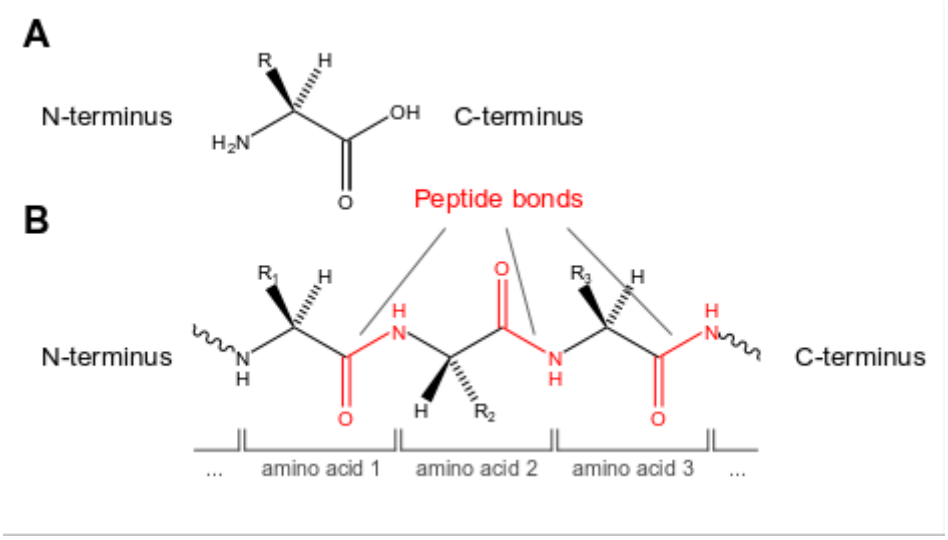


Figure 1.1. Chemical structures of a proteinogenic[1] amino acid (A) and a protein (B). All amino acids share the same base structure. They all contain an α-amino group (-NH_2) and an α-carboxyl group (-COOH) along with a rest group/side-chain (R). The part of the protein ending with the amino group is called the amino- or N-terminus, while the side ending in the carboxyl group is called the carboxyl- or C-terminus. Amino acids differ only in their rest group. In proteins, amino acids are joined together by peptide bonds (-CO-NH-, indicated in red).

[1] Proteinogenic amino acids are amino acids that are translationally incorporated into proteins. All proteinogenic amino acids, except glycine, have an L-stereoisomeric configuration. This means that, when the amino acid is oriented from its N-terminus to its C-terminus as in (A), the rest group (R) will be in front of the plane while the hydrogen atom (H) will be behind the plane. Glycine has no chiral center because its rest group is a hydrogen atom, so the central carbon (α-carbon) is only linked to three different atoms.

There are only 20 different standard amino acids[2] that can be incorporated in proteins. The actual sequence by which amino acids are joined together is encoded by the corresponding gene found in the genomic DNA (deoxyribonucleic acid). Indeed, in all living organisms, genes are transcribed to RNA (ribonucleic acid) molecules. RNA molecules are composed of only four different nucleotide building blocks holding the nucleobases guanine (G), uracil (U), adenine (A) or cytosine (C). In eukaryotic[3] cells, protein-coding RNA, or messenger RNA (mRNA), is then transported from the nucleus (where the grand majority of the DNA resides) to the cytoplasm, where it can be translated into proteins by ribosomes (Fig. 1.2). This unidirectional transfer of information takes place in every living organism: from DNA to mRNA to proteins and has come to be known as the central dogma of molecular biology. Even though some viruses violate this dogma by directly replicating RNA using an RNA template[4], or even generate DNA based on an RNA template[5], the translation of mRNA to proteins remains a one-way process. RNA is however not always translated into protein. Indeed, RNA itself can have regulatory functions (e.g. micro RNA (miRNA) and long non-coding lncNRA[6] can induce gene silencing [7, 8]), structural functions (e.g. ribosomal rNA (rRNA) is an important component of the ribosome [9]) and even catalytic functions (e.g. peptidyl transfer by rRNA, mRNA splicing, self-splicing [10, 11]).

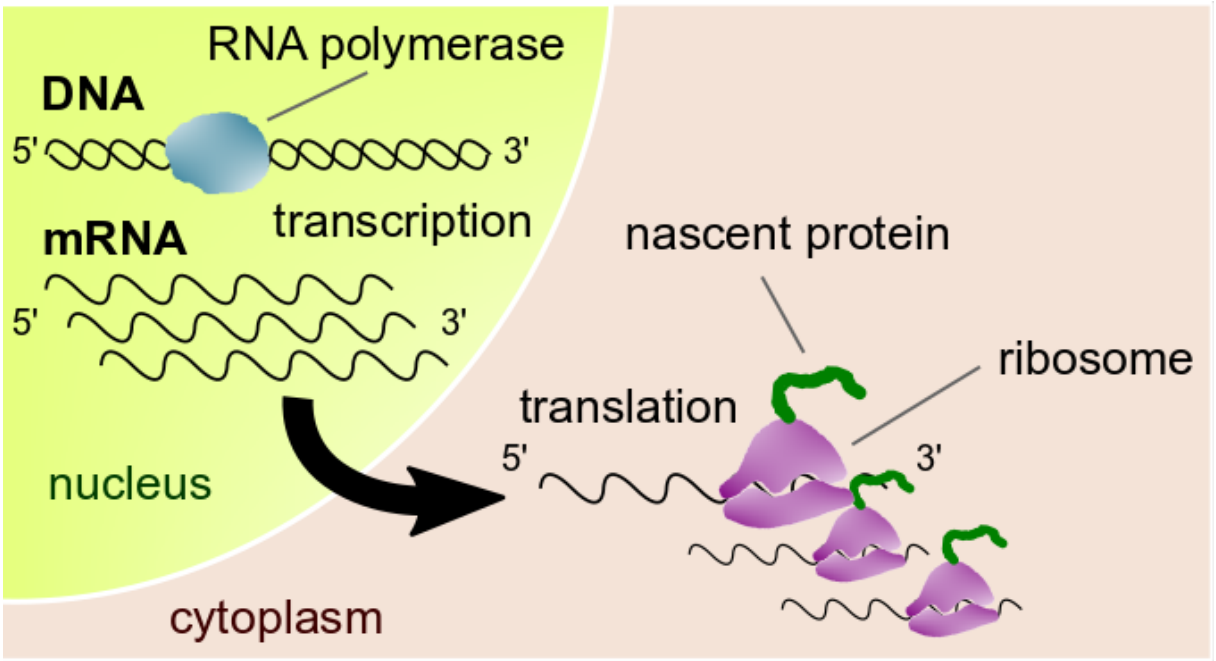


Figure 1.2. The grand majority of transcription occurs in the nucleus of a eukaryotic cell by a protein (or rather, enzyme) called RNA polymerase. Translation occurs in the cytoplasm where ribosomes, specialized cellular structures that are composed of ribosomal RNA and proteins, translate the mRNA into proteins. Both transcription and translation occur from the 5' to the 3' end of the oligonucleotides. 5'

---

[2] There are two additional very rare non-standard amino acids that are translationally incorporated into proteins. Selenocysteine (Sec, U) is present in all domains of life [1], while pyrrolysine (Pyl, O) is only present in 9 methanogenic *Archaea* of the *Methanosarcina* family and 15 *Bacteria* [2]. Both amino acids are present in only a few dozens of proteins.

[3] Eukaryotes are cells that, unlike *Bacteria* and *Archaea*, have a nucleus. All multicellular organisms (e.g. humans) are eukaryotic, though some eukaryotes are also unicellular.

[4] RNA viruses such as rhinoviruses (the most common causes of the common cold) and hepatitis C virus use a protein called RNA-dependent RNA polymerase to make new copies of their RNA genomes.

[5] HIV uses the protein reverse transcriptase to convert its RNA genome into DNA and subsequently integrates this DNA into its host's genome.

[6] Note that some short open reading frames in long "non-coding" RNA were shown to generate very small proteins [3-6].

and 3' refer to the conventional chemical names of the carbon atoms in the (deoxy)ribose rings. Proteins are always synthesized from their N- to their C-termini.

Transcription and translation are also unidirectional in space: they always occur from the 5' to the 3' terminus of the DNA and mRNA molecules respectively. Each group of three consecutive nucleotides (triplet) in the mRNA represents a codon and each codon represents one unique amino acid or encodes a stop codon. A general overview of the codons that translate into each of the 20 amino acids or are used as stop codons is given in Fig. 1.3. This genetic code is near-universal across the whole tree of life. Some minor exceptions include: yeasts from the CTG clade encode CUG partially or completely as serine instead of leucine [12] and UGA is sometimes translated into tryptophan or arginine instead of being used as a stop codon for some transcripts and some species [13, 14]. Note that all codons, except those for methionine and tryptophan, are redundant: e.g. UUA, UUG, CUU, CUC, CUA and CUG all code for leucine. This property is also called the "degeneracy" of the genetic code.

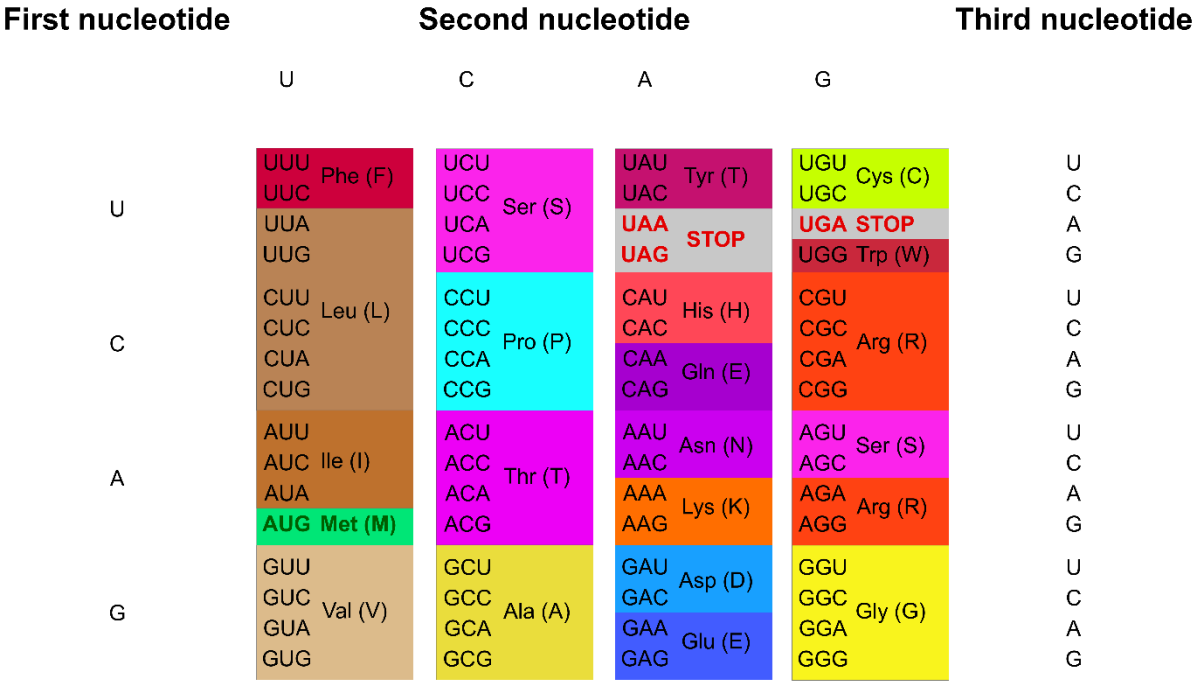


Figure 1.3. The genetic code. The four mRNA ribonucleosides are guanosine (G), uridine (U), adenosine (A), and cytidine (C). The 20 amino acids are phenylalanine (Phe, F), leucine (Leu, L), isoleucine (Ile, I), methionine (Met, M), valine (Val, V), serine (Ser, S), proline (Pro, P), threonine (Thr, T), alanine (Ala, A), tyrosine (Tyr, Y), histidine (His, H), glutamine (Gln, Q), asparagine (Asn, N), lysine (Lys, K), aspartic acid (Asp, D), glutamic acid (Glu, E), cysteine (Cys, C), tryptophan (Trp, W), arginine (Arg, R) and glycine (Gly, G). Note that each amino acid has both a three-letter abbreviation and a one-letter abbreviation. The AUG codon is the most common start codon and also codes for methionine. Hence, all nascent eukaryotic proteins start with a methionine at their N-terminus when being synthesized[7]. The three different stop codons (UAA, UAG and UGA) signal translation termination.

However, knowing the mRNA sequence sometimes does not suffice to predict the protein sequence. Indeed, in all domains of life, besides AUG, which is used in more than 80% of the

[7] In *Bacteria*, the start codon codes for N-formylmethionine, but this formyl group is cotranslationally removed [15]. In all domains of life, the N-terminal methionine is often cleaved off, causing more than 50% of the proteins not to have a methionine at the N-terminus [16-18].

cases, other codons are also used as start codons [19]. Such so-called near-cognate start codons include CUG, GUG, UUG, ACG, AUC, AUU, AAG, AUA and AGG [20]. In the bacterium *Escherichia coli* for example, up to 40 out of the 64 codons can be used as start codons, albeit in less than 0.1% of the cases [21]. Such alternative start codons also encode for methionine or N-formylmethionine (in the case of bacteria). If translation is initiated from a downstream[8] (near-cognate or canonical) start codon, a shorter protein is produced from the same mRNA template. Similarly, when translation is initiated from an upstream start codon, a longer protein is produced. More rarely, a stop codon can be replaced with another amino acid in a process called translational read-through [22].

Further, in both prokaryotes and eukaryotes, certain "slippery" mRNA sequences, such as AAAAAA, might, under certain circumstances, cause a frameshift, which means that translation continues in another reading frame [23] (Fig. 1.4).

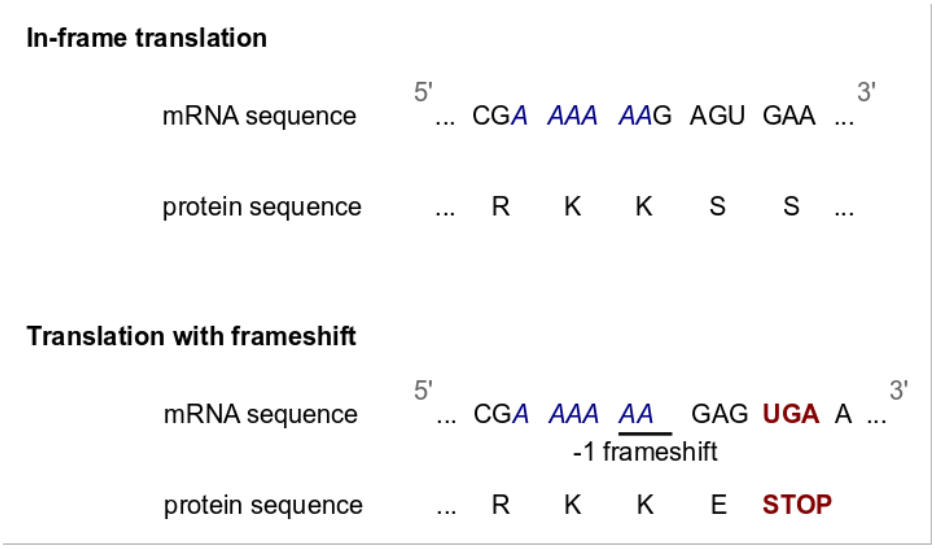


Figure 1.4. Example of a translational frameshift in the dnaX gene of the bacterium Escherichia coli. The dnaX gene encodes both the τ and the γ subunit of the DNA polymerase III protein. dnaX contains a slippery AAAAAA sequence (blue, italics). When the ribosome passes normally over this sequence (top), translation remains in frame and the τ subunit is produced. However, when the ribosome "slips" (bottom), a -1 frameshift occurs. Hereby, a premature stop codon (red, bold) is introduced resulting in the production of the shorter γ subunit. In the dnaX example, ribosome slipping is stimulated by the presence of a downstream stem loop structure in the mRNA that stalls the ribosome. An upstream Shine-Dalgarno like sequence helps repositioning the ribosome in its new reading frame. Example adapted from Dinman (2006) [24].

This ribosomal frameshift is rather rare in most organisms, but very common amongst viruses as it allows them to translate many proteins from a small genome that is limited by the size of the viral particles [24]. Similarly, slippage can already occur at the level of transcription, when the RNA polymerase introduces a variable number of nucleotides in long homopolymeric stretches [25]. Further, it is now also known that amino acids in bacterial proteins can be converted into their D-stereoisomeric form and recent work demonstrates that even a protein's backbone can be changed by introducing an α-keto-β-amino acid [26, 27].

Besides these rather infrequent phenomena discussed above, alternative RNA splicing is very common as more than 95% of all mammalian genes express alternatively spliced transcripts [28]. Splicing involves the removal of certain parts of an RNA molecule and, dependent on which parts are being spliced out from an mRNA molecule, different protein products can be

[8] Downstream means "in the 3' direction", upstream is "in the 5' direction".

generated. Similarly, certain protein sequences, called inteins, are able to post-translationally cut themselves out of a protein [29]. Moreover, proteins are known to carry, often transiently, a plethora of modifications (see also 1.1.3).

Because of these phenomena, many chemically different protein molecules can result from a single gene hence, the term "proteoform" was coined. Proteoforms are defined as "*Highly related protein molecules arising from all combinatorial sources of variation giving rise to products arising from a single gene. These include products differing due to genetic variations, alternatively spliced RNA transcripts, and post-translational modifications*" [30].

The side-chain of an amino acid determines its physicochemical properties. The nature of a protein will thus not only be determined by the amino acids it contains, but also by the sequence in which they are connected to each other. The specific chemical structures for selected amino acids are given in Fig. 1.5.

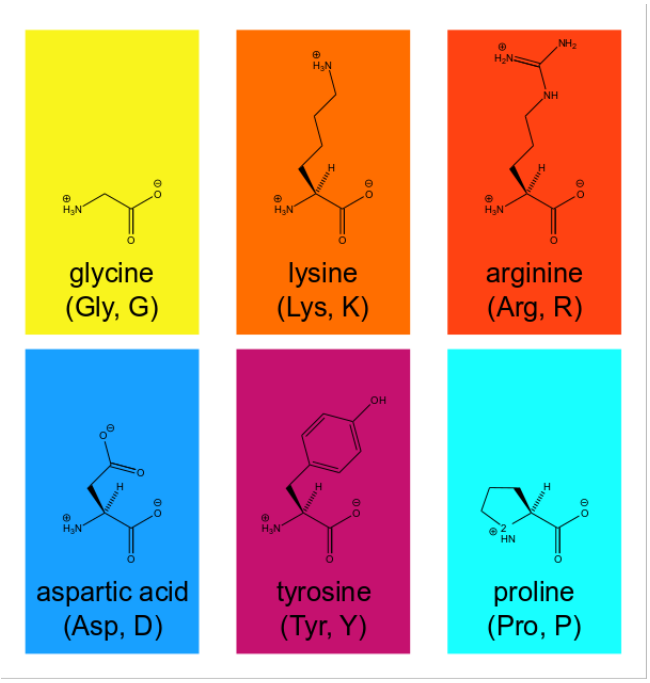


Figure 1.5. Examples of the diversity in chemical structures of amino acids: the structures of 6 different amino acids are given. The side-chains of lysine and arginine are long aliphatic chains that are positively charged at physiological pH[9]. Aspartic acid has a net negative charge at this pH. Tyrosine is a very bulky amino acid that is rather hydrophobic due to the benzene ring in its structure. It is however rather polar due to its hydroxyl (-OH) group. Glycine is the smallest amino acid and the only one that lacks a chiral center as its side-chain itself is a hydrogen atom.

1.1.2. Protein folding

Proteins arrange themselves into three dimensional structures. They are rather flexible, and it is mainly their sequence of amino acids that determines their final 3D structure. The complex interplay of the different chemical properties of the different amino acids will determine a protein's thermodynamically most favorable 3D conformation. Known physicochemical forces that play a role in protein folding include the hydrophobic effect[10], H-bridges, $n \rightarrow \pi^*$

[9] The average pH inside a cell, which is approximately 7.4 and thus close to the neutral pH 7.

[10] Since water is a polar solvent, apolar molecules do not mix well with water. Hence, apolar amino acid residues will often be found on the inside of a folded protein chain, away from the water that surrounds it. This property is called hydrophobicity ("being afraid of water").

interactions, van der Waals forces, formation of disulfide bridges, the gain of conformational entropy of water on protein folding and electrostatic interactions [31]. Hydrophilic, polar amino acids will mainly be found on the surface of a folded protein, while hydrophobic, apolar amino acids will typically be present on the inside. So-called chaperones are proteins that often aid nascent proteins during the folding process[11] to avoid aberrant folding [33]. Sometimes, multiple proteins cluster together to form a functional protein complex (Fig. 1.6).

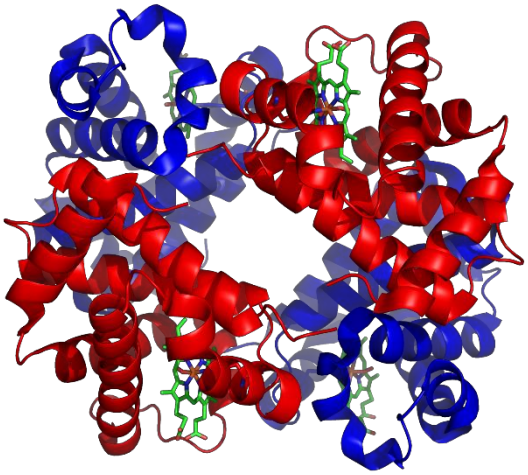


Figure 1.6. Three-dimensional structure of the protein hemoglobin. Hemoglobin consists of four folded proteins (two α subunits, red, and two β subunits, blue) that are held together by hydrogen bonds. Each subunit is folded such that it creates a pocket that strongly binds an iron-containing heme group (green). Image by Richard Wheeler (Zephyris) at the English language Wikipedia, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=2300973.

1.1.3. The JAK-STAT pathway as an example of a protein network

Cells are highly dynamic and thousands of biochemical processes are continuously going on in each cell. Proteins play a crucial role in nearly every one of these processes. The enormous diversity in 3D structures that are adopted by different proteins allows them to bind to other biomolecules with very high specificity. This in turn leads to an immense variety in protein functions. Well-known functions of proteins include, but are definitely not limited to, enzymatic reactions (e.g. trypsin, which digests other proteins in the stomach; kinases, proteins which add a phosphate group to protein substrates), DNA synthesis (DNA polymerases), DNA transcription (RNA polymerases, aided by different sorts of transcription factors), cellular structure (e.g. microtubules), muscle contraction (e.g. actin, myosin) and oxygen transport (hemoglobin).

Therefore, proteins interact both with each other and with other biomolecules. Indeed, most changes inside the cell are triggered by cascades of both stable and transient protein-protein interactions, termed signaling pathways. The JAK-STAT pathway is just one of many intracellular pathways and constitutes a classic example of how a cell responds to a stimulus coming from its environment (Fig. 1.7). JAK-STAT signaling actually is a simple signaling

[11] Note that some chaperones also work after translation. Some chaperones aid in stabilizing protein structures in response to a cellular stressor, others aid in protein unfolding or revert protein aggregation [32].

cascade and, in reality, such cascades are highly branched, as most proteins have multiple interaction partners.

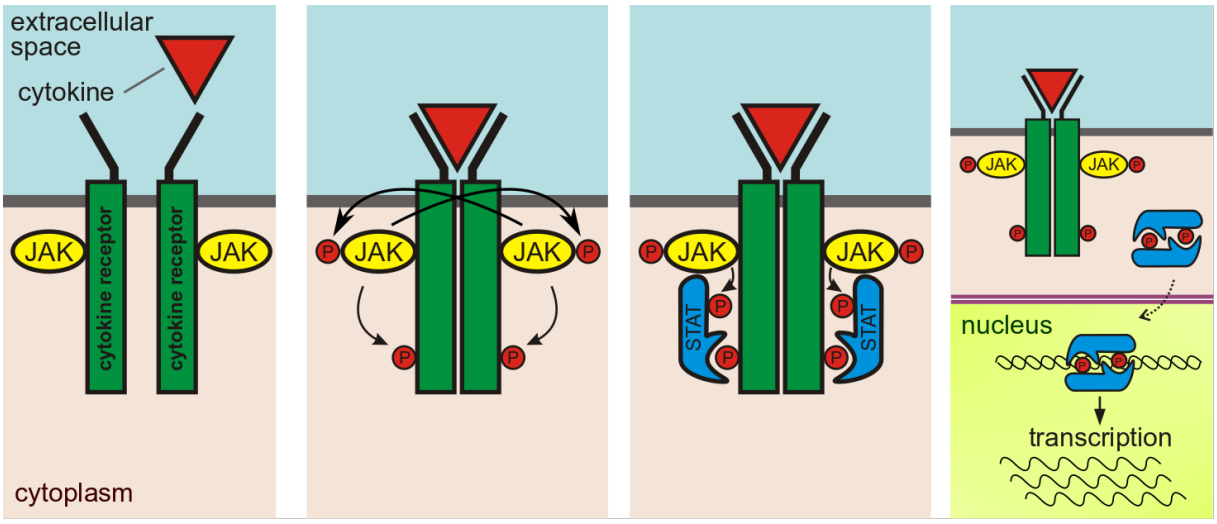


Figure 1.7. Simplified view of the JAK-STAT pathway. Certain events in the human body can trigger the release of small proteins, called cytokines, in the blood. Some cells are programmed to respond to these cytokines and are therefore equipped with specific cytokine receptors. After binding to an extracellular cytokine, the cytokine receptors dimerize, bringing the JAK kinases in close proximity of each other. The JAKs will subsequently phosphorylate each other on a Tyr residue. The phosphorylated JAKs will then phosphorylate a Tyr residue on the cytoplasmic side of the cytokine receptors. This allows docking of STAT proteins. These STAT proteins will also be phosphorylated by the JAKs. Phosphorylated STATs will dimerize and translocate to the nucleus, where they allow transcription of mRNA molecules encoding for proteins that are needed for the response to the stimulus. Modified after Peter Znamenkiy [Public domain], from Wikimedia Commons.

The JAK-STAT example involves phosphorylation as an illustration of a chemical group that is transferred by a protein (a kinase such as JAK) to another protein (a substrate, here the cytokine receptor, STAT, or JAK itself). In this example, phosphorylation on tyrosines occurs, though it can also occur on serine and threonine residues (as both contain a free hydroxyl (-OH) group) and on the nitrogen atoms of the imidazole ring in histidine residues [34, 35]. Phosphorylation on arginine, lysine, aspartate and glutamate are also known to occur, but are very labile in an acid environment and were therefore proven difficult to study by means of mass spectrometry [36, 37]. Recent work published on BioRxiv proposes a workflow at near-physiological pH that allows the identification of thousands of such non-canonical phosphorylation sites [38]. Moreover, next to phosphorylation, many more co- and post-translational modifications exist, and it is not uncommon for proteins to carry modifications across different residues. These modifications are not only important in signaling cascades, but can also affect protein stability and degradation, alter enzymatic activity and target proteins to membranes [39]. Many diseases (e.g. infectious diseases, auto-immune diseases, neurodegenerative diseases, …) have been linked to aberrant protein modification states [40-43]. A comprehensive overview of all possible protein modifications can be found in the UniMod database at: http://www.unimod.org/modifications_list.php [44].

1.1.4. Proteins in diseases

The extremely complex web of interactions of proteins with each other and with other biomolecules makes that disruption of a single protein's function often has severe outcomes [45]. Genetic diseases are often the result of a loss-of-function caused by the production of truncated or abnormally folded proteins. For instance, thalassemias are a family of genetic diseases in which an abnormal form of hemoglobin is produced that is less efficient in taking

up oxygen [46]. These diseases are caused by one or more mutations in the coding genes that lead to changes in hemoglobin's amino acid sequence. Such changes can cause substitutions of one amino acid by another but can also result in shorter or longer proteins when stop codons are respectively introduced or erased. They might also impact on mRNA splicing. Large gene deletions or insertions and even fusions with other genes have been reported to cause thalassemia [47]. Such mutations inhibit the production of one or more hemoglobin chains or result in the production of abnormally folded hemoglobin. Mutations can also cause diseases by interfering with the normal modification status of a protein. For example, the severe but rare disease mandibuloacral dysplasia is caused by a single mutation in the gene encoding for the protease ZMPSTE24, which results in the accumulation of toxic farnesylated prelamin A (Fig. 1.8).

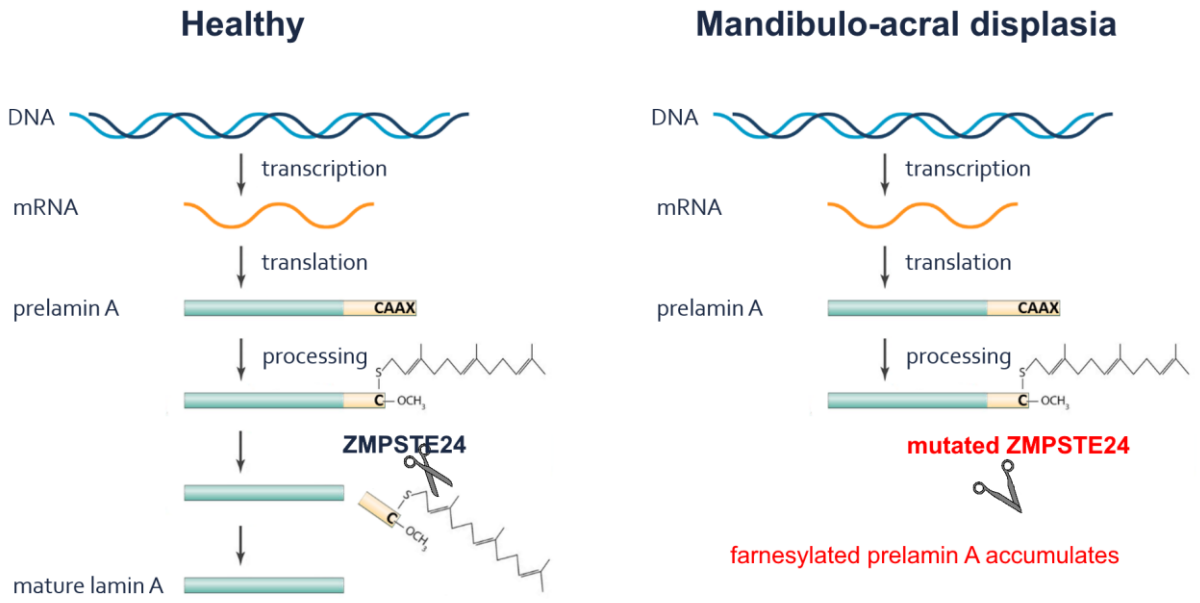


Figure 1.8. Left: overview of the maturation of the protein lamin A in healthy individuals. Normally, a hydrophobic farnesyl group is added to lamin A during its preprocessing. Lamin A's C-terminal end containing the modification is then cleaved off by the ZMPSTE24 protease (here shown as a pair of scissors). Right: in mandibuloacral dysplasia a homozygous mutation[12] in the *ZMPSTE24* gene causes loss-of-function, which results in the accumulation of toxic farnesylated prelamin A.

Conversely, for many diseases, genetics alone cannot fully explain disease onset. Late-onset Alzheimer's disease, for instance, has no single genetic cause[13]. At the protein level, it is characterized by the aggregation of the amyloid-β protein in the brain. Other incurable brain diseases, like Creutzfeldt-Jakob, are caused by a prion, an incorrectly folded protein that causes other proteins of the same kind to take over its aberrant shape, leading to some sort of a chain reaction and massive accumulation of misfolded proteins [50].

[12] Humans, like all mammals, have two copies of most genes [48]. ZMPSTE24 loss-of-function only occurs when both copies are affected.

[13] Many different genes influence susceptibility, and the overall genetic heritability is estimated between 60 and 80% [49].

1.1.5. Applications of protein research

Proteins are so abundantly being applied throughout our daily lives, that it is nearly impossible to give a complete overview of all their applications. Moreover, researchers continuously strive to improve and broaden protein applications.

In the medical field, proteins are used for diagnosis and disease monitoring. Examples of such biomarkers in blood and plasma approved by the US Federal Drug Administration (FDA) include HE4 for ovarian cancer, CA19-9 for pancreatic cancer and thyroglobulin for thyroid cancer, amongst many others [51], and researchers keep on developing novel biomarker assays.

Determining the 3D structure of proteins is pivotal to their characterization. Indeed, not only do these models allow the prediction of a protein's interactions with other proteins [52], they also aid in drug development. Nowadays, companies rationally design potential drug candidates by fitting them e.g. onto a protein's docking site [53, 54]. However, determining a protein's structure is a non-trivial task. Indeed, although the sequence of a protein will in the end determine its 3D structure, no algorithm exists that can accurately predict 3D structures solely based on amino acid sequences. X-ray crystallography and nuclear-magnetic resonance (NMR) are the most commonly used techniques to determine protein structures, although cryo-electron microscopy is also becoming a viable option [55]. Alternatively, proteins with more than 30% sequence homology are often assumed to have a similar structure [56]. Indeed, such proteins often show an evolutionary relationship and therefore hold a similar structure and function.

Proteins can also be used as therapeutics, with monoclonal antibodies forming the largest class of therapeutic proteins (48% of all FDA approvals in 2011 – 2016) [57]. Examples include antibodies against IL-5 for the treatment of asthma [58], anti-CD319 against relapsed multiple myeloma [59] and anti-VEGFR2 against gastric cancer [60]. New artificial ("recombinant") proteins can also be produced by modifying the DNA sequence of existing proteins. Examples include ocriplasmin against vitreomacular adhesion [61], glucarpidase against kidney failure [62] and recombinant von Willebrand factor against von Willebrand disease [63]. Proteins are also extensively used in basic biomedical research. Examples include green fluorescent protein (GFP) and its derivatives to visualize proteins inside a cell [64] and the use of Crispr-Cas9 to examine the function of genes by knocking them out or inducing targeted mutations [65].

However, the study of proteins is not only relevant for human diseases. Enzymes, proteins that catalyze biochemical reactions[14], are used in sectors as diverse as the pharmaceutical sector, the food industry, paper production, detergent manufacturing and biofuel production [67]. For the production of pharmaceuticals, enzymes aid in the production of precursors or in chemically modifying the final compounds to increase their stability and/or bioavailability [68]. In the food industry, α-amylase is used to convert starch into sugars [69], pectinase to clarify fruit juices [70] and lactase to produce lactose-free milk [71], amongst many others. In the paper industry, xylanase is used to loosen the structure of cellulose fibers, which improves paper quality [72]. Proteases, amylases and lipases are used in laundry detergents to help break down stains of biological origin [73]. Lipases are used in biofuel production to convert free fatty acids to methyl/ethyl esters [74].

[14] This acceleration often goes up to several trillions of orders of magnitude [66], allowing reactions that would naturally take millions of years to occur almost instantly. Previously mentioned proteins such as JAK and ZMPSTE24 are also enzymes: JAK catalyzes a phosphorylation reaction, ZMPSTE24 cleaves a peptide bond.

Finally, proteins are also extensively investigated in food crop research. The most well-known example is research on transgenic plants. Here, a DNA sequence coding for a protein with favorable properties is introduced into a commercial crop. This protein can for example promote crop yield or convey resistance against insects, pathogens or herbicides. A notorious example is the development of “golden rice”, a genetically engineered rice variant developed to combat vitamin A deficiency [75]. Indeed, per gram dry weight, the seeds of golden rice contain up to 37 µg β-carotene, a compound that is converted into vitamin A by the human body [76]. Although the rice genome can produce all enzymes needed for β-carotene production, four of these enzymes are not expressed in rice seeds. In golden rice, only two genes are introduced: the *psy* gene from maize, to produce the enzyme phytoene synthase and the *crtI* gene from the bacterium *Erwinia uredovora* to produce the enzyme carotene desaturase. Together, these enzymes restore the β-carotene pathway in the seeds, which results in a rice plant with the typical yellow seeds. Thanks to this simple genetic modification, golden rice holds the promise to save a substantial fraction of the 250,000 to 500,000 children that become blind every year due to vitamin A deficiency, half of which die within a year [77-80]. This is especially the case in Asian countries, where rice is a dominant portion of the standard diet.

Understanding how proteins interact with each other increases our understanding of a plant’s developmental pathways, which allows the breeding of high-yield variants as well as variants that produce stable yields under stress [81]. Indeed, various biotic and abiotic stresses can delay or even terminate plant growth, and even a relatively small, transient stress can markedly reduce crop yield [82]. Hence, researchers actively investigate protein networks involved in stress responses to explain why certain varieties are less stress-sensitive [83, 84]. This knowledge can then later be used for crop improvement through cross-breeding or genetic modification. The lab of my co-promoter is, amongst others, involved in research into protein signaling pathways during the germination of parasitic plants. This is expected to spur the development of germination inhibitors for these plagues [85].

1.2. The nature of mass spectrometry-based proteomics

Researchers need to obtain information about the proteome to understand and act upon all these protein-related processes. Proteomics is the study of the proteome and today, MS-based proteomics is the most important proteomic technology. Here, I start by giving a very brief overview of the general principles of chromatography and MS, followed by a discussion of MS-based proteomics workflows. Then, I will situate the proteomics field with respect to the other omics fields.

1.2.1. General principles of liquid chromatography and mass spectrometry

Like every other analytical technique, mass spectrometry has its limitations on the complexity it can efficiently cope with, both in terms of the number of distinct analytes and in terms of differences in the concentrations between these analytes. Separating analytes prior to further analysis is thus essential for reaching sufficient analytical depth. For MS-driven proteomics, the analytes – which are mainly peptides – are present in solution. Hence, liquid chromatography is the method-of-choice for separating peptides prior to analysis by means of mass spectrometry. In liquid chromatographic applications for proteomics, the peptides are first loaded onto a column (also called the stationary phase) in a buffered solution (also called the mobile phase). The composition of the column is such that most (if not all) of the peptides interact with and are thus withheld by this column. By now changing the composition of this mobile phase, the column-bound peptides will start to partition in the mobile phase and are thus eluted from the column at a given composition of this mobile phase. Liquid

chromatography can be used to separate peptides based on different physical characteristics such as size, charge and hydrophobicity. Separating peptides based on differences in hydrophobicity is the preferred chromatographic method that is linked to mass spectrometers. In the overall majority of applications, a hydrophobic stationary phase (e.g., chromatographic beads functionalized with C18-groups) is used to bind peptides via hydrophobic interactions. Here, the buffer used to load the peptides is an aqueous buffer. By now gradually increasing the concentration of a water-miscible organic solvent, peptides will start to favor being present in the increasingly organic mobile phase and elute from the stationary phase.

To further analyze the eluted peptides, these peptides need first to be brought into the gas phase and need to get ionized. This process happens in the ionization part of a mass spectrometer. “Soft” ionization techniques, i.e. ionization techniques that transfer little residual energy onto the ions and therefore cause only minimal ion fragmentation have been extremely important for the measurement of intact ions. Indeed, John B. Fenn and Koichi Tanaka were awarded the 2002 Nobel Prize in Chemistry for the development of electrospray ionization (ESI) and soft laser desorption (SLD), respectively. Next to an ionization part, a mass spectrometer also consists of at least one analyzer and at least one detector.

Different types of analyzers have been developed to separate and detect ions based on their mass-to-charge (*m*/*z*) ratio. In a time-of-flight (TOF) mass spectrometer, ions are accelerated by a fixed electric field into a vacuum tube. The time it takes for an ion to travel (or fly) through this tube and reach a detector depends on its mass and its charge. Indeed, given the formula of kinetic energy, the higher the mass of an ion, the lower its velocity and thus the later this ion will hit the detector. Conversely, the higher the charge of an ion, the higher its velocity and thus the faster it will hit the detector. The intensity of the signal recorded by this detector is used as a proxy for the initial abundance of the recorded peptide. These signals are recorded at discrete *m*/*z*-values at GHz resolution.

A quadrupole mass analyzer is an example of a more complex analyzer that works as an *m*/*z* filter. Quadrupoles consist of four rods, which have an alternating radio frequency voltage with an offset direct current. The frequency voltage and the offset can be tuned in such a way that only ions with a specific *m*/*z* value follow a stable trajectory throughout the quadrupole, while all other ions are pushed out (Fig. 1.9). When the quadrupole is used to scan a beam of ions over a certain *m*/*z* range, a mass spectrum is recorded.

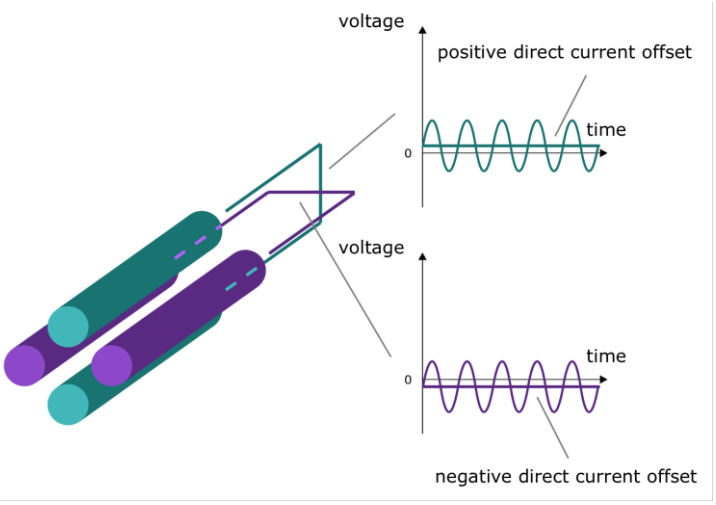


Figure 1.9. Working principle of a quadrupole. The rods have an alternating radio frequency voltage with a direct current offset. Due to inertia, ions with very high *m*/*z* values will be relatively unaffected by the alternating current but will be pushed out of the quadrupole by the non-zero direct current offset. Contrary, ions with very low *m*/*z* values will be pushed out of the quadrupole as they are strongly affected

by the alternating current. By carefully tuning the alternating and direct currents, a quadrupole works as a very specific ion filter. Figure based on Vékey *et al.* (2008) [86].

1.2.2. The MS-based proteomics workflow

Mass spectrometry-based proteomics is the method of choice for the high-throughput identification and quantification of peptides and proteins in a single analysis. Fig. 1.10 gives a general overview that encompasses the most frequently occurring steps in MS-based proteomics workflows.

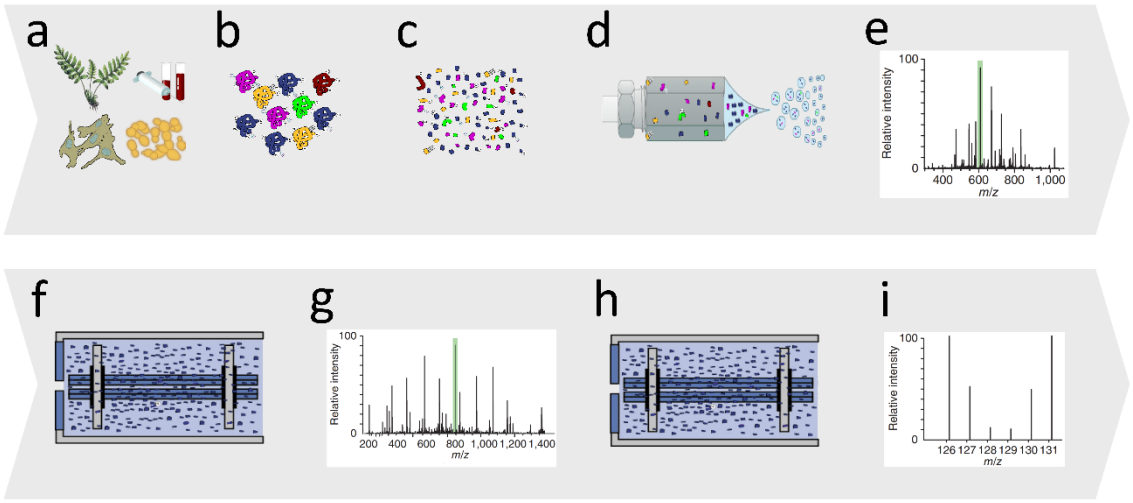


Figure 1.10. General overview of an MS-based proteomics workflow. The workflow starts with samples (a) from which proteins (b) are extracted. These proteins can e.g. be labeled and are, for most applications, digested into smaller fragments called peptides (c). The proteins/peptides are separated onto a high-performance liquid chromatography (HPLC) column and ionized (d). Then, an MS spectrum (e) is recorded. Some ions can be fragmented into smaller ions (f) after which the resulting spectrum is recorded (MS^2 spectrum, g). Some of these fragments can again be targeted for fragmentation (h), after which an MS^3 spectrum can be recorded (i). Note that the presence or absence of some steps in this workflow depend on the type of analysis that is performed.

Every MS-based proteomics workflow starts with one or more samples. The types of samples can vary widely, ranging from plant material over animal tissues, plasma samples, cell cultures or recombinantly produced proteins. If the sample is not a solution of (purified) proteins, the proteins will need to be extracted first. The techniques used for extraction and purification depend on the sample type. For mammalian cell cultures, for example, a rather simple lysis buffer will suffice [87]. For organisms with thick cell walls or for membrane proteins, (additional) mechanical disruption, e.g. sonication with glass or metal beads, might be necessary [88]. Techniques like ultracentrifugation or ammonium sulfate precipitation are then used to separate the proteome from other unwanted cellular components (e.g. DNA, RNA, lipids) and possible detergents that were used to disrupt cells. Solubilized intact proteins can be directly analyzed by MS (top-down proteomics) [89], but complex protein mixtures are generally enzymatically digested into smaller fragments, termed peptides, for so-called bottom-up proteomics or shotgun proteomics[15].

[15] Note that the term “bottom-up proteomics” refers to any LC-MS proteomics technique that uses prior digestion of proteins to peptides, while the term “shotgun proteomics” specifically refers to LC-MS proteomics techniques whereby complex protein mixtures are digested into peptides.

Peptides have the advantage that they are more chemically tractable, more easily separated by liquid chromatography and more easily ionized and fragmented as compared to intact proteins [90-92]. To limit the number of possible peptides in bottom-up proteomics to a reasonable computational search space, the commonly-used protease is highly specific. Trypsin, for instance, is ideally suited, as it cleaves with high specificity after lysine and arginine residues.

Alternative enzymes (e.g. pepsin, chymotrypsin, and the endoproteinases LysC, LysN, AspN, GluC and ArgC) have also been used, as they generate different sets of peptides and therefore reveal complementary parts of a proteome's sequence space [93-99]. Parallel digestion of the same proteome with multiple proteases can also give a strong boost to the coverage of modification sites [100, 101]. Alternative proteases with more infrequent cleavage specificities will generate longer peptides that can be studied with middle-down proteomics [92, 102, 103]. Nonetheless, trypsin remains the dominant digestion enzyme in bottom-up proteomics. On November 2014, more than 96% of all raw files deposited in the PRIDE repository were using trypsin [102] and there is little reason to assume this percentage has drastically changed today.

To facilitate proteolytic digestion, proteins are often first denatured with urea, which disrupts a protein's hydrogen bonds causing the protein to denature and unfold. This results in a destruction of protein-protein interactions and a solubilization of hydrophobic lipid-bilayer bound proteins. Urea has the advantage that it can easily be removed by reverse-phase chromatography [104]. However, adding too much urea might also partially denature the protease and hence reduce the digestion efficiency. Further, at higher temperatures, urea partially decomposes into isocyanic acid which carbamylates primary amine groups and thus introduces artefactual amino acid modifications that also block enzymatic digestion [105, 106]. Note that detergents such as sodium dodecyl sulphate (SDS) that are commonly used to lyse cells or to denature proteins prior to polyacrylamide gel electrophoresis (SDS-PAGE), should generally be avoided as they are incompatible with liquid chromatography (LC)-MS [91]. Indeed, even though small amounts of SDS facilitate enzymatic digestion, this detergent suppresses ion signals even at very low concentrations (< 0.01%). SDS cannot be removed with a reverse-phase high-performance liquid chromatography (HPLC) separation step [106-109], although the recently-introduced suspension trapping filter S-Trap^{TM} includes a washing step that makes it compatible with SDS denaturation [110]. Nevertheless, most digestion protocols omit the use of denaturants [111]. Other contaminants may also adversely affect the analysis [112]. It is therefore strongly advised to discuss all preprocessing protocols with proteomics specialists prior to the experiment.

If one is interested in only a subpart of the proteome, additional purification of certain proteins or peptides, e.g. by immunoprecipitation might be needed. Similarly, many protein modifications are transient, have low occupancies, are chemically unstable during standard sample preparation procedures and/or decrease a peptide's ionization efficiency [113-116]. Therefore, detecting specific modifications requires optimized enrichment protocols [117].

Proteins or peptides are sometimes labeled to facilitate identification and quantification (see section 2.1). These labels can either be small chemical groups or heavy isotopes. They can either be incorporated during the growth of the organism (metabolic labeling, see 2.1.1) or after protein extraction (post-metabolic labeling, see 2.1.2).

Sample pre-fractionation is an option when samples are very complicated and enough protein material is available [118]. Pre-fractionation can be done with a method that is orthogonal to the standard reverse-phase liquid chromatograph that is coupled to the mass spectrometer. Strong cation exchange (SCX) chromatography is a popular pre-fractionation strategy [119]. Although the complexity of each fraction will be reduced, separation is never perfect and many

proteins will be present in more than one fraction, which may complicate protein quantification [120]. Each of these samples (or fractions in the case of pre-fractionation) is subsequently analyzed by the mass spectrometer. If the samples are very simple (e.g. a single protein), the proteins/peptides can be directly analyzed by matrix-assisted laser desorption (MALDI) ionization coupled to a time-of-flight mass spectrometer, by which a mass spectrum for the entire sample is obtained [121]. If the samples are more complex, the proteins/peptides are first separated onto a reverse phase liquid chromatography column that is coupled to the mass spectrometer. Upon elution, the proteins/peptides are ionized, typically with electrospray ionization (ESI) [122]. Traditionally, positively charged ions are generated, while neutral molecules and negatively charged ions are filtered out [123]. At discrete time points, the mass spectrometer will measure the mass-to-charge ratios for all the ion species eluting from the column. In the commonly used Orbitrap analyzer, this is achieved by trapping the ions in an orbital motion around a spindle-like electrode, hence the name. The ions are moved back and forth and the fluctuations in charge caused by the movement of the ions are recorded by a detector. This wavelet signal is subsequently converted into a mass spectrum by Fourier transformation. The resulting spectrum is termed an MS or MS^1 spectrum. It is important to note that due to the natural occurrence of heavy isotopes of all chemical elements in a fixed ratio, every ion species generates multiple isotopic peaks in the MS spectrum, resulting in a so-called isotopic envelope. Each ion's charge state is then calculated from the m/z-distance between the peaks in such an isotopic envelope. The summed-up intensities of all ions in an MS spectrum is called the total ion current (TIC). The evolution of the TIC over time is used as a measure of quality control.

In most workflows, an MS spectrum will be insufficient to identify the ion species. Therefore, a single peak in the isotopic envelope of the ions of interest will be selected by the mass spectrometer and targeted for fragmentation. Typically, high intensity peaks are targeted for fragmentation to avoid selecting noise, which would lead to significant losses in operating time. To avoid sequential fragmentations of the same ion during its elution, all previously targeted m/z values are often excluded from being targeted again for a certain amount of time (e.g. 20 seconds). This setting is termed dynamic exclusion and substantially increases the coverage of the mass spectrometer by freeing MS time for the targeting and fragmentation of less intense MS peaks [124].

In shotgun proteomics, collision-induced dissociation (CID) [125] or higher-energy collisional dissociation (HCD) [126] are by far the most common fragmentation methods, while electron-transfer dissociation (ETD) gains popularity for phosphoproteome studies (see 1.3.2) [127, 128]. Negative electron-transfer dissociation (NETD) [123] and photo-dissociation [129-132] are examples of infrequently used fragmentation methods. With CID, ions are collided with noble gasses such as helium and argon that increase the ions' internal vibrational energy and eventually lead to fragmentation [133]. With HCD fragmentation, the collision energies are higher than 1 keV [134]. Collision therefore occurs in a separate collision cell, typically with a heavier gas, such as dinitrogen [135]. If the peptide's backbone is fragmented, six types of ions (a, b, c, x, y and z) can be formed, as shown in Fig. 1.11. The mass spectrum of the fragment ions is termed an MS/MS or MS^2 spectrum.

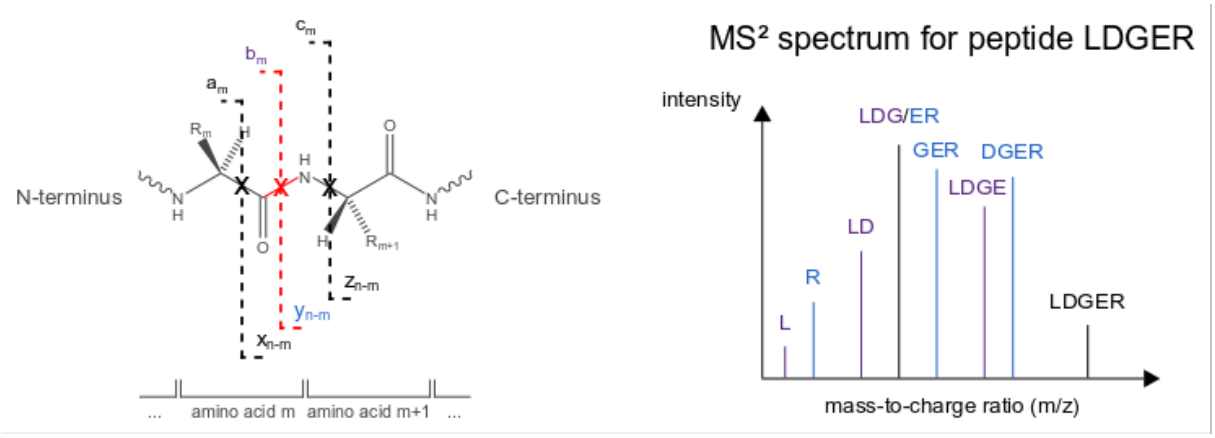


Figure 1.11. Left: overview of the different types of ions that can be formed after fragmentation of a peptide ion's backbone. a-, b- and c-ions are formed if the charge is retained on the N-terminal peptide fragment. Conversely, x- y- and z- ions are formed if the charge is retained on the C-terminal fragment. Breaking the peptide bond (red) is by far the most energetically favorable fragmentation pattern. Therefore, b- and y-ions will be the most abundant ion species in every MS² spectrum generated by CID or HCD. n is the total number of amino acids in the protein. Modified after Steen and Mann (2004) [136]. Right: example fragmentation pattern for the peptide LDGER. b-ions are indicated in purple, y-ions are indicated in blue. Fragments LDG and ER have the same average mass and therefore form a single peak in the MS² spectrum.

With CID and HCD, b- and y-ions will be far more intense than the other types of ions (a-, c-, x- and z-ions) [137]. For tryptic peptides, CID fragmentation generates both b- and y-ions, while y-ions are much more prominent with HCD fragmentation [137-139].

When the intensities of certain fragment ions are to be used for quantification, MS² fragment ions can optionally be isolated and subjected to mass spectrometry (MS³ spectrum) to prevent interference of other fragment ions with the intensity of the fragment ion of interest. Such MS³ workflows are mainly useful for isobaric labeling (see 2.1.2). Identification of an ion species is typically achieved by searching the fragment ion spectra against a database (see section 3.1).

Note that MS instrumentation is evolving extremely rapidly. Only a few years ago, 20 MS² spectra per second was considered state-of-the-art [140], but the most recent machines now exceed a scanning speed of over 40 spectra per second [141]. This implies that if one MS spectrum is recorded per second, more than 40 peaks in this MS spectrum can be fragmented and thus potentially identified.

1.2.3. Proteomics in relation to other omics

Many omics techniques can provide some information about the state of the proteome, but proteomics is the most suited for this purpose. The fields of genomics, epigenomics, transcriptomics and ribosome profiling largely rely on the analysis of DNA and RNA molecules. RNA can easily be reverse transcribed to DNA and even a single molecule can be readily amplified to billions of copies with a routine polymerase chain reaction (PCR). Moreover, advanced techniques have been devised to sequence and characterize even single RNA and DNA molecules. These possibilities are currently lacking for other types of biomolecules, including proteins. Present-day proteomics relies heavily on mass spectrometry and with the ever-increasing resolutions and operating speeds of contemporary mass spectrometers, the proteomics field has evolved rapidly over the past few years [142, 143]. Other mass spectrometry-based omics fields, such as lipidomics and glycomics, are considered to be still in their infancies.

Since the completion of the Human Genome Project, we have an adequate view on most of the protein-coding genes[16] and the genomes of more and more organisms are almost routinely being added to the ever expanding genome databases [145-149]. However, the genome only provides a view on which proteins can potentially be expressed. Indeed, a retina cell is very different from a muscle cell, despite sharing the same genomes. Similarly, a caterpillar is very different from a butterfly because they both activate different genetic programs. Thus, the genome alone provides little information about the current proteomic state of a cell.

Epigenomics is the study of the modifications (e.g. methylation, acetylation) of the DNA and its associated proteins (histones amongst others) [150]. These chemical markers change the chromatin's structure and dictate the accessibility of each gene for the transcription machinery. Within the same species, different cell types have different epigenetic patterns. However, epigenomics only provides an overview of how easily genes can be accessed, but it does not provide answers to how much of each proteoform is actually being produced.

Transcriptomics is a routine technique to quantify the amounts of (m)RNA molecules derived from each gene [151]. It can therefore be used as a rough estimate for protein production. However, a substantial amount of mRNA is not translated and there is a variety of mechanisms that modulate protein synthesis at the translation step[17]. The ribosome profiling technique provides a quantitative snapshot of which mRNA is getting translated. Therefore, ribosome profiling provides a much better estimate of protein translation than mRNA sequencing [153].

Metabolomics and lipidomics, the large-scale analyses of metabolites and lipids respectively, can elucidate complex metabolic processes related to health and disease [154]. Metabolic conversions are not only catalyzed by enzymes, but also closely regulated by various protein signaling cascades. Therefore, metabolomics and lipidomics can provide additional indirect information on the state of the proteome [155].

However, none of these omics' techniques are able to determine which proteoforms will be produced, nor can they provide any information about protein degradation, the other side of the balance that governs protein steady-state. They are also not well-suited to assess protein localization and are unable to provide information about a protein's interaction partners. The only technique that is able to assess with high-throughput the properties of a protein within a proteome is MS-based proteomics.

1.3. Applications of mass spectrometry-based proteomics

MS-based proteomics is the preferred method for solving numerous biological questions from the "proteome angle". Identifying and later also quantifying proteins were and are still the main initial applications of MS-based proteomics. Over the last decade, the application of proteomics for the identification and quantification of protein modifications has also gained significant attention. The proteomics field has also been expanded to study amongst others, protein-protein interactions [156-159], protein-compound interactions [160-163], cellular protein localization [164-167] and protein structure determination [89, 160, 168-173]. However, the identification and quantification of proteins remains the most important application of MS-based proteomics and many of these applications rely on the quantitative ability of the mass spectrometer. Since my thesis focuses on protein quantification, I will elaborate here on protein quantification and the quantification of protein modifications.

[16] Although the detection of many small "hidden" proteins remains challenging [144].
[17] These include, amongst others, RNA splicing, reading frame shifts, translational read-through, sequestration of mRNA and mRNA degradation [152].

1.3.1. The analysis of protein and peptide abundance

After a peptide or protein is identified, quantification is the next logical step. Typically, peptides from different biological conditions are labeled either isotopically or chemically in order to induce a mass shift in the MS, MS^2 or MS^3 spectrum. Alternatively, if labeling is omitted (label-free proteomics), mass spectra from different runs should be compared. The peak intensities are a proxy for peptide (and hence a protein) abundance. The peak intensities registered during the elution of a peptide therefore allow for protein quantification in each biological condition. Alternatively, quantification can be done by the less accurate, but simpler spectral or peptide counting. Relating intensities to protein concentrations is challenging because peptides can have very different ionization efficiencies, which are difficult to predict. Nonetheless, there were some attempts at absolute protein quantification [174, 175]. Since the analysis of protein abundance is the focus of my work, I have kept this section intentionally brief since more details about quantitative proteomics are given in chapters 2-4.

1.3.2. The analysis of protein modifications

Both bottom-up and top-down proteomics can be employed to study protein modifications as these cause shifts in the masses of affected peptides that can readily be detected.

As mentioned earlier under 1.1.3, proteins can carry a plethora of modifications. Most of these modifications occur too infrequently to be detected or prove to be very chemically labile. Other modifications, such as the frequently occurring, but often labile phosphorylation, are also difficult to detect because of the poor ionization of the phosphoryl group, which carries a negative charge. Therefore, when the main research goal is to study a certain protein modification, specific enrichment procedures are often required [117].

Phosphorylation is both the most common and the most intensively studied protein modification [176]. Indeed, phosphorylation is a key modification in many signaling cascades and phosphoproteomics has contributed enormously to our understanding of these pathways [177-179]. However, the negative charge of the phosphoryl group and the default usage of positive electrospray ionization makes it difficult to generate and therefore detect positively-charged phosphorpeptides. Thus, phosphopeptides are enriched via pre-fractionation with hydrophilic interaction liquid chromatography (HILIC), Strong cation exchange (SCX) chromatography or strong anion exchange (SAX) chromatography, as well as immunoprecipitation, immobilized metal affinity chromatography (IMAC), metal-oxide affinity chromatography (MOAC), Phos-Tag chromatography, polymer-based metal ion affinity capture (PolyMAC), hydroxyapatite chromatography, enrichment by chemical modification, and/or phosphopeptide precipitation [180]. Negative electron-transfer dissociation (NETD) can also increase the MS intensities of phosphopeptides [123].

Acetylation and methylation are modifications that were first discovered on histones, proteins that are associated with the DNA and package it into structural units called nucleosomes [181, 182]. Histone acetylation is generally associated with open chromatin [183], while histone methylation, dependent on the location of the methylation site, can both be associated with open and closed chromatin [184]. Acetylation can occur both on protein N-termini and on lysine residues [183], while methylation can also occur on arginine residues [184].

Acetylation is important for protein localization [185-188], protein folding [189], protein stability [190] and interactions with other proteins [191]. Methylation is a major factor in important cell signaling pathways such as JAK-STAT, MAPK, WNT, Hippo and BMP [192] and overall protein methylation seems to be strongly intertwined with the organism's metabolic state [193].

Again, enrichment procedures by e.g. (immuno)affinity purification or chromatography are preferable when assessing acetylation or methylation on a proteome-wide scale [194-196].

Glycosylation is another very important post-translational protein modification as over 50% of all mammalian proteins are glycosylated [197][18]. Glycosylation plays an important role in protein structure and stability [199-201]. The presence of a glycan structure can block other modifications such as phosphorylation [202]. Glycan structures can be sensed by other proteins (leptins) and play an important role in e.g. cancer and immunity [203, 204]. They are also intensively studied in the context of therapeutic proteins, as many of these proteins carry glycan structures [200, 205]. The enormous complexity of many glycan structures has spurred the development of a new field called glyco(proteo)mics [206].

Ubiquitin is a small 8.6 kDa protein that can be covalently linked to lysine residues of other proteins. However, non-canonical ubiquitination can also occur at Ser, Thr and Cys residues and at a protein's N-terminal amine group [207]. The earliest known function of K48-linked[19] polyubiquitination is the degradation of the modified proteins in the proteasome. This discovery resulted in the 2004 Nobel Prize in Chemistry for Avram Hershko, Aaron Ciechanover, and Irwin A. Rose [208]. Although K11-linked polyubiquitination can also result in proteasomal degradation [209], ubiquitin is also an important scaffold that allows recruitment of other proteins and plays an important role in cellular processes as diverse as e.g. protein trafficking, mitophagy and cell cycle control. Next to K48 and K11; K6, K27, K29, K33 and K63 linkages have also been described. Many proteins are mono-ubiquitinated, but ubiquitin chains can also be linear and branched and even combined with other small ubiquitin-like modifiers such as SUMO and NEDD-8 [210].

Detection of ubiquitination by MS is challenging because ubiquitin is often either quickly degraded by the proteasome or part of very dynamic signaling pathways [211]. Moreover, ubiquitin is much bigger than most other small modifications. Being a protein, ubiquitin is also degraded by trypsin, leaving only a small GG or LRGG tag behind [212, 213]. Finally, ubiquitination does not seem to occur on well-defined amino acid sequence motifs [214-216]. Nonetheless, protocols for proteome-wide detection of ubiquitination are now readily available thanks to innovative purification and chemical tagging strategies [211, 217]. Similarly, protocols have also been developed for the proteomic analysis of other less characterized small ubiquitin-like modifiers like SUMO [218, 219], NEDD-8 [220, 221] and ISG [222, 223].

[18] Note that not only proteins, but also lipids can be glycosylated [198].
[19] K48-linkage means that each ubiquitin in the chain is linked to the previous ubiquitin via the lysine (K) residue at position 48 (starting from the N-terminus).

2. TECHNICAL CONTEXT

As amply demonstrated in the previous chapter, mass spectrometry-based proteomics has become an invaluable tool for protein researchers. In chapter 2, I will discuss the different flavors of mass spectrometry-based proteomics with a focus on their technicalities. I will specifically emphasize label-free shotgun proteomics, as all my PhD work revolves around this particular technique. This technical overview will provide a handle for chapter 3, where I will link the statistical challenges to the technology.

2.1. Label-based mass spectrometry-based proteomics

Traditionally, quantitative proteomics has made use of stable-isotope coded labels. This used to be a necessity, as the peptides that were identified often differed strongly between LC-MS/MS runs. Indeed, even minor differences in electrospray voltages and/or chromatographic flow rates increase the run-to-run variability in signal intensities and hence reduce the precision of label-free quantification [224, 225]. By metabolically labeling proteins from different samples and subsequently pooling these together for a single analysis, it became possible to identify the peptides in each sample analyzed and remove the inter-sample variability. In fact, one distinguishes two categories of label-based proteomics: metabolic and post-metabolic labeling.

In metabolic labeling, labels are incorporated during cell or organism expansion. By contrast, in post-metabolic labeling, a label is added after protein extraction or even after protein digestion, typically by means of a chemical reaction that targets specific reactive groups in proteins or peptides. The main advantage of metabolic labeling is that the labeled proteins can immediately be pooled together, thus any random or systematic experimental errors that occur after pooling will affect all samples equally [226]. Thereby, this unwanted variability can be factored out from the analysis, which increases the overall precision. Post-metabolic labeling, by contrast, does not affect the biology of the organism under study (see below), as labeling only occurs after protein extraction. And, although post-metabolic labeling requires more protein material, it is more broadly applicable [227]. Fig. 2.1 gives an overview of the experimental stages in which labels are introduced and samples are mixed for different labeling protocols.

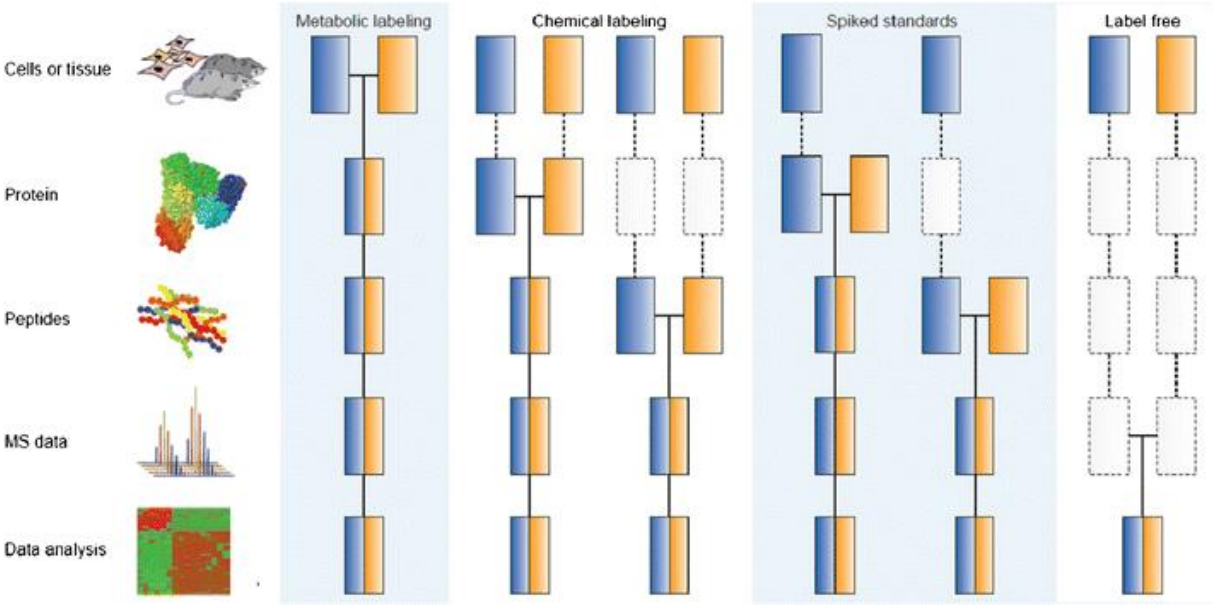


Figure 2.1. Schematic overview of the labeling workflow for two experimental conditions (blue and yellow). Horizontal lines indicate when samples are combined, dashed lines indicate the protocol steps where experimental variation unequally affects both samples. Reprinted with permission from Bantscheff *et al.* (2012) [228], copyright © 2012, Springer-Verlag.

2.1.1. Metabolic labeling

In metabolic labeling, a cell culture or even an entire organism is expanded in a medium that contains nutrients composed of heavy isotopes. Metabolic incorporation of radioactive isotopes was first used to quantify proteins in gels [229], whereas ^{15}N labeling was the first MS-based quantitative proteomics approach [230]. Here, cells were grown in a medium with nutrients containing the stable ^{15}N isotope, which allows this isotope to be incorporated into the biomolecules of the cell. The heavy-labeled proteins have higher masses compared to proteins from cells expanded in a medium containing the natural ^{14}N isotope. This results in a mass shift for each peptide that is detected in the mass spectra. ^{15}N labeling is thus well-suited to compare the levels of all identified proteins between two conditions: one grown in normal ("light" or "unlabeled") medium and the other in heavy medium. Indeed, peptides originating from digesting both proteomes are pooled and analyzed by mass spectrometry. The intensities of the ^{14}N and ^{15}N peptides are compared to infer differences in protein abundances. This approach assumes that cellular metabolism remains unaffected by the heavy isotopes. Nonetheless, several studies indicate that there might be biological effects of heavy isotopes given differential preferences of enzymes for certain isotopes [231-234]. To guard against statistical confounding due to these kinetic isotopic effects and unavoidable differences in quality[20] between the light- and the heavy-labeled medium, the heavy and light conditions are routinely swapped in experimental repeats. ^{15}N labeling was originally used to quantify baker's yeast (*Saccharomyces cerevisiae*) proteins and was later also applied on bacterial and mammalian cell cultures [235], and even on whole organisms [236-238].

[20] E.g. small deviations from the stated isotopic content might bias quantifications. Also, since the heavy and light media are stored in separate bottles, there might be small differences in biological effects since fetal bovine serum, an essential serum supplement for most cell cultures, is often added separately to each bottle. Moreover, the quality of both media might also diverge over time.

Incomplete labeling is a major disadvantage of all metabolic labeling strategies. Indeed, natural medium still contains a non-negligible albeit very low amount of heavy isotopes (e.g., 1% ^{13}C and 0.4% ^{15}N) [239]. Similarly, medium highly enriched in heavy isotopes still contains a fraction of lighter isotopes. As peptides differ in their number of amino acids and different amino acids have different numbers of atoms, isotopic labeling will generate a plethora of different partially unlabeled variants for every peptide analyzed [240]. Therefore, the mass shift of a peptide after labeling depends on its composition, which complicates downstream peptide identification and quantification.

Stable Isotope Labeling by Amino acids in Cell culture (SILAC) is a metabolic labeling approach where amino acids, typically essential amino acids, containing one or more stable heavy isotopes are used for differential protein quantification [241, 242]. The use of amino acids as opposed to isotopically labeled nutrients has greatly simplified data analysis for MS-based protein quantification, and this because the labeled amino acids are incorporated into proteins, implying that a peptide's mass shift can be directly derived from its amino acid composition. An important point is that the organism in the heavy condition needs to be exposed to the heavy-labeled amino acids for long enough to allow the complete replacement of the organism's natural (unlabeled) amino acids by the supplied heavy amino acids. For cell cultures, seven doubling times seems to be sufficient to allow full incorporation, even for proteins with very slow turnover rates [240]. By contrasting light (e.g. $^{12}C_6{}^{14}N_4$-arginine), medium (e.g. $^{13}C_6{}^{14}N_4$-arginine) and heavy (e.g. $^{13}C_6{}^{15}N_4$-arginine) SILAC labeling, three different conditions can be directly compared [243]. In 2010, the introduction of 5-plex SILAC allowed the comparison of up to five different conditions in a single MS run [244]. When more than five conditions need to be compared with SILAC, a heavy-labeled standard proteome can be spiked into every condition and be used as a reference to allow calculation of the protein ratios for every comparison (super-SILAC) [245] [246]. SILAC-labeling was initially confined to cell cultures, but over time, it has been expanded to organisms such as the worm *Caenorhabditis elegans* [247], the fruit fly *Drosophila melanogaster* [248], mice [249] and plants [250]. Fig. 2.2 gives an overview of a typical SILAC workflow.

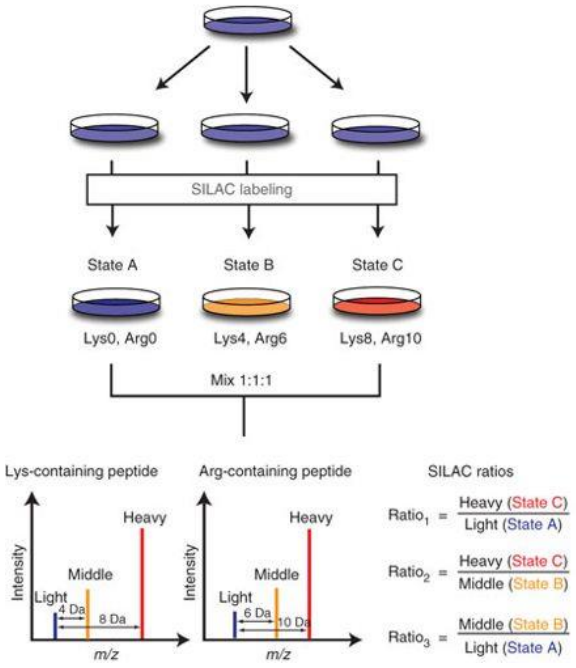


Figure 2.2. Overview of a SILAC workflow where three experimental conditions need to be compared. The proteomes from the light, medium-labeled and heavy-labeled conditions can be pooled together and analyzed in a single MS run. The intensity ratios of the triplets for each peptide ion species can then

be directly calculated from its corresponding MS spectrum. Reprinted with permission from Geiger *et al.* (2011) [251], copyright © 2011, Springer-Verlag.

Also, the isotopic purity of the SILAC medium is of utmost importance. When the fraction of "heavy" amino acids is insufficiently high, the hence-labeled proteome contains a substantial fraction of "light" amino acids, leading to quantification biases in the direction of the "light" condition [240]. Further, when using trypsin to digest an arginine-SILAC labeled proteome, a possible caveat is introduced by the metabolic conversion of arginine into proline and glutamate [252], which introduces extra stable isotopes in proline-containing peptides, thereby obstructing accurate quantification. This can be prevented by adding an excess of unlabeled proline [253, 254] and/or by reducing the arginine concentration [255].

SILAC requires dialyzed fetal bovine serum as no other variants of the selected essential amino acids other than the labeled variants must be present in the medium. Dialysis however results in the loss of growth factors, which will cause a cellular stress response that might bias the response to the treatment of interest or even completely prevent cell or organism growth altogether [227, 256].

Despite its main advantage of factoring out run-to-run variability from each comparison, SILAC seems to be slowly getting out of favor for high-throughput protein quantification. The main reason is that the quantification depth (i.e. the number of proteins identified) of SILAC is reported to be ~30 to 60% lower than the quantification depth of label-free quantification (see 2.2.1) [257, 258], although this difference also depends on sample complexity and instrument resolution. Indeed, as multiple SILAC-labeled samples are jointly analyzed in a single MS run, the amount of protein analyzed per experimental condition is two to five (for 5-plex SILAC) times lower, while the number of peaks in each MS spectrum increases with the same factor [258, 259][21]. In such complex spectra, the peptide ion signals will be lower and might even become indiscernible from the background noise. Also, the more peaks, the higher the chances of co-fragmentation: i.e. two (or more) peptide ions with overlapping isotopic envelopes are fragmented together, increasing the risk of unidentifiable spectra [260]. Moreover, as all isotopic variants of the same peptide can be targeted for identification, less MS peaks might be identified due to limitations on the MS^2 sampling rate [261]. Another disadvantage of metabolic labeling is that its dynamic range is generally smaller than those of label-free and isobaric quantification approaches (see 2.1.2 and 2.2.1) [262-265]. Also, isobaric labeling appears to have a higher precision [265].

Recently, neutron-encoded (NeuCode) labeling has been proposed as a promising new metabolic labeling approach that relies on the ability of modern high-resolution MS to distinguish extremely small mass differences (in the orders of mDa) [261, 266]. However, NeuCode did not gain a lot of traction yet because of the extremely high cost of its reagents, which seriously limits its throughput.

2.1.2. Post-metabolic labeling

In post-metabolic labeling, a different chemical label is added to each sample after protein extraction or digestion. Post-metabolic labeling strategies are often used for large experiments in which many samples need to be compared because of their superior opportunities to simultaneously measure multiple experimental conditions in a single MS run ("multiplexing"). In the context of high-throughput protein quantification, labels are added after digestion

[21] The reason is that the total amount of peptides (in µg) that is spiked onto the mass spectrometer is a fixed constraint. Under-spiking will result in low signal intensities and hence low proteome coverage, while over-spiking will result in signal saturation.

because this allows labeling of peptides that are buried within a protein's 3D structure, resulting in a higher peptide coverage.

Dimethylation is one of the oldest post-metabolic labeling strategies, though still widely used as it is easy to multiplex and fairly cheap [267]. Indeed, dimethylation requires relatively cheap reagents such as isotopically labeled formaldehyde and cyanoborohydride [268]. Here, all lysine side chains and peptide N-termini are dimethylated, except if the N-terminus starts with proline, in which case it will be monomethylated [269]. Originally performed in duplex [270], dimethylation was expanded to 3- [271], 4- [272] and 5-plex labeling [273]. By combining dimethylation with SILAC, 6-plex labeling has been achieved [274]. Compared to SILAC, dimethylation has a similar dynamic range and accuracy [265, 275]. Thus, the additional experimental variability introduced by post-metabolic labeling as compared to metabolic labeling seems to be limited in practice. Deuterium was used in the dimethylation approaches described above. However, compared to hydrogen-1, deuterium binds less strongly to the hydrophobic stationary phase due to the lower amplitude of its vibrational frequencies [276]. This results in a chromatographic shift as deuterium-labeled peptides elute somewhat earlier than their hydrogenated counterparts, which increases the uncertainty on the ratios of their peak intensities [277, 278]. Therefore, most contemporary isotopic labeling strategies avoid deuterium and favor e.g. ^{13}C, which does not induce a noticeable chromatographic shift [279].

^{18}O labeling is an enzymatic post-metabolic labeling approach that was originally used to improve peptide identification [280-282]. Soon after, ^{18}O labeling also became a tool for protein quantification [283-286]. Here, protein digestion is performed either in normal water or in ^{18}O-rich water. Two ^{18}O-atoms are incorporated at the peptides' C-termini, resulting in 4 Da mass shifts. Nowadays, ^{18}O labeling is not very common anymore as ^{18}O incorporation efficiency is variable and the technique cannot be multiplexed [287].

Isobaric labeling relies on mass differences in the fragment ions of isotopologic tags to quantify peptide ions via their MS^2 spectra. Indeed, the isobaric labels in each condition have the same total nominal mass, but a specific fragment ion, the reporter ion, has different masses for each label. Differentially labeled peptides thus coincide in the MS spectrum, but the reporter ions allow for quantification in the MS^2 spectrum. iTRAQ [288] and TMT [227] are the most well-known and frequently used reagents for isobaric labeling (Fig. 2.3).

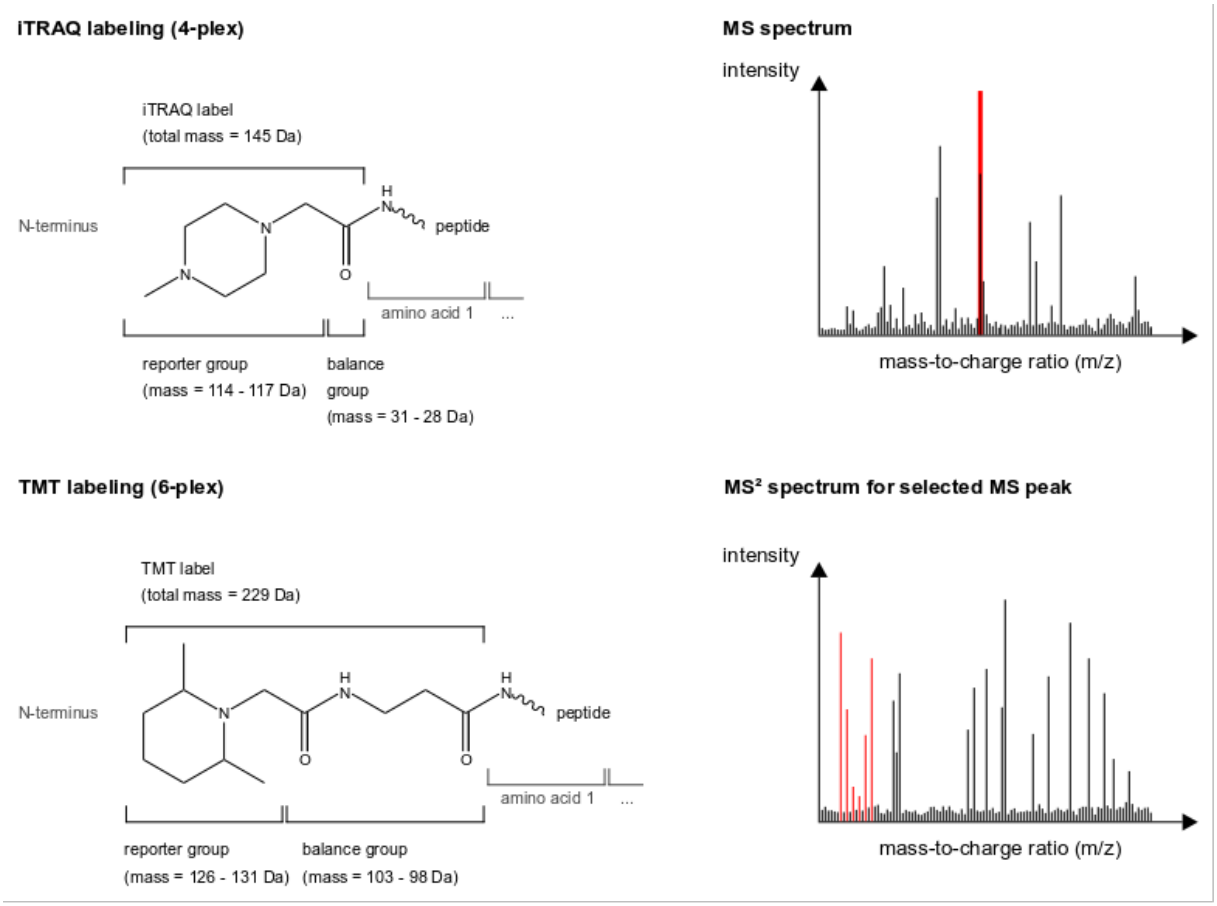


Figure 2.3. Left: chemical structures of 4-plex iTRAQ- (top) and 6-plex TMT- (bottom) labeled peptides. In isobaric labeling, peptide mixtures from different conditions are labeled with different isobaric labels that have the same total nominal masses but differ in the distributions of heavy isotopes within their structures (i.e. the isobaric labels are isotopologues of each other). Indeed, the reporter groups of the different isobaric labels all have different nominal masses (due to differential usage of heavy isotopes). This mass difference is balanced out by the balance group to ensure equal total nominal masses. The reporter group is constructed so that it detaches easily during fragmentation. It also has a strong preference to carry the positive charge after fragmentation so that it can be readily detected. Right: the monoisotopic[22] forms of differentially labeled peptide ions with identical amino acid sequences and charges will generate a single peak in the MS spectrum (top, red). However, after fragmentation of the isobaric labels, the mass difference of the reporter fragment ions will allow differential quantification in the MS² spectrum (bottom, red).

The main advantage of isobaric labeling is that it does not increase the complexity of the MS spectra and is therefore highly suitable for multiplexing [289]. Indeed, TMT readily allows 6-plex quantification [290], while iTRAQ goes up to 8-plex [291]. By making use of isotopologues, TMT-labeling has now even reached 11-plex [292, 293]. However, despite the market dominance of iTRAQ and TMT, some recent isobaric labeling alternatives such as DiArt [294] and DiLeu [295] labeling are gaining attention due to their easy synthesis, low cost, labeling efficiency and improved fragmentation efficiency. With DiLeu, up to 12 samples [296] can be compared in a single run. Isobaric labeling has also been combined with SILAC which increases its multiplexing potential even further [297].

However, in isobaric labeling, distortion of the reporter ion intensities in the MS² spectra is a major issue. Furthermore, isotopic impurities in the isobaric labeling reagents lead to isotopic

[22] The monoisotopic form is the isotopic variant wherein all atoms of a molecule are in their most abundant isotopic form.

overlaps in neighboring reporter ion intensities. Prior to data analysis, each reporter ion peak should therefore be corrected for isotopic overlap using correction factors supplied by the reagent manufacturer [298]. Another issue is that co-eluting near-isobaric peptide ions are isolated and co-fragmented with the target peptide ion, skewing the ratios of the reporter ion intensities towards the median value over all proteins, which is usually very close to one[23]. For this reason, this phenomenon is also called ratio compression [298-300].

Several technical solutions have been developed to reduce ratio compression. Gas-phase purification [301] and ion mobility separation [302] remove interfering ion species prior to fragmentation. MS^3 prevents co-fragmentation by isolating an MS^2 peptide fragment ion and fragmenting it again to produce an MS^3 spectrum. This additional step massively reduces the chance that the selected fragment ion will again be co-isolated with a fragment ion from another peptide species. As this MS^2 fragment ion contains all the isobaric labels, the reporter ions in the MS^3 spectrum enable more accurate and more reproducible protein quantification [303].

Even with these improvements, around 8% of the MS^3 spectra remain affected by co-fragmentation [265]. The introduction of synchronous precursor selection MS^3 (SPS-MS^3), in which multiple MS^2 fragment ions of a precursor peptide are co-isolated, strongly boosts MS^3's sensitivity [304]. However, despite clear progress, MS^3 still requires complex instrumentation and has a slower duty cycle, resulting in a lower identification depth compared to MS^2 quantification [289]. TMTc and EASI tag are new strategies whereby the complement ions (i.e. peptide fragments that remain attached to the balance group) are quantified in the MS^2 spectrum [289, 305, 306]. These approaches avoid both the complexities of MS^3 and the unwanted ratio compression because they examine the isotopic envelope of a particular labeled MS^2 peptide ion fragment.

Disadvantages of isobaric labeling include the loss of quantification depth and the poor accuracy for quantifying low-intensity peaks due to the lower signal-to-noise ratio in MS^2 spectra compared to MS spectra [265, 307]. The former can be partially alleviated by pre-fractionation. However, this results in longer run times and a more complicated data analysis [308, 309]. MS^2-based quantification also implies that only those peptides that are selected for fragmentation can be used for quantification [266]. Hence, for the many proteins that are identified with few peptides, it is difficult to assess the precision on their differential abundance estimates. Finally, iTRAQ and TMT labeling are relatively costly [275].

2.2. Label-free mass spectrometry-based proteomics

My PhD work is centered around the quantification of data-dependent label-free discovery-based shotgun proteomics data. Therefore, I will here give an extensive overview of the label-free shotgun proteomics workflow.

2.2.1. The label-free proteomics workflow

An overview of a generalized proteomics workflow has already been given in 1.2.2. In this section, I will focus on the aspects that are specific to label-free proteomics, and more specifically on those aspects that pose challenges to the ensuing data analysis.

A label-free proteomic workflow consists of the following steps: (1) proteins are extracted from the sample, (2) these proteins are digested into peptides, (3) the peptides are separated by reverse phase HPLC (RP-HPLC), (4) eluting peptides are ionized, (5) an MS spectrum is taken, (6) a precursor ion is selected, (7) this precursor ion is fragmented and (8) an MS^2 spectrum

[23] The total amounts of peptides loaded and analyzed are usually equal for all samples.

of its fragments is taken. Steps 6-8 are typically repeated several times for different precursor ions before a new MS spectrum is recorded. The peak intensity in the MS spectrum is used as a proxy for a peptide ion's abundance, while its corresponding MS² spectrum is used to identify the precursor. An overview of this procedure for a quadrupole Orbitrap instrument is given in Fig. 2.4.

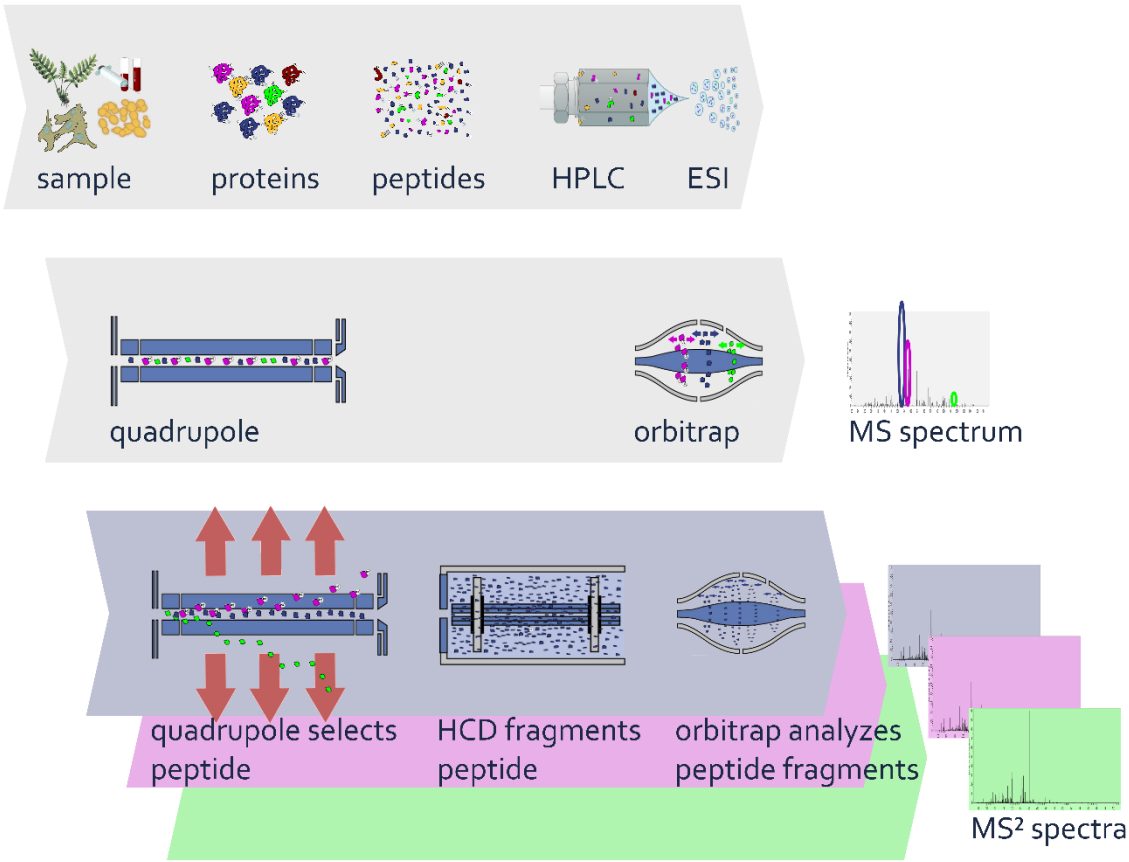


Figure 2.4. Overview of a typical label-free mass-spectrometry-based proteomics workflow on a quadrupole Orbitrap instrument. First, proteins are extracted from a sample. Then, the extracted proteins are digested into peptides, using a protease with a strong cleavage specificity (typically trypsin). This peptide mixture is loaded onto the instrument. A high-performance liquid chromatography column allows separation of the peptides by their hydrophobicity (RP-HPLC). The eluate from the column is ionized by electrospray ionization (ESI). The quadrupole ensures that only positively charged ions pass through. In the Orbitrap, an MS spectrum is taken. Each peak in the spectrum corresponds to a peptide ion. In a next step, the quadrupole will select a peptide ion with a sufficiently high peak in the MS spectrum. This peptide is fragmented and an MS² spectrum of its fragments is taken. This process is repeated for other peptides with high enough MS intensities.

Several of these steps create particular issues that are important to consider when analyzing the data. Table 2.1 presents a non-exhaustive overview of common issues in label-free proteomics workflows that have an important impact on the ensuing data. In what follows, I will elaborate on the most important issues of each data analysis step.

Table 2.1. Non-exhaustive list of issues that arise during a label-free shotgun proteomics workflow and their impact on the data.

Analysis step	Issues	Consequences
Protein extraction	Differences in extraction efficiencies	Variable protein concentrations over different samples
Protein digestion	Unequal enzymatic cleavage efficiencies, non-canonical cleavage, peptide ragging	Unequal concentrations of peptides originating from the same protein
RP-HPLC separation	Technical variability in elution times	Difficulties in matching unidentified peptides to identified peptides across different runs
	Peptide carry-over due to peptides from previous runs that did not elute from the column	Detection of peptides that were not present (worst case), peptide intensities influenced (increased) by the order in which the samples were run
ESI ionization	Strong differences in peptide ionization efficiencies	Strong differences in MS intensities for peptides originating from the same protein
	ESI saturation: peptides compete for ionization	Ionization suppression: a peptide's MS intensity (partially) depends on the nature and amount of other co-eluting peptides
		Suppression of high MS intensities resulting in an upper limit of quantification
	Peptides ionize in different charge states	Intensities for different charge states need to be taken into account when quantifying a peptide
	Gas-phase peptide ions undergo chemical reactions	Artificial peptide modifications, which are possibly not accounted for, resulting in unidentifiable spectra and missing values
	Changes in electrospray voltage during and across different runs	Increased variability in peptide ionization efficiencies
MS recording	Discrete MS spectrum recording	MS intensity not recorded at the elution peak
	MS detector saturation	Suppression of high MS intensities resulting in an upper limit of quantification
MS^2 recording	Limited amount of MS peaks targeted for fragmentation	Different peaks fragmented across different runs: missing values
	Preferential targeting of high-intensity MS peaks	Intensity-dependent missing values
	Loss of unexpected peptide fragments	Difficulties in MS^2 spectrum identifications and thus missing values
	Co-fragmentation of peptide ions with similar m/z values	Unidentifiable chimeric MS^2 spectra and thus missing values

The first step in the protocol, protein extraction, was shown to be responsible for 72% of all technical variability [310]. However, small differences in protein extraction efficiencies generally should not pose too much problems as an equal amount of total peptides is loaded onto the instrument for every sample. As noted in 1.2.2, it is imperative that all sample preprocessing is compatible with RP-HPLC-MS. It is therefore important to limit or avoid the use of urea [105, 106] and to avoid the use of SDS [91, 106-109] as well as other contaminants [112].

The digestion step is also critical. If the digestion efficiencies differ between samples, certain peptides might be formed to different extents. Consider for example the proteolytic cleavage of the following protein:

SESNAHFSFPKEEEKEFLESYPQNCPPDALPGTPGNLDSAQELEGFQIPTNLDWAGTSQAR

The most commonly used protease, trypsin, cleaves at the C-terminus of lysine (K) and arginine (R) (indicated in red) [311]. Depending on the digestion efficiency, varying amounts of the peptides SESNAHFSFPK and SESNAHFSFPKEEEK can be identified. Both the very short EEEK peptide and the very long EFLESYPQN…R peptide are barely or not detected by the mass spectrometer. Peptide SESNAHFSFPKEEEK is said to have one missed cleavage site, but peptides can have multiple missed cleavages. The presence of missed cleavages not only depends on the duration of the digestion, but also on the quality of the enzyme [312]. However, peptides can also be formed at non-canonical cleavage sites (i.e. not at lysine's or arginine's C-terminus) and modifications (either of biological origin or artifacts formed during sample preprocessing) may affect enzymatic digestion efficiency [311]. Next to these elements, the amino acids adjacent to a possible cleavage site, the location of a peptide in a protein's 3D structure and chemical degradation also play a role in the kinetics of the generation and degradation of different peptide species. Moreover, unwanted chemical reactions during sample preprocessing and/or the activities of exopeptidases that might be present in the sample or as contaminants in the commercial protease batch may cause N-terminal or C-terminal peptide degradation ("peptide ragging") [313]. For these reasons, the concentrations of most peptides are not equal to the initial protein concentration [314]. This stresses the importance of uniformity in the digestion conditions over the different samples.

Since the RP-HPLC column is typically packaged with porous silica beads coated with apolar alkyl chains (e.g. C18), peptides will be separated by their differences in hydrophobicity [315]. Protein modifications can however influence the column retention of an affected peptide [316]. Further, there is inherent variability in the chromatographic retention of a given peptide over different MS runs. This is an important factor to bear in mind when matching an identified peak in one MS run to an unidentified peak in another run; so-called matching between runs [317]. Comparing retention times over different runs is routinely done by algorithms for retention time alignment [318]. Retention times are often monitored as a general quality control measure by spiking in a known peptide or peptide mix [319, 320] and several algorithms have been developed to predict the retention times of known peptides [315, 321-323]. In order to maximize the mass spectrometer's operation time, HPLC columns are generally not replaced between analyses. Therefore, peptides from a previous injection might partially remain retained on the column and elute during a subsequent run, thus leading to the detection of peptides that were not present in that last sample and/or biased quantifications. Little has been published in the literature about sample carry-over, even though this is a non-negligible problem [324].

Peptide elution happens continuously, while the mass spectrometer only generates mass spectra at discrete time points. As peptide elution takes around 5 – 25 seconds [325, 326] and modern mass spectrometers typically tend to record an MS spectrum every second, the intensity peak for a single peptide ion will be recorded in sequential MS spectra. To allow for

an accurate quantification, it is important to record at least one MS peak at or near each peptide's elution apex. High-resolution RP-HPLC separation concentrates peptides in narrower elution peaks and is highly beneficial for different reasons [141]. First, repeated fragmentations of the same peptide ion over its elution window are highly reduced, thus winning MS analysis time. Next, highly focused peptide elution increases ion intensities and the ability to target monoisotopic peaks, which facilitates identification. Finally, an improved separation reduces co-elution of peptides, preventing ion suppression and co-fragmentation (see below).

The quality of ionization is of utmost importance as poor ionization leads to lower quality spectra (lower signal-to-noise ratio), making it more difficult to identify and quantify peptides. The electrospray voltage should also be kept as constant as possible over different runs as even small voltage deviations aggravate the variability in ionization efficiencies of peptides over different runs, which reduces the precision of label-free quantification [225].

During electrospray ionization (ESI), peptides are endowed with a positive charge (when working in positive ionization mode) [122]. However, the ionization efficiency of peptides differs. For example, highly acidic peptides and phosphopeptides ionize more poorly [123]. Such different ionization efficiencies leads to different MS intensities across different peptides, even if their concentrations are equal. In addition, the ionization efficiency of a peptide is also influenced by the nature and the amount of co-eluting peptides as these may suppress each other's ionization, a phenomenon known as ionization competition or ionization suppression [122, 327]. For these reasons, peptide ionization efficiencies are still difficult to predict [328]. Finally, the total amount of peptides that can be ionized simultaneously is also limited due to saturation effects during ionization [329, 330], which results in a suppression of high MS intensities and therefore an upper limit on the amount of peptides that can be quantified.

Ionization also generates different ion species from a single peptide, which differ in charge state. Indeed, ESI typically generates multiple ions with different charge states from the same peptide. For tryptic peptides, a double positive (+2) charge state is the by far most common after ESI ionization (around 75 - 90% of all ions), followed by triple positive (+3, around 10 - 20% of all ions) and single positive (+1, around 5% of all ions) [331]. The ionization step might also affect the chemical composition, and hence the mass, of an ion. Examples include in-source peptide oxidation and the loss of water or ammonia from N-terminal glutamate or glutamine residues, respectively, to form pyroglutamate [332, 333].

In the MS spectrum, the intensity of every ion that eluted from the HPLC column at that particular point in time is recorded in function of its mass-to-charge (m/z) ratio. Ion intensities can be seen as proxies for abundance, with the *caveat* that the ionization efficiencies differ strongly between different ion species. The detector of the mass spectrometer is also sensitive to saturation, which might also result in suppression of high MS intensities [329, 330].

After the generation of the MS spectrum[24], a high-intensity ion species is isolated with the quadrupole and targeted for fragmentation, a process that is repeated several times before a new MS spectrum is recorded. High-intensity MS peaks are intentionally targeted for fragmentation to avoid the targeting of noise peaks and hence increase the depth of the analysis. However, the fact that not all peaks in an MS spectrum can be targeted for fragmentation and the fact that high-intensity peaks are preferentially selected, results in intensity-dependent missing values. Dynamic exclusion, whereby previously-targeted m/z

[24] Or sometimes during the generation of the MS spectrum, if the machine has two mass analyzers.

values are temporarily excluded from fragmentation increases the quantification depth but does not alleviate the issue of intensity-dependent missingness.

Fragmentation also commonly results in neutral losses of water and ammonia, or even partial or complete amino acid side chains [137]. Such losses need to be taken into account when identifying a peptide ion based on its MS^2 spectrum. To facilitate identification, it is also preferable that the monoisotopic peak is correctly determined. For short peptides, the monoisotopic peak is often the highest one, but for longer peptides, this peak might be considerably smaller. This sometimes leads to determination of the wrong peak in the isotopic envelope and hence a wrong selection of candidate peptides during database filtering. Finally, when one or more peptides of about the same m/z elute together with a target peptide ion, such neighboring ions can be isolated as well, resulting in a mixed MS^2 spectrum [334]. Such a mixed fragmentation spectrum will also cause problems regarding identification and quantification.

2.2.2. Advantages and disadvantages of label-free MS-based proteomics

An obvious advantage of label-free MS-based proteomics is that the laborious and often costly labeling is omitted altogether [226]. Moreover, label-free analyses have a deeper coverage compared to label-based strategies. Indeed, compared to label-based quantification at the MS level, a significantly deeper coverage in label-free proteomics is caused by the decrease in spectral complexity [257, 258]. Compared to isobaric labeling strategies, the increase in coverage mainly results from the higher signal-to-noise ratios in MS spectra compared to MS^2 spectra [265, 307]. As noted before, label-free approaches also have a higher coverage and a higher dynamic range than label-based approaches [263, 308]. Another major advantage of label-free quantification is that there are no restrictions on the sample type. Indeed, metabolic labeling cannot be applied to every type of sample (e.g. patient blood samples), and, although the use of a labeled reference standard (e.g. super-SILAC) might be an option in certain cases, this application is destined to fail when the samples in an experiment are too different from each other. Isobaric labeling can be applied on any sample type and has reached multiplexing capacities up to 12 [296]. However, when more samples need to be compared, between-run comparisons will need to be made [335]. Moreover, for any label-based strategy, all samples should be generated prior to the MS analysis of the first sample. Label-free quantification allows for the analysis of an unlimited number of samples, even retroactively, as long as the machine's working conditions have not been changed dramatically [317, 336].

Since labeling is omitted, each sample is analyzed separately. This results in longer overall run times and thus a lower throughput [306, 336]. Inevitably, precision will also be lower compared to label-based approaches because random run-to-run variability cannot be factored out. Thus, robustness of the instrumentation and the analytical conditions is imperative for successful label-free analyses [336]. However, after balancing the pros and the cons, most research publications that compared label-based quantification to label-free quantification seem to express a general preference for the latter [257, 258, 308, 309], although the preference of individual researchers is often driven by their own experience, i.e. the instrumentation and quantitation technology they are most comfortable with.

2.2.3. Other label-free approaches

Next to label-free discovery-based data-dependent shotgun proteomics, there are two other major label-free proteomic techniques: targeted proteomics and data-independent proteomics. For the sake of completeness, and because I briefly refer to them in my future perspectives, I will here outline their major points.

All techniques outlined so far in this chapter aim to identify and quantify as many proteins as possible. However, none of these techniques guarantees that a particular protein of interest will be identified, especially if this protein is very low abundant (e.g. many transcription factors) [337] or is difficult to solubilize (e.g. membrane-inserted proteins) [338]. In fact, some proteins or protein regions have never even been identified by mass spectrometry, the so-called "hidden" or "dark" proteome [337, 339, 340]. Moreover, even if a protein of interest is identified in a certain run, it is not guaranteed to be identified in another run due to inherent run-to-run variability.

Targeted proteomics aims to reproducibly monitor and precisely quantify one or more selected proteins [341]. Indeed, if the retention times and *m*/*z* values from peptide ions of interest are known, a mass spectrometer can be programmed to detect, fragment and record MS^2 spectra only for those peptide ions. This technique is called selected reaction monitoring (SRM) and is typically performed on a triple quadrupole instrument (QQQ) instrument. In this instrument, the first quadrupole is used to isolate the peptide ion, the second quadrupole serves as a collision cell to fragment this ion and the third quadrupole is used to isolate selected fragment ions. Since the QQQ only isolates selected fragment ions, SRM on a QQQ instrument has a high dynamic range and is very selective and sensitive [342]. Note that it is also possible to operate other machines in SRM mode [343]. In parallel reaction monitoring (PRM), the third quadrupole is substituted with a highly accurate mass analyzer to record all fragment ions [344]. PRM recently gained a lot of traction as PRM assays have a similar performance compared to SRM assays, but are easier to develop and possibly even more specific [345].

In data-independent acquisition (DIA), the intensity of the MS peak does not determine which ions are being fragmented. Shotgun CID, MS^E and All-ion Fragmentation are DIA methods in which the complete *m*/*z* range of the MS spectrum is targeted for fragmentation [135, 346, 347]. Other approaches divide the *m*/*z* range of the MS spectrum into predetermined *m*/*z* isolation windows and sequentially co-fragment all ions within such windows before going on to record the next MS spectrum [120, 348, 349]. DIA methods are immensely promising because they record all fragment ion spectra. Theoretically, it is thus possible to record every possible peptide in a sample as long as its fragment ions can be identified. Moreover, DIA data can be retroactively queried for new peptides of interest (e.g. modified forms of peptides). DDA search engines have been modified to unravel the very complicated MS^2 spectra that result from DIA [346, 348]. However, deconvoluting such multiplexed spectra remains challenging [350].

Out of all DIA methods, Sequential Windowed Acquisition of all Theoretical fragment ion mass spectra (SWATH-MS) is the most popular [351-353]. In SWATH-MS, specific MS^2 fragment ion peaks are tracked over time without deconvoluting the spectra [351]. On the new Q Exactive HF-X, a single SWATH-MS run identifies on average more than twice the number of peptides of a DDA run [354]. Moreover, in a benchmark experiment where 12 human proteins were spiked in a HEK-293 background, SWATH-MS analysis had only 1.6% missing values compared to 51% missing values in the DDA analysis of the same sample [355]. SWATH-MS is also highly reproducible across different labs [356]. Compared to targeted proteomics, it has a much higher coverage, but remains slightly less sensitive [351, 357-359].

The major disadvantage of DIA is that knowledge on the chromatographic and mass spectrometric behavior of peptides of interest is required to build a spectral library needed to identify the peptides [359]. This often implies that a DDA analysis is performed prior to the DIA analysis to obtain the necessary information on the peptides expected in the sample.

3. FROM SPECTRA TO DATA

In the previous chapter, I described the bottom-up shotgun proteomics workflow up to the generation of MS and MS² spectra. Given the continuous elution of peptides from the HPLC column, most peptide ions will be recorded in multiple, consecutive MS spectra. The peaks in these spectra must be converted into meaningful qualitative and quantitative information about the peptides in the samples. In this chapter, I will describe how MS and MS² spectra are used to identify peptides in label-free DDA MS-based proteomics. Then, I will show how peptides are assigned to proteins and I will describe how a single intensity value is assigned to each peptide followed by a description of the nature of the MS data. I will conclude this chapter with an explanation for the need for benchmarking and the description of the CPTAC Study 6 benchmark dataset.

3.1. Peptide ion identification

In MS-based proteomics, peptide ions (also called "features") are identified based on information present in both the MS and the MS² spectrum and only features that are deemed to be reliably identified are typically passed on to the quantification stage. Do note that a feature's identification is actually not required for its quantification. Indeed, some workflows, such as those proposed by Sieve (Thermo Fisher Scientific) and Progenesis (Nonlinear Dynamics) adopt a feature-based quantification whereby identification efforts are primarily focused on features that are flagged as differentially abundant after the quantification stage. However, my thesis primarily focuses on the protein quantification, for which prior identification of features is required. Therefore, the remainder of this thesis describes the identification-based workflow.

The charge state and the mass of a feature can be readily determined from its MS spectrum. Indeed, charge state variants elute simultaneously and are thus present in the same MS spectra. As mass-to-charge ratios (m/z values) are recorded in MS spectra, an ion with, say, charge state +3 will have a peak at an m/z value exactly equal to 2/3 times the m/z value of the same ion with charge state +2. As mentioned in 1.2.2, the natural occurrence of stable heavier isotopes generates an isotopic envelope for each charge state of an ion. Since isotopes differ in mass by a natural number of neutrons, the minimal difference in mass between two isotopic variants is approximately 1 Dalton. This knowledge is used to infer the charge state of an ion: a difference of m/z = 1/2 Th between the peaks of an isotopic envelope denotes a +2 charge state, whereas a difference of m/z = 1/3 Th points to a triply positively charged ion. From this charge state, the monoisotopic mass of a peptide ion can be readily calculated by multiplying its corresponding monoisotopic m/z value by its charge state. However, knowing the mass of a peptide is not enough to identify it. Indeed, not only do all permutations of a particular amino acid sequence have exactly the same masses[25], some entirely different combinations of amino acids also have the same nominal masses[26]. Also note that modifications change a peptide's mass and that modified peptides often do not co-elute with their unmodified counterparts [333]. For these reasons, an MS² spectrum of a peptide is required to identify the amino acid sequence corresponding to a peptide's MS peak.

An MS² spectrum contains the m/z values for all the fragment ions obtained by fragmenting a single peptide ion. To enable identification, the selected peptide ion should by preference be the monoisotopic peptide ion. Indeed, any other peak in the isotopic envelope is in fact a

[25] E.g. the amino acid sequences LDGER, DGERL, GERLD, etc. all have a 588 Da mass.
[26] E.g. the amino acid sequences LGD and ER both have a nominal mass of 303 Da.

mixture of numerous isotopic variants. For example, the *m*+1 peak represents a mixture of ions that consist of a single heavy isotope together with only light isotopes. However, this heavy isotope can be for example a heavy carbon in the first amino acid, or a heavy carbon in the second amino acid, or a heavy nitrogen in the third amino acid, and so on.

As described in 1.2.2, the type of fragmentation determines the main type of fragment ions that will be found in the MS² spectrum (both y- and b-ions for CID, mainly y-ions for HCD [137-139]). In the early days, when mass spectrometers only generated a few 100 MS² spectra, peptide identification was often done manually by printing out the MS² spectra and measuring the distances between the fragment ion peaks. The sequence could then be inferred through some sort of "ladder sequencing" as demonstrated in Fig. 3.1.

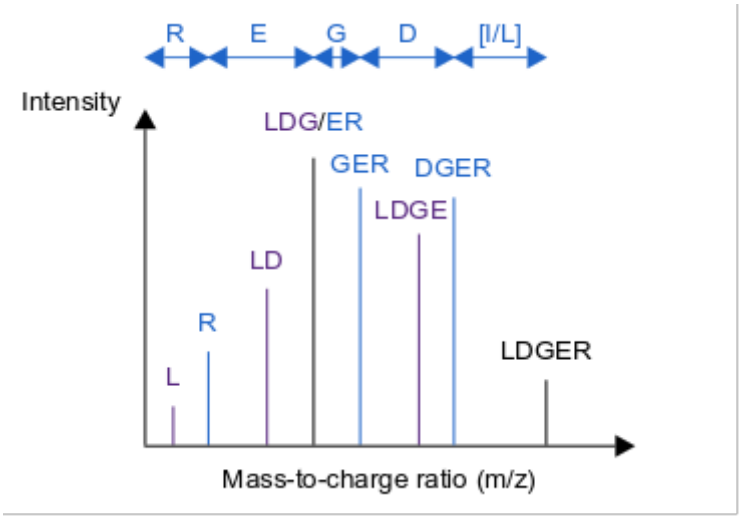


Figure 3.1. Theoretical demonstration on the inference of a peptide sequence based on some sort of ladder sequencing if all b- and y-ions are present (shown in purple and blue, respectively). The mass differences between the peaks give information about the amino acid sequence of the peptide. Since leucine (L) and isoleucine (I) have the same empirical chemical formula and thus the same masses, no direct distinction can be made between both. y-ions are often somewhat more abundant than b-ions, especially in HCD fragment spectra. Therefore, peptide sequencing based on y-ions is shown. The presence of the b-ions complicates the analysis somewhat but can also provide supporting evidence. Note that this representation is somewhat simplistic. In reality, a-, c-, x- and z-ions might also be present in the fragmentation spectrum. Moreover, some b- and y-ions might not exceed noise levels, while noise peaks (either true noise or fragments from co-isolated peptides) might obfuscate the analysis. Modifications will also influence the masses of those fragments that carry the modified amino acid.

Nowadays, manual inspection of the sheer number of MS² spectra is impossible as high-quality experiments typically generate well over 5,000 MS spectra and over 50,000 MS² spectra in a single run.

Because of the sheer amount of data, bioinformatics software is now indispensable for feature identification. One example is PeptideProphet [360], which makes use of the SEQUEST algorithm [361]. The idea behind SEQUEST is rather simple. First, a database with protein sequences is provided to the algorithm. This database clearly depends on the sample: when the sample is e.g. a human cell culture, one uses the human proteome from the UniProt Knowledgebase, whereas for an *Arabidopsis* sample, the TAIR database is often used [362]. All proteins in the database are then *in silico* digested according to the cleavage rules of the protease used. For trypsin, all protein sequences are thus split at lysine (K) or arginine (R) residues. Next, only those peptides are selected from the database which lie within a narrow mass tolerance range around the observed mass value of the MS peak. Then, all possible b- and y-ions are calculated for each of these peptides and a score is calculated. This score

increases for each theoretical b- or y-ion that, again within a certain mass tolerance, can be matched to an observed MS² peak. Similarly, the score decreases for each theoretical peak that cannot be matched to an observed peak and for any observed peak that cannot be matched to a theoretical one. All theoretical peptide candidates are then ranked according to their scores and the highest-ranking peptide is assumed to be the correct one if it scores significantly higher than the second-highest ranking peptide. This coupling of a peptide ion species to a certain MS² spectrum (and hence the corresponding MS peak) is termed a peptide-to-spectrum match (PSM) [363]. The simple idea of calculating a score based on theoretical spectra from *in silico* digested proteins still remains the basis of all other database search algorithms today.

In reality, the SEQUEST algorithm is more complicated than described above as it also takes neutral losses such as water, ammonia and carbon dioxide into account. In its earliest version, modifications could only be considered if they were assumed to be present at every occurrence of the modification site. This was achieved by simply shifting the masses of the *in silico* peptides. Later on, variable modifications were also introduced. This means that the algorithm searches both for the presence and the absence of certain modifications on certain amino acids in certain peptides. Phosphorylation of serine, threonine and tyrosine residues is an example of a variable modification that is often included in a search. One problem with searching for variable modifications is that they massively inflate the computational search space, especially if peptides are allowed to carry multiple modifications at multiple amino acids. For example, a variable search for phosphorylation alone may lead to a 67-fold increase in the search space [364]. Such inflations do not only drastically increase computational search times, but also increase the uncertainty on PSMs that map to unmodified peptides because of the increased probability that an *in silico* spectrum of an unrelated modified peptide matches against the spectrum of an unmodified peptide. Researchers are therefore often advised to not search for more than two or three variable modifications. Nonetheless, search space inflation can be largely fended off with some clever optimizations because the masses of modifications rarely coincide with amino acid masses, a property that can be exploited in high-resolution MS data [365]. Other optimizations include performing a two-pass search in which the second search only probes for modifications of peptides for which an unmodified counterpart was already identified in the first search [366]. Alternatively, one could limit the search to those modifications and modification sites that have previously been confirmed and are stored in a curated database [364]. Interestingly, it has been shown that unaccounted modifications are responsible for about 20–50% of all false positive identifications [367]. Therefore, open modification search engines have been developed that allow users to search for mass differences [365]. That way, much more modifications and even peptides that differ by a single amino acid from the canonical sequence in the database can be detected.

The Mascot search engine was the first to introduce probability-based scoring, whereby the probability of an identification was weighted against the probability that a match between the theoretical and the observed spectrum occurred by random chance [368]. This was later on improved by incorporating the concept of target-decoy matching [369, 370]. In target-decoy matching, a set of nonsensical peptides (decoys) of the same size as the theoretical database peptide set (targets) is added to the search space. This nonsensical set can be obtained in multiple ways, e.g. by random scrambling of the original protein sequences, but is mostly obtained by simply reversing the protein sequences followed by *in silico* digestion. In case of palindromic sequences, forward and reverse sequences would overlap, but palindromes are extremely rare in practice because the protease's specificity typically requires peptides to end in very few specific amino acids (e.g. arginine (R) or lysine (K) in the case of trypsin). Therefore, a palindromic sequence should already display a missed cleavage after the first amino acid, which is rather uncommon. Next, the observed peaks are matched against the combined

database as described before. Then, an identification false discovery rate (FDR) threshold is calculated for each target-PSM. This FDR is simply calculated by dividing the number of equally- or higher-scoring decoy-PSMs by the number of equally- or higher-scoring target-PSMs. Finally, an FDR threshold is set (typically at 1%, rarely at 5%) and all PSMs under this threshold are passed on to the quantification stage, while the PSMs that exceed the threshold are removed from the data as these are deemed not certain enough. Note that for two-pass searches, this FDR calculation is way too liberal because the enrichment of targets in the first pass makes it more likely for the spectra to match to targets in the second pass. This constitutes a violation of the assumption that false matches should be equally likely to match a target or a decoy. Therefore, two-pass searches require an adjusted target-decoy database [371].

Nowadays, there are many different algorithms that enable peptide identifications from MS² spectra. These include Tide, a fast implementation of the SEQUEST algorithm [372], X!Tandem [373], MS-GF+ [374], MS Amanda [375], MyriMatch [376], Comet [377], Andromeda [378], OMSSA [379], Novor [380] and DirecTag [381]. A tool like SearchGUI [382], developed by the compOmics lab, unites all these algorithms in a graphical user interface. Furthermore, tools such as PeptideShaker [383] can be used to combine results of different search engines to boost identifications. The MaxQuant software package, which uses the Andromeda search engine, is very popular nowadays thanks to its integrated pipeline from identification to quantification and its user-friendly graphical user interface [384].

3.2. Protein inference

In a typical shotgun proteomics workflow, the aim is to identify and quantify as many proteins as possible, but the data used for this are at the PSM level. Therefore, PSMs should be first assigned to one (or more) protein(s). Protein inference is straightforward for peptides that can be uniquely mapped to a protein sequence stored in a database. However, this becomes more complicated when a peptide can be mapped to several protein sequences [385], which frequently occurs for different protein isoforms that result from alternative splicing and for proteins that originate from paralogous genes[27]. Such peptides are called “shared peptides”, “degenerate peptides” or “razor peptides”.

PeptideProphet was the first algorithm to propose a solution to this problem [386]. Here, protein identification probabilities are first calculated under the assumption that all peptide matches are independent. Then, peptide identification probabilities are updated based on the protein identification probabilities. This updating of peptide and protein identification probabilities is then repeated until convergence. Others statistical models have also been proposed in an attempt to more reliably assign shared peptides [387, 388]. A conceptually simple way to deal with shared peptides is Occam’s razor approach. Here, each shared peptide is simply assigned to the protein which already has the highest number of identified unique peptides assigned to it. This is the approach currently implemented in MaxQuant [384]. MaxQuant also groups proteins that share a large fraction of their peptides in so-called “protein groups”. As the abundance of a shared peptide might reflect the combined abundance of multiple proteins, shared peptides are almost always removed from the dataset prior to quantification.

Another issue in protein inference is the occurrence of so-called “one hit wonders”: proteins that are identified by a single peptide. It is generally considered unreliable to infer a protein

[27] Paralogous genes are genes that descend from the same ancestral gene within a species and therefore often have a high sequence homology. Their resulting protein products often execute similar functions.

based on a single peptide because if this identification is incorrect[28], a protein is quantified that might not even be present in the sample. Therefore, such proteins are often removed from the dataset after using the so-called “two-peptide rule” [385], which however has also been criticized. It was indeed shown that it is more likely to find a protein with two mediocre-scoring peptides in the decoy database than it is to find a protein with a single high-scoring peptide in the decoy database [389, 390]. It was therefore proposed to abandon this two-peptide rule in favor of an approach in which the protein identification FDR is the sole criterion to accept a protein as being identified. Indeed, although in the past, all PSMs that passed a certain FDR threshold were passed on to the quantification stage, it was soon shown that if the PSM FDR is controlled at the 1% level, the protein FDR is much higher and should therefore also be taken into account [391, 392]. This is because the chance that a false positive PSM maps to a protein in the database is in theory random and therefore only dependent on that protein’s number of theoretical (tryptic) peptides. However, a protein that is truly present in a sample will typically generate multiple PSMs that will correctly map to that protein. Therefore, a fraction of the proteins in the dataset (i.e. the true positives) will be enriched in true positive PSMs, while a relatively large fraction of false positive proteins will have very little PSMs assigned to them [393].

Many different methods have been developed to estimate protein identification FDRs [386, 394-396]. However, when calculating a protein FDR, it is important to clearly define how this value is to be interpreted. As noted by The *et al.* (2016), a distinction should be made between defining a false discovery as a protein that is inferred from an incorrect PSM versus defining a false discovery as the incorrect identification of a protein that is in reality not present in the sample [397]. These are not the same, as many proteins might be present in the sample that are not assigned any correct PSM but can be assigned to one or more incorrect PSMs by random chance. These authors noted that protein FDRs can strongly differ depending on the definition.

The protein inference problem is reviewed more in depth in Huang *et al.* (2012) [398] and Serang and Noble (2012) [399].

3.3. Peptide quantification

A first step in the quantification procedure is the determination of the abundance of all identified peptides. This can be done in two ways. The first way, summing up MS^2 intensities, is now largely deprecated for label-free DDA data. The second and by far the most common way is by using the MS intensities.

The reason why label-free approaches in which MS^2 fragment ion intensities are summed to determine a peptide ion’s intensity perform worse than MS peak-based methods in terms of reproducibility, missing data, quantitative dynamic range and quantitative accuracy [307] is because MS^2 intensities are highly data-dependent. Due to dynamic exclusion, a particular peptide ion can be targeted for fragmentation when it is relatively far from its elution apex. This makes summed-up MS^2 intensities much more variable from run to run and hence less suited for reliable quantification.

When quantifying peptides based on MS intensities, it is important to realize that peptides elute continuously from the RP-HPLC column, while MS spectra are recorded at discrete time points. Indeed, if an MS spectrum is recorded every second and peptide elution typically ranges from 5-25 seconds [325, 326], most peptide ion peaks will be recorded in multiple, sequential MS

[28] At a 1% PSM identification FDR, on average 1% of all PSMs is expected to be wrong, but for each specific PSM, this probability might be significantly higher or lower.

spectra (see also 2.2.2). Here, every peptide ion is recorded in each MS spectrum as an isotopic envelope due to the natural isotopic occurrence. A simple way to determine an identified peptide ion's intensity would be to sum the intensities of the isotopic envelopes in each MS spectrum and select the MS spectrum in which this summed intensity is the highest (i.e. near the peptide's elution peak). However, this results in rather imprecise quantifications because for one peptide ion, an MS spectrum might be recorded very close to its elution peak, while for another ion, the MS spectrum with the ion's highest intensity peak might be recorded up to 0.5 seconds[29] before or after its elution peak. As shown in Fig. 3.2, the elution profile of a peptide can be fairly easily reconstructed based on the sampled MS intensities of this peptide. Indeed, by fitting a curve to these summed intensities over time, a peptide's elution profile can be reconstructed, which results in a much more reproducible quantification [400, 401].

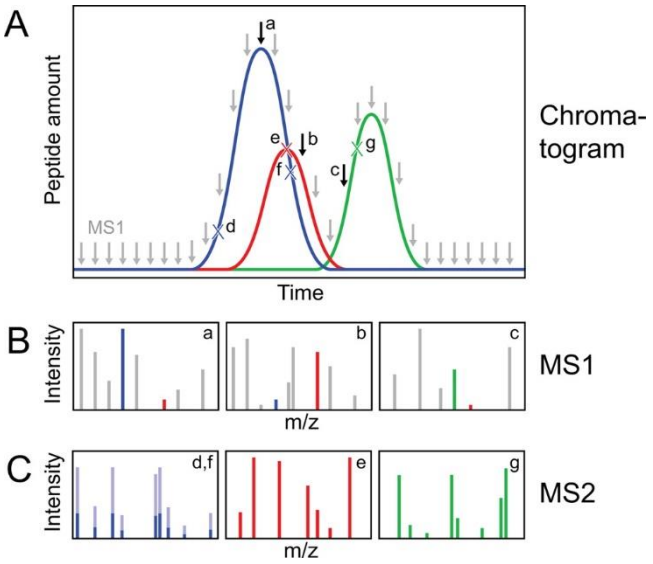


Figure 3.2. Theoretical example demonstrating the challenges in MS¹ and MS² peptide quantification. A. Illustration of the elution profile of three peptides (blue, red and green). Arrows denote the discrete time points at which an MS spectrum is taken. Crosses on the elution profile indicate when that specific peptide is selected for fragmentation, which results in an MS² spectrum. MS spectra at time points a, b and c are shown in B; MS² spectra are shown in C. Due to its high abundance, the blue peptide was selected for fragmentation early in its elution profile (MS² spectrum d: dark blue). Because of dynamic exclusion, this peptide was not re-selected for fragmentation again until far beyond its elution peak (MS² spectrum f: light blue). Reprinted with permission from *Krey et al. (2014) [402]*, © 2014 American Chemical Society.

Some methods are even more sophisticated in trying to increase the quantitative accuracy. For example, the Andromeda search engine, incorporated in MaxQuant, integrates the peak intensities of each ion's isotopic envelope during peptide elution. More specifically, Andromeda fits a Gaussian peak to the three most central data points in each isotopic envelope. These 2D peaks are then smoothed to 3D peaks in the retention time dimension and the total intensity for a particular ion is then set equal to the volume of its corresponding 3D peak [384]. Fig. 3.3 gives a 3D visualization of the increase in intensities in the isotopic envelope at the start of the elution of a peptide ion.

[29] This is, given that an MS spectrum is recorded every second, a typical user-defined setting.

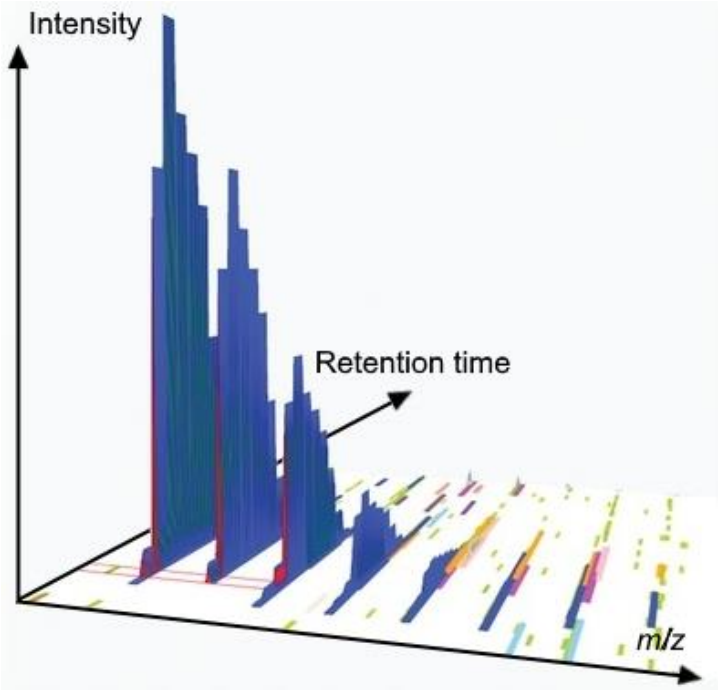


Figure 3.3. 3D view of the isotopic peaks during the elution of the doubly charged peptide ion CCSDVFNQVVK in sequential MS spectra in the MaxQuant Viewer tab. Image adapted from Tyanova *et al.* (2015) [403]. Proteomics. Published by Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, CC BY-NC-ND 4.0.

A disadvantage of Andromeda's algorithm is that it is computationally intensive. Moreover, until very recently, MaxQuant was only available on Windows, which impeded its inclusion in automated Linux server pipelines [404]. These were the main motivations for the development of moFF, a platform-independent quantification algorithm that only quantifies the apex of the elution profile, making it fast, but without compromising on quality [405, 406]. Do note that many other algorithms have been developed to calculate peptide ion intensities, some of which are reviewed in Sandin *et al.* (2014) [407].

3.4. The nature of the data

The previously described workflow results in specific data properties that are typical for MS-based proteomics and that are important to bear in mind when evaluating protein levels.

In bottom-up proteomics, PSMs do not correspond to peptides, but to peptide ions, whereby multiple peptide ions can map to the same peptide sequence. Indeed, the same peptide can often be identified under different charge states and/or with different modifications. Sometimes, a given modification can be detected at more than one location within the same peptide. As the (unmodified) backbone amino acid sequences of such PSMs are identical, these ion species are expected to behave more alike compared to unrelated ion species. Therefore, the intensities of these species are correlated with each other. Similarly, ion species of the same charge states will also be correlated. On a higher level, there is correlation between all peptides that are mapped to the same protein. The highest level in the hierarchy is the correlation between proteins. Indeed, since proteins interact with each other in numerous pathways, proteins that closely interact with each other, that are part of the same signaling pathway or even reside in the same subcellular location tend to behave more similarly than totally unrelated proteins. Orthogonal to the correlations between peptides and proteins, there is a strong within-run correlation due to the relatively large run-to-run variability in label-free proteomics. To keep this run-to-run variability minimal, it is necessary to tightly control the

instrumentation. This makes raw label-free proteomics data hierarchical with correlations on many levels in the data. Moreover, it is possible that for the same protein in the same MS run, e.g. two peptides are observed with only one PSM, three peptides with two PSMs and one peptide with three PSMs.

Even for peptides originating from the same protein, differences in intensities are often substantial. One reason for this is that differences in proteolytic cleavage efficiency cause some peptides to be generated more efficiently than others, rendering individual peptide levels not equal to protein levels. Moreover, intensities of different peptide ions also strongly differ because of large differences in ionization efficiency. Finally, the latter is context-dependent as the nature and the amount of co-eluting peptides also drive the efficiency by which a peptide is ionized (see chapter 8). Such ionization competition can be a more important driver of a peptide's intensity than its actual abundance [408]. As noted in section 3.2, the intensities of shared peptide sequences arise from an unknown combination of peptide ions coming from different proteins. Therefore, shared peptides are often removed from the dataset prior to quantification.

While the raw data is at the PSM-level, quantification is typically done at the protein level. PSM-level data thus needs to be summarized to the protein level. It is possible to summarize the PSMs directly to the protein level. However, this might be suboptimal because of the hierarchical nature of the data and the missing values that make the data unbalanced. Indeed, if the PSMs would be summarized as if they were independent observations, a bias will be introduced because some peptides will be "overrepresented" in the protein's abundance estimate, while other peptides will be "underrepresented". Therefore, summarization is sometimes done in two steps: in a first step, the data are summarized to the peptide level (i.e. each peptide sequence is the summary of all its potential charge states and peptideforms). MaxQuant, for example, outputs peptide-level summaries as summed raw PSM intensities. In a second step, the peptide-level data are summarized to the protein level (discussed in detail in 4.1.5). Contrary to protein summarization based on peptide-level values, PSM to peptide summarization has not been studied in detail, and commonly-used data analysis pipelines such as the MaxLFQ algorithm in MaxQuant [317] and MSstats [409] calculate protein-level summaries based on PSM intensities, ignoring possible correlation between PSMs that map to the same peptide sequence. From here on, I will assume that all data are summarized to the peptide level, unless specifically mentioned otherwise. However, note that all of the criticisms regarding peptide-to-protein summarization outlined in 4.1.5 can also be applied to PSM to peptide summarization.

The linear dynamic range of the mass spectrometer also affects the data. Indeed, within a certain concentration range, an increase in peptide concentration will result in a multiplication of the MS signal with more or less the same factor. This range is termed the linear dynamic range. Above this range, the ESI spray ionization and/or the MS detector becomes saturated, leading to a plateau in the ion's MS intensity signal. Below the linear dynamic range, a peptide ion will not be observed. Note that the linear dynamic range will be different from peptide to peptide due to their different ionization efficiencies. Fig. 3.4 gives a graphical representation of the linear dynamic range.

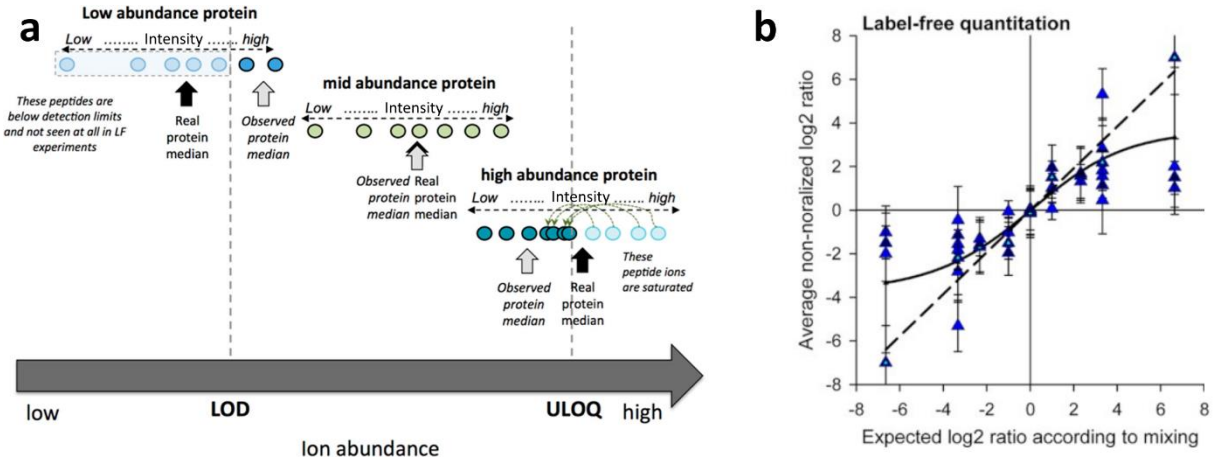


Figure 3.4. The concept of linear dynamic range in proteomics. (a) Peptide ion intensities for three theoretical proteins: a low abundant, middle abundant and highly abundant protein. For the low abundant protein, many of its peptides will fall below the limit of detection (LOD) and will therefore not be identified. Its abundance will only be estimated based on the identified peptides and might therefore be over-estimated. For the middle abundant protein, most of its peptides are observed and the protein's estimated abundance will be close to its true abundance. For the highly abundant protein, some of its peptides will be more abundant than the upper limit of quantification (ULOQ), and their ion signals will be lowered. Hence, the highly abundant protein's abundance might be under-estimated. (b) The dynamic range in practice. In this experiment, different known amounts of proteins were compared to each other. When the differences in concentration become large, the difference between the estimated protein abundances and the true protein abundances also increase. Such deviations from linearity are substantial for large differences in abundance. Image (a) modified after Jarnuczak *et al.* (2016) [330], © 2016 by American Chemical Society ("ACS"), CC BY 4.0 and image (b) adapted from Arsova *et al.* (2012) [410] © 2012 by The American Society for Biochemistry and Molecular Biology, Inc.

The large amount of missing values in the data is a major issue in label-free DDA shotgun proteomics. Upon searching the public repository PRIDE [411] for MaxQuant datasets that applied shotgun proteomics to full or partial proteomes, we found 16 to 82% missing values at the peptide level (see chapter 10). There are multiple reasons for this missingness. A first reason is that a protein might simply not be present in certain samples. A second reason is that experimental reasons (e.g. different tissues with different protein abundance profiles are compared), biological reasons (e.g. downregulation or degradation of a protein) or technical reasons (e.g. inferior quality of a certain run) might cause the MS intensity of a peptide that is truly present to fall below the background noise level. For label-free shotgun proteomics, the limit of detection was estimated to lie around 1 fmol [412]. A third important reason is in the data-dependent nature of the sampling: a peak that was selected in one run, might not be selected for fragmentation in a next run. Indeed, the height of the peak mainly dictates if a peak is targeted for fragmentation or not [413]. This type of missingness is thus largely intensity-dependent, although ionization also renders this type of missingness context-dependent. A fourth reason for missing values in label-free shotgun proteomics is that on average around 75% of all MS^2 spectra is not identified [414]. This happens when an MS^2 spectrum's highest-scoring PSM falls below the pre-set identification FDR threshold or when no distinction can be made between two or more high-scoring alternatives. Alternatively, the PSM could pass the FDR threshold but might in reality be misidentified. Poorly ionizing peptides are particularly at risk for failed or incorrect identifications. Peptides carrying a phosphate-modification, for example, ionize notoriously poorly because of the phosphoryl group's default double negative charge state. Moreover, CID fragmentation of peptides with a phosphorylated serine residue often results in a dominant neutral loss of phosphoric acid, leaving too little energy for the efficient fragmentation of the precursor's peptide bonds [415]. Furthermore, most peptide modifications with a biological or an artefactual origin are often unsearched for due to the

strong inflation in computational search space when allowing for too much modifications to occur on the peptides. It has been estimated that at least one third of all spectra cannot be assigned to a peptide due to the presence of sub-stoichiometric post-translational modifications [416]. Another cause for failed MS^2 identification is co-fragmentation [334]. The chimeric MS^2 spectra that result from co-fragmented peptide ions are either not identified, or misidentified, resulting in a missing value, or mapped onto only one of the precursor ions. As the corresponding MS peak is a mixture of more than one ion, the abundance of this ion will be estimated higher than it actually is.

All of this stresses the importance of MS^2 fragmentation. Indeed, all MS peaks that were not targeted for fragmentation or for which no peptide could be matched to the MS^2 spectrum remain unidentified. One way to alleviate such cases of missing values is by applying a feature alignment or "match-between-runs" algorithm [400, 417, 418]. Here, unidentified MS peaks in one MS run are matched, based on their masses, charges and retention times, to identified peaks of another run in which a similar sample was analyzed. Important for these algorithms is that they should be able to accurately align retention times over different runs and keep their retention time windows (i.e. the deviations in retention time to allow a match between two MS peaks) narrow enough to prevent incorrect matches. This again stresses the importance of a stable analysis workflow and sufficient quality control [419]. A match-between-runs algorithm is also frequently applied in MaxQuant searches [420].

3.5. The need for benchmarking

As explained above, peptide intensities are used for differential analysis of protein abundances. However, there is an enormous variety in differential protein abundance analysis workflows, and each step in these workflows has its impact on the result. Hence, the performances of the different workflows might also differ considerably. This difference in performance is an important point. Indeed, if suboptimal workflows are used, biologically relevant proteins might remain under the radar. Therefore, I will here explain the need for benchmarking to allow comparison of the performances of different workflows.

To evaluate the sensitivities and specificities of different quantitative pipelines, a dataset is needed in which the true relative amounts of proteins in all samples ("the ground truth") are known. Such a dataset can either be generated by simulation or by spiking in proteins of a certain organism into another organism's proteome. Simulations have the advantage that they are very simple to generate: one only needs a computer. However, simulated datasets often do not reliably capture the complex data structures of true biological experiments. In a spike-in experiment, a set of proteins from one organism is spiked into a complex protein background from another organism at different concentrations. Hence, when two different spike-in conditions are compared, only the spiked-in proteins are differentially abundant. Generating a spike-in dataset requires setting up a wet-lab experiment. Here, it is important to pick two organisms that are genetically very distinct to avoid a large number of shared peptides between both organisms.

In what follows, I will use the CPTAC study 6 dataset to demonstrate the effect of each preprocessing step. In the 6th study of the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC), a trypsin-digested mix of 48 human proteins (Universal Protein Standard 1, UPS1) was spiked in 5 different concentrations into a mix of trypsin-digested *Saccharomyces cerevisiae* proteins. These concentrations were: 0.25 fmol/µL (sample 6A), 0.74 fmol/µL (sample 6B), 2.2 fmol/µL (sample 6C), 6.7 fmol/µL (sample 6D) and 20 fmol/µL (sample 6E). These samples were sent to five different labs and analyzed on six different mass

spectrometers[30]. For convenience, I only used the data from the LTQ-orbitraps at sites 56, 65 and 86. The study was conceived to provide a benchmark dataset to compare the power of different approaches to detect differential abundance and it is the most well-known quantification benchmark dataset. Indeed, Paulovich *et al.* were cited 121 times in Scopus on February 7th, 2019.

This dataset is ideal to demonstrate the statistical concept of blocking. Indeed, a typical proteomics experiment is often restricted by spatiotemporal constraints that prevent the generation of all samples simultaneously, with the same equipment, etc. Such variation causes unwanted technical variability (noise) and makes it harder to detect the effect of interest. Blocking is the experimental design choice to assign the experimental units (e.g. MS runs) to different "blocks" (e.g. batches, periods in time) in such a way that the treatments that need to be compared are present within each block. This enables to estimate the treatment effect within each block, in which the noise is lower. Blocking is therefore a way of removing unwanted variability and thus increases the power of a statistical analysis. In the CPTAC dataset, we retained the data from three different laboratories. Since the between-lab variability is not of interest, "lab" can be considered as a blocking factor, which allows to remove the between-lab variability from the analysis.

Note that there are issues with ionization competition in the CPTAC dataset [421]. Due to the relatively high spike-in concentrations of the UPS1 mix, ionization of UPS1 peptides will partially suppress the intensities of the yeast peptides. This might lead to the false positive calling of yeast proteins as DA. We indeed noticed that most quantification methods could not control their quantification false discovery rates at the 5% level when large differences in spike-in concentrations were compared [422]. Fig. 3.5 gives an overview of CPTAC Study 6, which was used in all my first-author manuscripts.

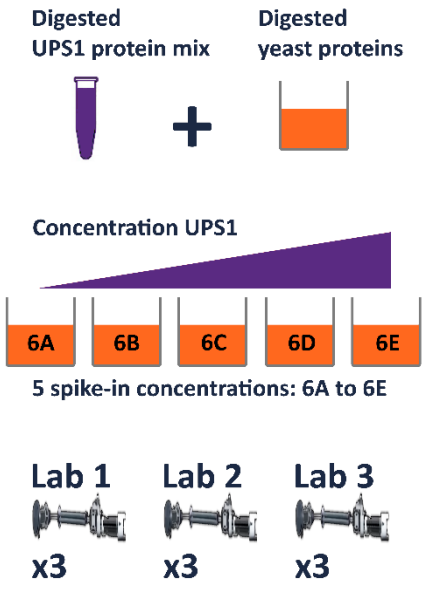


Figure 3.5. Overview of the subset of the CPTAC Study 6 that will be used throughout this thesis. Digested human UPS1 proteins were spiked in five different concentrations (6A – 6E) into digested

[30] "Lab 65" analyzed the samples on two different machines

yeast (*Saccharomyces cerevisiae*) proteome. These samples were sent to three different laboratories. Each laboratory analyzed the samples in technical triplicates.

Other, more recent spike-in datasets were published in Ramus *et al.* (2016) [423] and Jarnuczak *et al.* (2016) [330].

There might be some confusion on the usage of the term “sample”. In this section, I denoted the different spike-in conditions as “samples”, which were repeatedly analyzed. However, in a biological experiment, multiple samples will be taken for each treatment condition of interest and each of these samples will be analyzed in one or more technical replicates on the mass spectrometer. Because of the ever-increasing sequencing depth, technical replication is often omitted in modern experiments, which causes each “sample” to correspond to a single MS run. In the literature, indicators referring to “sample” or “MS run” are therefore almost always used interchangeably. To avoid confusion, I will use the terms “condition” or “treatment” to refer to the spike-in conditions from now on.

4. DIFFERENTIAL PROTEIN ABUNDANCE ANALYSIS

Once peptides are identified, linked to a protein and assigned an intensity value, the data can be used for the analysis of differential protein abundance. In this chapter, I will go deeper into the different steps in a typical differential protein abundance analysis workflow. I will start with data preprocessing, followed by a demonstration of the most important methods to use the preprocessed data for differential protein abundance analysis.

4.1. Preprocessing

Because of the nature of the peptide-level data described in section 3.4, preprocessing is required before proteins can be quantified. There is a plethora of preprocessing workflows and many of these are often constructed *ad hoc*. Nonetheless, in most workflows, the data typically undergo some kind of (log-)transformation, filtering, normalization, imputation and summarization. Note that the order in which each of these preprocessing steps are executed impacts on the final results. Transformation, in principle can be executed at any stage in the preprocessing workflow, but typically occurs early because it makes the data easier to handle. Similarly, filtering is done early in the workflow to remove those observations that are deemed unreliable and/or unwanted for various reasons. Furthermore, Karpievitch *et al.* (2012) showed that normalization followed by imputation generally outperforms imputation followed by normalization [424]. The rationale behind this is that imputing missing values obscures possible bias trends and therefore renders normalization less efficient. Moreover, imputing missing values does not make sense when unaccounted systematic bias is still present in the data. Finally, data summarization is best done after imputation as it is much more difficult to summarize data that contains missing values [425]. Nonetheless, some workflows, such as the default MaxQuant-Perseus workflow impute the data only after summarization (see 4.1.4). In this section, I discuss each of the preprocessing steps in more detail.

4.1.1. Transformation

The distributional properties of raw intensity measurements are often unfavorable for direct statistical modeling. Therefore, nearly every quantitative proteomics workflow involves a transformation of the raw intensity values. Log-transformation is a logical choice because raw intensity and concentration measurements are often more or less log-normally distributed: their values are always positive, they are skewed to the right and their variances increase with the mean. The strong right-skewness of the raw intensities is shown in Fig. 4.1: there are many relatively low intensities and only a few very high intensities. After log-transformation, the distributions become much more symmetrical.

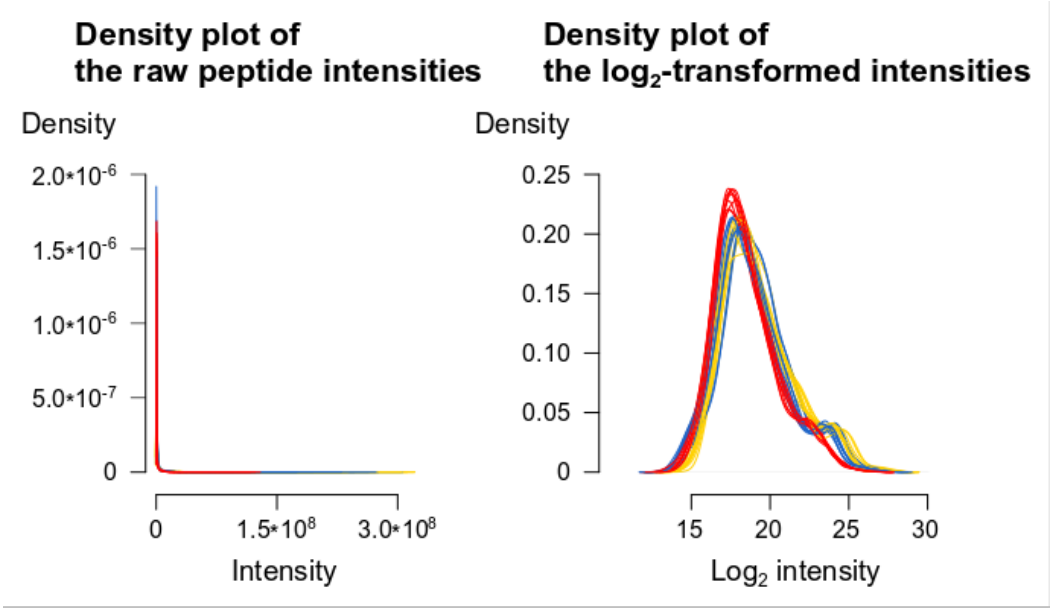


Figure 4.1. Impact of $\log_2$ transformation on the raw peptide intensities in the CPTAC dataset *[426]*. Left: the densities of the raw peptide intensities are strongly skewed to the right. Right: the densities of the $\log_2$-transformed peptide intensities are much more symmetrical. The densities are colored according to lab: red corresponds to the orbitrap at site 56, yellow to the orbitrap at site 65 and blue to the orbitrap at site 86.

Moreover, the variance structure of the raw intensities is often multiplicative: the variability in the data tends to be higher for higher intensities than for lower intensities (Fig. 4.2). Log-transformation will stabilize the variances by transforming a multiplicative error structure into an additive error structure [263]. In an additive error structure, the variability in the data is independent of the mean. The statistical property of equal variances is called homoscedasticity (as opposed to heteroscedasticity: unequal variances). The assumption of homoscedasticity opens the way to use the standard toolbox for statistical inference, such as linear regression. Such classic inference methods often provide closed-form solutions for their estimators, which makes the estimation procedures much simpler and faster. In practice, there sometimes remains a mildly positive mean-variance correlation in the $\log_2$-transformed intensities.

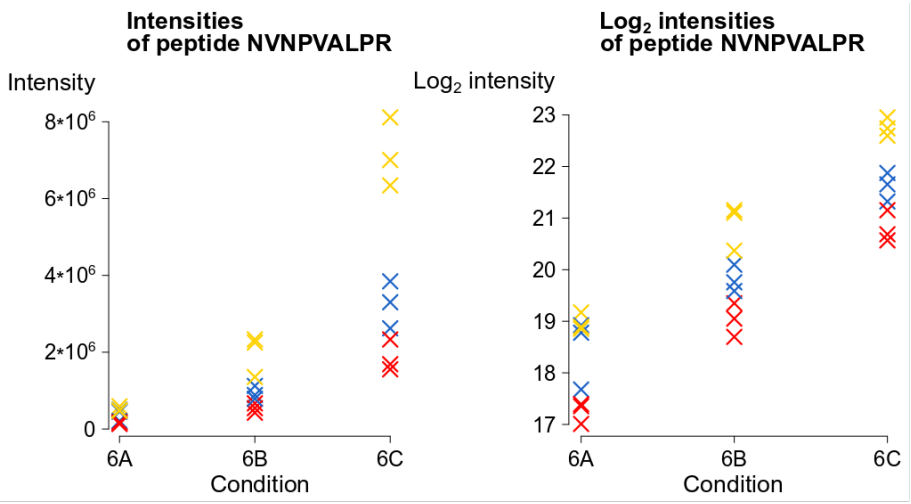


Figure 4.2. Raw intensities (left) and $\log_2$-transformed intensities (left) of peptide NVNPVALPR, which is part of the human UPS1 protein P08311 (cathepsin G) in the CPTAC dataset [426]. The variability in the peptide's raw intensities increases with higher spike-in concentrations but remains constant for the $\log_2$-transformed intensities.

A final argument in favor of the log-transformation is that by modeling PSM intensities at the log-scale, one models proportional differences at the biological scale, which is also more relevant from a biological point of view. Most researchers choose a log-transformation with a base of 2. A one-unit increase in $\log_2$-abundance will then be equivalent with a factor 2 increase in protein abundance.

4.1.2. Filtering

The data are filtered prior to differential analysis to remove peptides and proteins that are *a priori* uninformative from a biological or statistical perspective. Removing uninformative peptides increases the power to detect differential abundance in their corresponding proteins. Removing uninformative proteins reduces the number of proteins that needs to be tested for differential abundance. This will result in a less severe multiple testing correction and thus in a higher statistical power [427].

Examples of peptides or proteins that are typically filtered out include:

- decoy peptides
- typical contaminants such as keratin, which originate from the operator's skin amongst others
- other highly abundant and possibly less informative proteins (e.g. RuBisCo in plant samples)
- shared peptides, being peptides that map to more than one protein. Sometimes, these shared peptides are assigned to a protein (group) with the most unique peptides, but this practice should be discouraged because a shared peptide's intensity could very well represent the combined intensity of multiple proteins.
- proteins or peptides identified in only a few samples
- proteins identified with only one or a few unique peptides (e.g. the two-peptide rule)
- proteins for which no peptides without a modification site are identified. The rationale here is that the identification FDR is more difficult to calculate for a peptide carrying a modified site or amino acid mutation. Some search engines, such as MaxQuant, are less certain about the identification of a peptide with a modification and prefer to filter out proteins that are identified with only modified peptides, rather than doing inference on a protein of which they are less certain that it is really present in the data.
- peptides with very variable retention times over different runs, as this might be an indication for misidentification [428].

According to the vignette of the R package genefilter [429], a good filtering criterion should adhere to three criteria:

1. It should be statistically independent from the test statistic under the null hypothesis (i.e. the protein is not differentially abundant).
2. It is correlated with the test statistic under the alternative hypothesis (i.e. the protein is differentially abundant).
3. Filtering based on the criterion does not notably change the dependence structure (if it exists) of the joint test statistics.

The second property provides a benefit for filtering as it enriches for differentially abundant proteins after filtering. The first and the third criterion are necessary to keep control over the false discovery rate of the subsequent analysis at the pre-specified level (see 4.2.5). Indeed, as long as these criteria are fulfilled, filtering will not result in a biased analysis. The

aforementioned filtering procedures all fulfill the first criterion[31]. However, filtering on a criterion such as fold change estimates would induce a biased downstream quantification, as the fold change is an integral part of almost every test statistic and hence strongly correlates with it, also under the null hypothesis. The effect of filtering on different quantification methods has been studied in Belouah *et al.* (2019) [430].

4.1.3. Normalization

Even in a very clean, synthetic dataset as CPTAC, where there is no biological variability and only the spiked 48 UPS1 proteins are differentially abundant, the marginal peptide distributions are quite distinct. Considering all proteins in the dataset, there are considerable effects between samples with different spike-in concentrations, even within the same lab (Fig. 4.3, left). Moreover, for replicate measurements of the same samples, there is considerable lab-to-lab variability, and even the within-lab variability is non-negligible (Fig. 4.3, right).

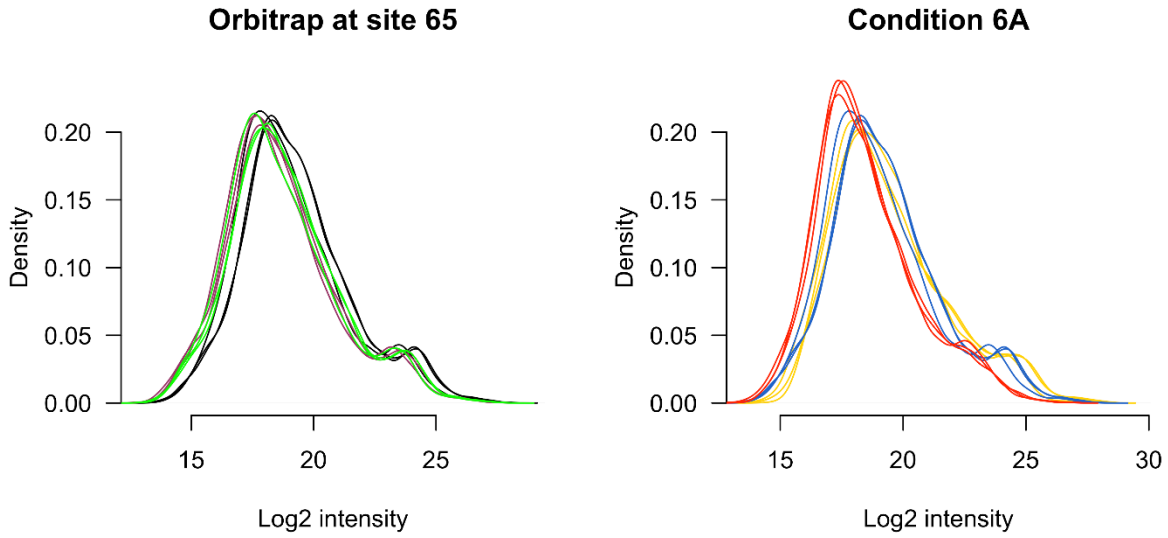


Figure 4.3. Left: density plot of the $\log_2$ peptide intensities for the orbitrap at site 65 in the CPTAC dataset [426]. The densities are colored according to spike-in condition (black: spike-in condition 6A, dark red: spike-in condition 6B and green: spike-in condition 6C). Right: density plot of the $\log_2$ peptide intensities for spike-in condition 6A. The densities are colored according to lab (red: orbitrap at site 56, yellow: orbitrap at site 65 and blue: orbitrap at site 86).

Normalization aims to remove, or at least dampen, this potentially large, unwanted variability. Center mean and center median normalization subtract the respective means or medians from each distribution (Fig. 4.4). These simple normalization approaches aim to remove non-biological variability by centering the peptide intensity distributions and do not impact on their shapes [431]. However, it is clear from Fig. 4.4 that the shapes of the distributions are also affected by technical variability. Hence, more advanced normalization procedures are needed.

A method that has proven to work well for microarray data is quantile normalization [432, 433]. Quantile normalization will impose the same density distribution upon each MS-run. Here, the peptide intensities are sorted from low to high. The lowest peptide intensity for each run is then set equal to the mean of the lowest peptide intensities over these runs. Similarly, the second-lowest peptide intensity in each run will be equal to the mean of the second-lowest peptide

[31] In practice, criterion 3 is rarely problematic, as most filtering procedures do not noticeably change the correlation structure of the tests [429].

intensities in each run, and so on. Missing values are handled based on the assumption that the data are missing at random.

Linear regression is a versatile statistical framework that can also be used to normalize the data (more about linear regression in section 4.2) [424, 434, 435]. VSN normalization is an example of a regression-based normalization approach that simultaneously executes transformation and normalization [436]. In brief, VSN normalization assumes that the different raw intensities from the different MS runs can be brought onto the same scale through linear mappings. VSN normalization assumes that the variance of the raw intensities v_p depends on the mean μ_p as follows:

$$v_p = v(\mu_p) = (c_1\mu_p + c_2)^2 + c_3, \qquad \text{(Eq. 4.1)}$$

with $c_3 > 0$. Based on these assumptions, a transformation h is proposed such that the variance is approximately independent of the mean. Then, the following statistical model is proposed:

$$h_r(Y_{pr}) = \mu_p + \varepsilon_{pr}, \qquad \text{(Eq. 4.2)}$$

for all $p \in p^{\text{null}}$. Here, Y_{pr} is the raw intensity for peptide p in MS run r, and p^{null} is the set of non-differentially abundant peptides. The parameters of the model are estimated with a least trimmed sum of squares regression under the assumptions that $\mathrm{E}[\varepsilon_{pr}] = 0$ and that the variance of the error term is constant: $\mathrm{Var}[\varepsilon_{pr}] = \sigma^2$. A more detailed explanation of the VSN algorithm can be found in Huber *et al.* (2002) [436]. Both older and more recent publications that compared the performance of different normalization methods seem to indicate that linear regression-based methods, such as VSN, generally outperform other normalization methods [435, 437, 438], although the relative performance of different normalization methods is also strongly dataset-dependent [439].

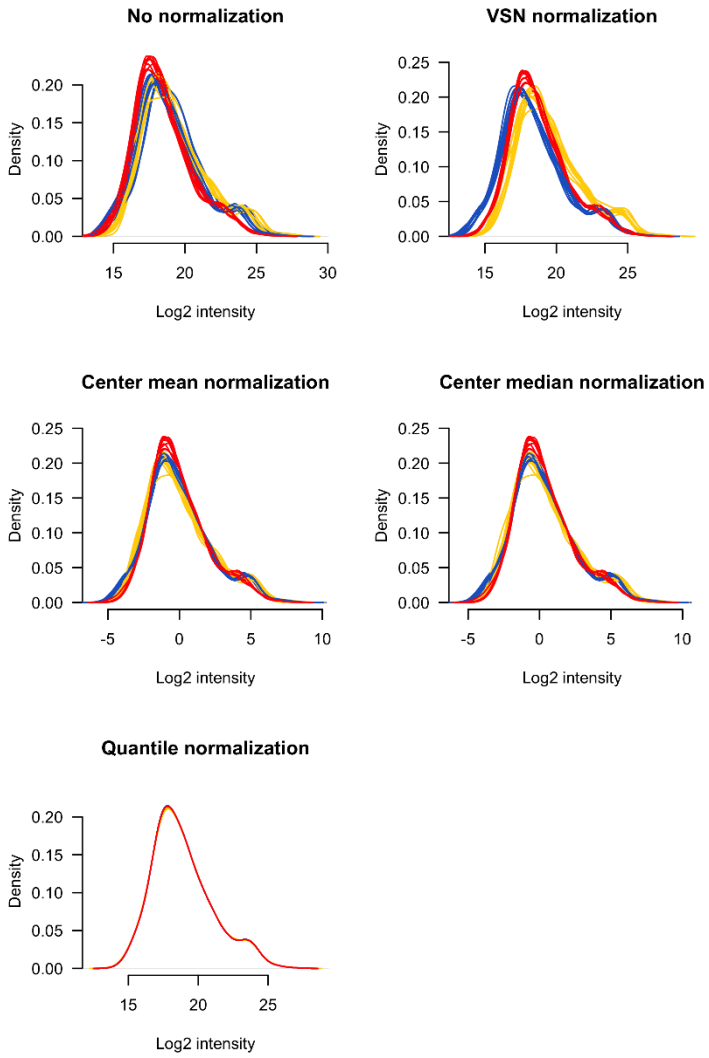


Figure 4.4. Overview of the effects of different types of normalization on the peptide intensity distributions in the CPTAC dataset [426].

The MaxLFQ summarization algorithm described in 4.1.5 combines summarization with normalization, although MaxLFQ summaries are sometimes still normalized afterwards.

An important assumption of all these normalization approaches is that the abundance of the large majority of the peptides remains unchanged over the different treatments. This assumption is often reasonable because researchers are mostly interested in biological perturbations that affect very specific pathways in the cell, thus affecting a minority of proteins. Moreover, most normalization methods can tolerate quite large fractions of differentially abundant peptides, as long as there is a more or less equal number of up- and downregulated peptides. This premise is also quite reasonable since the total amounts of peptides (in µg) analyzed in each run are as equal as possible.

The assumption of no major changes in the bulk of the proteome is however problematic for specific studies. For example, in AP-MS studies, proteins are purified that specifically interact with a protein of interest (bait). The control group contains only “background proteins”, i.e. proteins that non-specifically interact with the bait. Therefore, a large fraction of the identified proteins is more abundant in the samples with the bait, while the background proteins are, in

theory, equally abundant between control and bait samples. In such kinds of experiments, normalization should be performed with extreme caution [407].

Note that even though normalization intends to remove the overall run-to-run variability, considerable block effects can persist at the level of the individual peptides. Fig. 4.5 shows a multidimensional scaling (MDS) plot after quantile normalization for the CPTAC dataset. Even with a radical normalization method like quantile normalization, which literally forces the $\log_2$-transformed intensity distributions to be equal in each MS run, the runs clearly cluster together per lab. This demonstrates that it will be necessary to also correct for blocking effects further down the analysis pipeline.

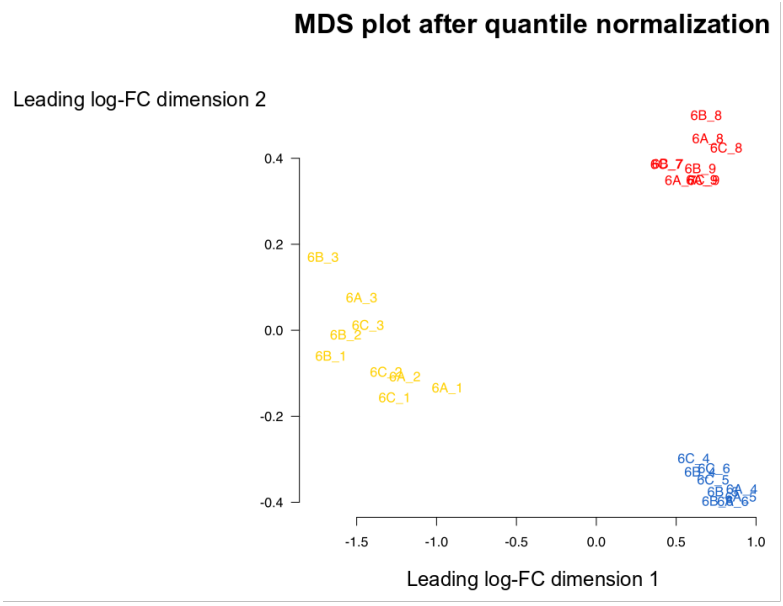


Figure 4.5. Multidimensional Scaling (MDS) plot after quantile normalization for the CPTAC dataset [426]. The MDS plot shows each MS run in such a way that the distance between each pair of runs is equal to the root-mean-square deviation for the top 500 peptides that are the most distinct between the pairs of runs.

4.1.4. Imputation

To demonstrate the important aspects of missingness, we investigated the amount of missing values at the peptide level in 73 recent label-free shotgun proteomics datasets. We showed that on average 44% of all values at the peptide level are missing (see chapter 10). To cope with such large amounts of missing values, they are often replaced with substitute values in a process called imputation.

Classically, three types of missingness can be defined: missingness completely at random (MCAR), missingness at random (MAR) and missingness not at random (MNAR) [440]. MCAR assumes that the missing values cannot be explained by the nature of their underlying true values, nor by any known covariate: every value in the data matrix has an equal probability of being missing. MAR is a type of missingness whereby the probability of an observation to be missing is dependent on one or more observed covariates, but independent of the nature of the underlying values themselves. MNAR are all cases where the missingness is dependent on the underlying values (and optionally also on one or more known covariates): some values (e.g. very low values, very high values) have a higher probability of being missing than others.

Missingness in label-free shotgun proteomics datasets is a combination of missingness completely at random (MCAR) (e.g. an enzymatic modification in one experimental condition

might cause a peptide to be unidentified if that modification was not accounted for during the search), missingness at random (MAR) (e.g. certain peptide sequences ionize more easily than others; therefore, missingness is much more likely for poorly-ionizing peptides) and missingness not at random (MNAR) (e.g. more abundant peptides simply have a higher chance of getting fragmented and thus being identified). Note that this MNAR is exacerbated as the probability for a peptide to be identified is also context-dependent: when co-eluting with many other highly abundant peptides, a peptide will have a smaller chance of getting identified than if these other peptides would be absent or lower in abundance.

When choosing an imputation strategy, it is important to keep in mind the assumptions of that imputation strategy as most imputation strategies make use of either a MCAR or a MNAR assumption, but not both.

k-nearest neighbors (kNN) imputation is an example of an MCAR imputation strategy. In kNN, a Euclidean distance metric is calculated on all peptide intensities. Based on this distance matrix, the *k* most similar peptides (neighbors) are identified for each peptide that has at least one missing value. All missing values for that peptide are then imputed with the average of the corresponding (non-missing) values from the *k* neighbors [441, 442].

Quantile Regression Imputation of Left Censored data (QRILC) imputation is an example of an MNAR-based imputation strategy [443]. In QRILC, missing values are imputed with random draws from a truncated distribution with parameters that are estimated using quantile regression. QRILC has been implemented in the MSnbase R/Bioconductor package for manipulation, processing and visualization of proteomics data [444].

The popular proteomics computational platform Perseus also makes use of an MNAR-based imputation strategy. In Perseus, imputation is achieved by imputing the data with random draws from a rescaled, down-shifted normal distribution [445]. The characteristics of this distribution are calculated based on the data. More specifically, its mean is equal to the average of all the observed data minus *d* times the standard deviation of the observed data. Its standard deviation is equal to *w* times the standard deviation of the observed data. The default values for *w* and *d* are 0.3 and 1.8, respectively.

The current version of the popular Bioconductor package MSstats (version 3.12.2) [446] uses a more advanced, model-based approach to impute missing values under a MNAR assumption. Their accelerated failure time (AFT) model (see 4.2.3) does not incorporate a random missingness component as the authors argue that due to the improved technology, the proportion of random missing values has become negligible.

Choosing for no imputation, MCAR-based imputation or MNAR imputation can have a big impact on the downstream analysis. For example, the MCAR-based kNN is more suited when the majority of the missing values is not intensity-dependent [447]. Contrary, MNAR-based methods like QRILC, Perseus and AFT model imputation, perform better in a context with relatively more intensity-dependent missing values. These MNAR-based methods might however perform poorly in detecting special cases, e.g. where a long isoform of a certain protein is absent, but a smaller isoform is strongly upregulated (Fig. 4.6). In such cases, imputation with low-intensity values might dilute the signal and obscure the classification of the given protein as DA.

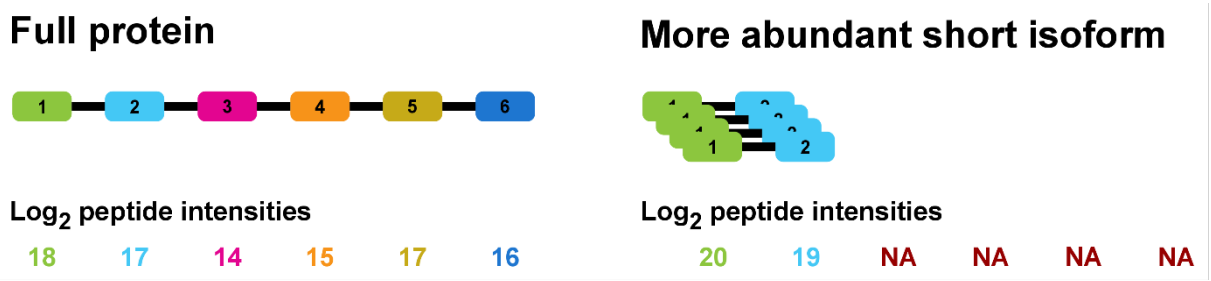


Figure 4.6. Imputation with low-intensity values can dilute signals present in the data. In this theoretical example, a protein consisting of 6 tryptic peptides is present in the left condition, while only a short isoform of the same protein giving rise to tryptic peptides 1 and 2 is 4 times more abundant in the right condition. When assessing peptides 1 and 2, the protein $\log_2$ fold change is equal to 2. However, imputing the missing values ("not assigned", NA) with either MCAR or MNAR methods will dilute this signal.

Indeed, although imputing missing peptide values was suggested in the proteomics literature, imputation should always be used with caution. When nothing is known about the nature of the missing values, it has been suggested to use MCAR imputation approaches based on local similarity, as these perform well on average [424]. It has to be noted however, that the performance of an imputation approach is highly dataset-dependent [422, 447-449]. In reality, missing values are often caused by an unknown mix of intensity-dependent and -independent mechanisms, which is strongly dataset-specific [424, 447] and choosing the wrong imputation method for the dataset at hand can result in a severe backlash in performance [422].

Some imputation methods try to combine MCAR and MNAR imputation. For example, one of the imputation strategies in the DEP Bioconductor package by Smits and Huber suggests an imputation method whereby proteins for which the values are completely missing in one or more experimental conditions are imputed with a MNAR method, while the other missing values are imputed with a MCAR method [450]. However, this distinction is rather arbitrary, since for some proteins, all values in an experimental condition might also be missing due to random chance. Conversely, some missing values for proteins which are detected in all experimental conditions might still be due to low intensities.

A final issue with imputation is that, even if the true mechanism of missingness would be known, the uncertainty caused by replacing a missing value by a fixed value from a certain distribution is essentially ignored. A correct data analysis strategy should take this uncertainty into account. This problem might be solved by using a multiple imputation strategy [451, 452] in which the dataset is imputed multiple times and each of these imputed datasets is subsequently analyzed. The variability in the outcomes gives a good idea of the impact of the imputation on the analysis. Unfortunately, multiple imputation has not yet been widely adopted in the field. Note that it is also possible to model mechanisms of missingness explicitly (see section 4.2.3) [440].

4.1.5. Summarization

As noted in section 3.4, differential analysis mostly takes place at the protein level, but the data are at the peptide level. Therefore, most workflows involve some kind of summarization. In this section, I will focus on peptide- to protein-level summarization to show the effects of different summarization techniques. A simple way to summarize is by summing up all raw peptide intensities that correspond to each protein in each MS run [453]. Alternatively, mean summarization involves taking the mean of the peptide intensities to obtain a protein-level summary. Median summarization is also very common as a median is insensitive to outlying peptide intensities [384]. For the same reason, weighted means [454, 455] or medians [456],

whereby the outlying peptides are given less weight, or trimmed means [299] have also been proposed.

Although these techniques are very simple to apply, one needs to consider a few things. The first one is the intensity-dependent missingness. Indeed, in samples with a high concentration of a particular protein, more of its poorly ionizing peptides are expected to be found compared to samples with a lower concentration of that protein. However, such poorly ionizing peptides reduce the protein concentration estimates in the samples where the protein is highly abundant. Therefore, a naive mean or median summary that does not correct for peptide ionization efficiency produces fold change estimates that are biased towards 0. This fold change bias is demonstrated in Fig. 4.7.

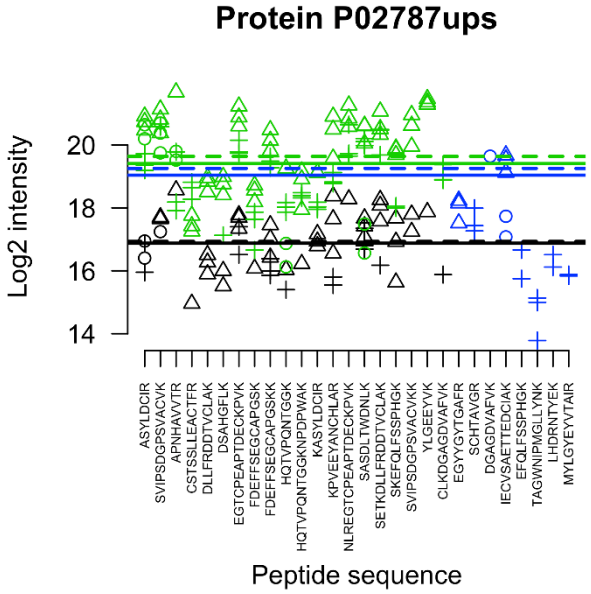


Figure 4.7. Effect of intensity-dependent missingness on mean and median summarization. The figure shows the $\log_2$-transformed intensities for all identified peptide sequences of a UPS1 protein in the CPTAC dataset [426] in the low spike-in condition 6A (black) and the higher spike-in condition 6C (green and blue). Symbols denote different labs (plus: site 56, triangle: site 65, circle: site 86). All peptides identified in condition 6A were also identified in condition 6C. Blue are the peptides which are exclusively identified in condition 6C. Full lines denote the mean summaries, dashed lines the median summaries. The black lines are the summaries for condition 6A, the blue lines the summaries for condition 6C. The green lines are the summaries for condition 6C when the peptides exclusively identified in condition 6C (blue) are omitted. Omitting those peptides increases both the mean and median summaries for condition 6C.

To avoid this issue, it is of course possible to base the protein summaries only on the overlapping peptides. However, when many different conditions are compared, the number of peptides that is identified in every condition tends to be very low, which makes it impossible to obtain such a protein-level summary for many proteins in the dataset.

MaxLFQ, the algorithm that is used to summarize proteins in MaxQuant, addresses this issue by making use of only those PSMs that overlap between each pair-wise MS run comparison. [317]. A schematic overview of the MaxLFQ algorithm is given in Fig. 4.8.

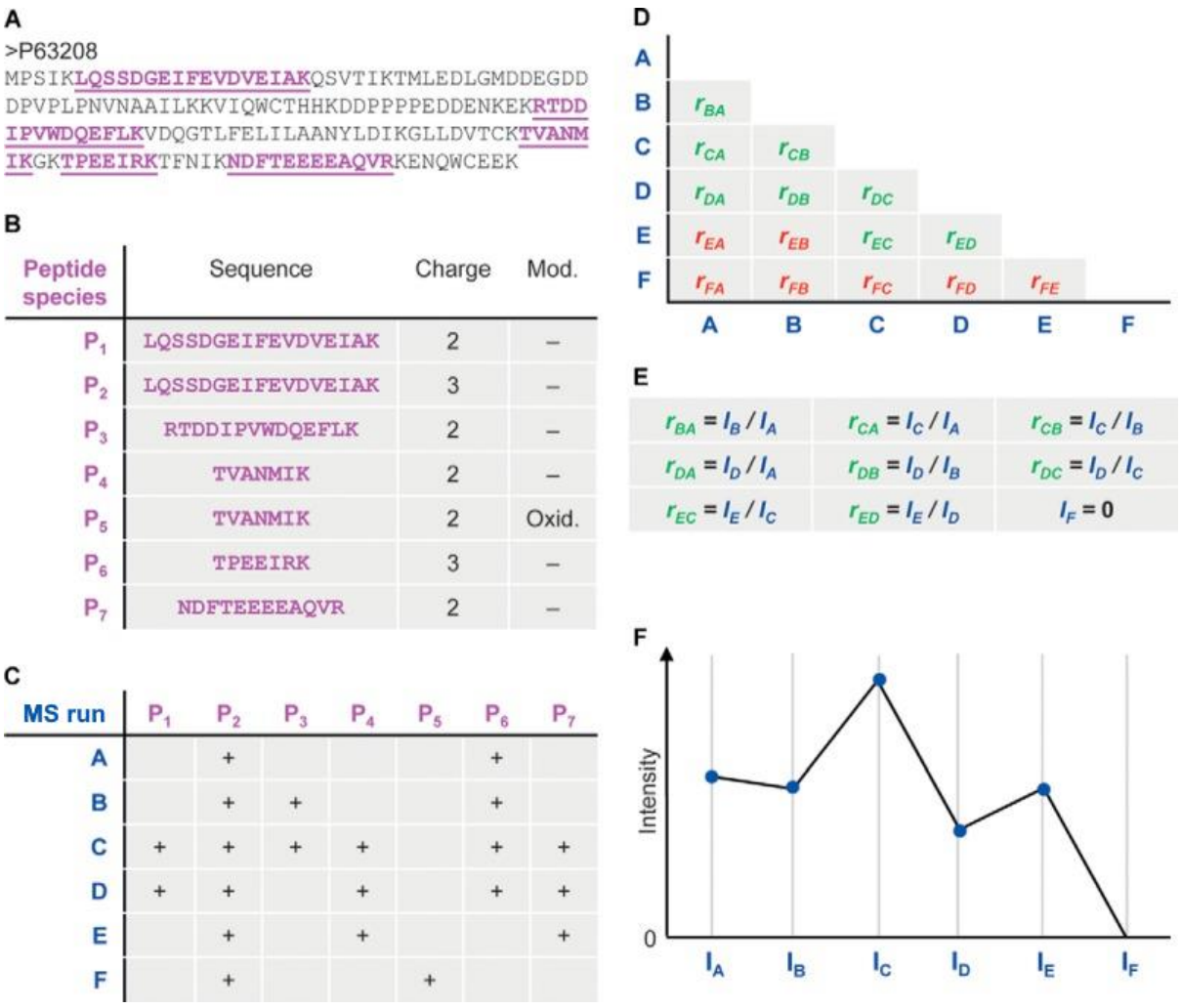


Figure 4.8. (A) Example of a protein of which five peptide sequences (indicated in magenta) are detected. (B) These five peptides were identified as seven different PSMs ("peptide species"). (C) Occurrence of each of the seven PSMs in 6 exemplary samples A – F, each of which was run once on the mass spectrometer. (D) Matrix with the pairwise protein ratios. Protein ratios are calculated by taking the median of all valid normalized pair-wise PSM ratios. Valid protein ratios are ratios for which two or more PSMs are in common between both runs (green). Protein ratios for which less than two PSMs are in common are considered invalid (red). (E) System of equations that needs to be solved to obtain the MaxLFQ protein intensities per run. MaxLFQ intensities for runs for which no valid protein ratios exist (e.g. run F) will be set to zero (i.e. a missing value on the log-scale). (F) Run-wise MaxLFQ protein intensities for the given protein. MaxLFQ intensities are calculated by solving the equations in (E) through least squares and rescaling the result to maintain the total summed intensity over all runs. Image adapted from Cox *et al.* (2014) [317], © 2014 by The American Society for Biochemistry and Molecular Biology, Inc., CC BY 4.0.

The rationale behind MaxLFQ is the following:

The intensity for each PSM is calculated as the area under the isotopic envelope at the maximum intensity over the retention time profile multiplied by a run-wise normalization factor. These normalization factors are calculated by least-squares minimization of the overall pair-wise log fold changes for all PSMs between all runs. As with most normalization methods, the assumption is made that the large majority of the proteome is not differentially abundant.

Then, the common PSM intensities between each run pair q and r are used to calculate PSM ratios. The pair-wise protein ratio ϱ_{qr} between runs q and r is then equal to the median of all pair-wise PSM ratios between runs q and r.

Based on all pair-wise protein ratios, it is then possible to calculate log-transformed protein-level intensities y_q for each run q. This is done by performing the following protein-wise least-squares analyses for each "valid" pair of runs q and r:

$$\log \varrho_{qr} = y_r - y_q + \varepsilon_{qr} \qquad (\text{Eq. } 4.3)$$

Herein, ϱ_{qr} is the pair-wise protein ratio between runs q and r, y_q the log-transformed protein intensity in run q and y_r the log-transformed protein intensity in run r. ε_{qr} is a random error term. Pairs are considered valid if they have at least two PSMs in common. Finally, the whole profile of the estimated run intensities $\hat{y}_q$ is rescaled to maintain the total summed intensity for a protein over all runs. Important to note is that in this procedure, summaries are calculated solely based on the PSMs that are common between each pair of runs. Therefore, MaxLFQ does not suffer from a downwards bias in its summary estimates.

It is also possible to reformulate the summarization problem as follows:

$$y_{fr} = \beta^0 + \beta_f^{\text{feature}} + \beta_r^{\text{run}} + \varepsilon_{fr}, \qquad (\text{Eq. } 4.4)$$

Herein, y_{fr} is the log-transformed intensity for PSM (feature) f in run r, β^0 is the intercept, which corresponds to the average log-transformed intensity of a certain reference PSM in a certain reference run. β_f^{feature} is the effect of PSM f relative to the intercept and β_r^{run} is the effect of run r relative to the intercept. ε_{fr} is a random error term. The MaxLFQ procedure is in fact an *ad hoc* procedure to fit such a model for the ratio y_{fr}/y_{fq}, conditional on all PSMs f that are in common between runs q and r. The disadvantage of the MaxLFQ procedure is that if the PSM overlap between runs q and r is very limited, the MaxLFQ estimates become very imprecise. This is the reason that MaxLFQ requires an overlap of at least two PSMs before allowing a ratio to be valid.

It is however more efficient to fit model $(\text{Eq. } 4.4)$ as it is, as this model uses the information in all the PSMs, not only those that overlap, and still corrects for peptide-specific effects thanks to the β_f^{feature} effect. $\beta^0 + \beta_r^{\text{run}}$ can then be interpreted as the average protein intensity in run r for the reference PSM.

Median polish [457, 458] is a robust way of fitting model $(\text{Eq. } 4.4)$ that is also implemented in the current version of MSstats, but still seems show a slightly downwards bias. MaxLFQ, conversely, produces nearly unbiased protein-level estimates (Fig. 4.9).

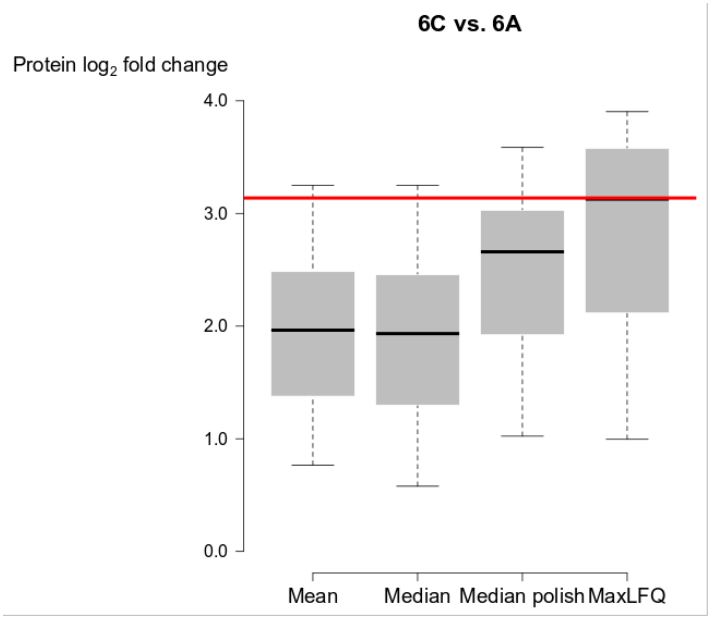


Figure 4.9. Overview of the $\log_2$ fold change estimates between condition 6C and 6A for the 36 UPS1 proteins for which these estimates could be calculated based on $\log_2$-transformed PSM-level intensities with four different summarization methods: mean summarization, median summarization, median polish and MaxLFQ. The red line denotes the true $\log_2$ fold change based on the known spike-in concentrations (2.2 fmol/µL for 6C, 0.35 fmol/µL for 6A). Mean and mean summarization strongly underestimate the true fold change. Median polish shows a smaller downwards bias, while MaxLFQ is nearly unbiased.

To avoid the downwards bias introduced by intensity-dependent missingness, MSstats imputes the data under a missing-by-low-intensity assumption prior to median polish summarization. However, such an assumption is not always valid, as already discussed in 4.1.4.

Note that both MaxLFQ and MSstats start from PSM-level intensities without taking into account the fact that PSMs mapping to the same peptide sequence are correlated.

4.2. Methods for differential protein abundance analysis

Differential analysis here aims at identifying those proteins that are differentially abundant. In proteomics, there are three main methods to perform differential analysis: summarization-based methods, counting-based methods and peptide-based methods. However, differential analysis is only meaningful if the design of the study allows for it. Therefore, I will start this section with a note on the importance of the study design.

4.2.1. The importance of study design

In its early days, mass spectrometry was tedious, time-consuming and costly. The main reason for this was the low duty cycle of the mass spectrometers, implying that, within a given time frame, very few peptides got selected for fragmentation and could thus be identified. Intelligent approaches such as MudPIT [119] and ICAT [459], countered this by peptide pre-fractionation or by selecting for so-called protein-representative peptides respectively. The former increased the overall analysis time, whereas the latter relied on expensive reagents that also tended to interfere with peptide fragmentation and peptide identification. Hence, samples were often analyzed only once on mass spectrometers. Proteins were then declared “significant” solely based on a fold change threshold [241, 243]. Publications using this approach are sometimes still accepted in high-impact journals [460]. Alternatively, a normal distribution was fitted to all fold change estimates and fold changes for proteins in the upper and lower 2.5% quantiles

were declared “significant” [455, 461]. Some authors even developed advanced empirical Bayes methods to deal with single-run experiments [462]. However, these methods provide little to no evidence about which proteins are truly differentially abundant. Indeed, with no information on the biological variability between biological repeats, it is impossible to assess how an estimator varies from experiment to experiment. For all we know, a protein with a very high fold change estimate can be in fact a protein whose abundance is highly variable, but unrelated to the studied treatment [463]. Hence, experiments without biological repeats make it impossible to infer the results towards the population. It is thus of utmost importance to design a study in such a way that samples are included from multiple, independent subjects from the population on which one aspires to do inference. For instance, if properly conceived, a study containing only BALB/c mice should provide results that are valid for all BALB/c mice. However, if a researcher wants to extrapolate these results towards other mice strains, he or she should have included at least a few different mouse strains in his/her experimental design. Fig. 4.10 demonstrates how different levels of replication contribute to the total variability in the system.

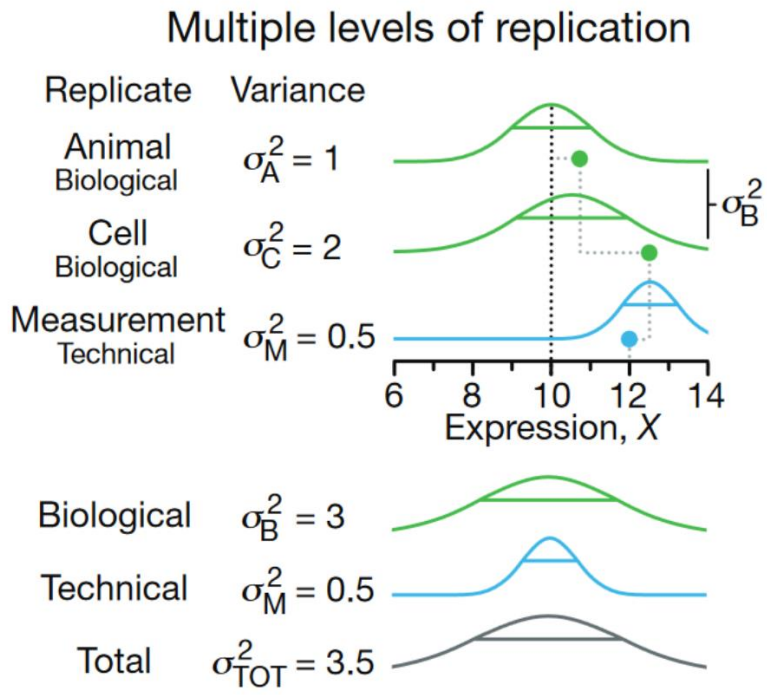


Figure 4.10. Different levels of replication do not contribute equally and independently to the total variability in the system. In the given example, there are two biological levels of replication (animal and cell) and one technical level of replication (measurement). Although the average expression of all animals in the population is equal to 10, each of these replication levels contributes to the total measurement variability by introducing, in this example, a random error that follows a normal distribution with variances 1, 2 and 0.5 respectively (the corresponding standard deviations are shown as horizontal lines). Note that in proteomics data, these levels of replication hold for every single protein. Moreover, proteomics data has multiple levels of technical replication: MS run, peptide and PSM, as discussed in section 3.4. Reprinted with permission from Blainey *et al.* (2014) [464], copyright © 2014, Springer-Verlag.

Still too often, researchers limit themselves to conducting a few “biological” replicates on the same cell line, often from the same vial, or worse, they run only one sample in a few technical replicates on the mass spectrometer. In the first case, the results can only be extrapolated to that specific cell line in that specific lab, but at least, the experimental variability was taken into account (i.e. difference due to slightly different handling of the cells, a slightly different temperature because the repeats were performed on a different day, etc.). In case only technical replicates are used, the results can only be extrapolated to that specific sample.

These types of improper study designs are, in my opinion, one of the reasons for the replication crisis that plagues the biological sciences. An extensive overview of the statistical considerations to bear in mind when designing an MS-based proteomics experiment can be found in Oberg and Vitek (2009) [465].

4.2.2. Summarization-based methods

Summarization-based methods for differential analysis start from protein-level summaries. In this section, I will explain a few of the most commonly used methods.

Perseus is one of the most popular software packages amongst mass spectrometrists to perform differential analysis [445]. It seamlessly imports MaxQuant output and is equipped with a user-friendly GUI that allows for a variety of data manipulations, statistical analyses and visualizations.

Perseus' default way to perform differential analysis between two groups is via t-tests on MaxLFQ-summarized, $\log_2$-transformed and preprocessed protein intensities. A t-test relies on three assumptions:

1. Independence. The information about any of the observations does not provide additional information about any of the other observations after correction for the treatment. This implies that all observations should be at the same level of hierarchy and that no pair of observations can be assumed (by design) to be more similar to each other than any other pair of observations (after correction for the treatment).
2. Normality. The normality assumption demands that the observations in both conditions are realizations of a normally distributed population.
3. Homoscedasticity. Homoscedasticity or equality of the variances means that the population variances in both conditions are equal. When the homoscedasticity assumption is not met, it is however still possible to use the Welch two-sample t-test (see below).

If the rigid assumptions of the t-test are not met, there is no guarantee that the inference will be correct. In practice, however, researchers seldom assess these assumptions, especially in high-throughput omics contexts. The rationale behind a t-test is to weigh the fold change of a protein by its natural variability in abundance. Indeed, as explained above, a high fold change is not very meaningful if a protein's abundance is very variable from sample to sample (Fig. 4.11).

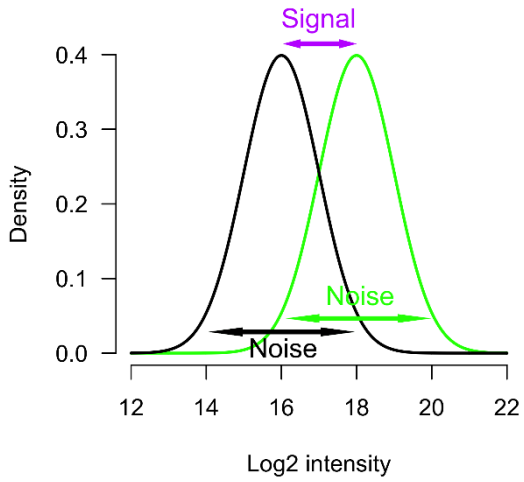


Figure 4.11. Illustration of signal and noise. If the signal increases, the confidence that a protein is differentially abundant between the green and the black condition will also increase. However, if for a constant signal, the noise increases, the confidence that a protein is differentially abundant will decrease. Hence, the signal-to-noise ratio is an ideal statistic to assess differential abundance.

The t-test will wrap the ratio of signal (fold change estimate) to noise (estimate of the variability in protein intensities) in a single test statistic t that estimates the signal-to-noise ratio. The t-test statistic performs superior compared to the use of simple fold change cut-offs because it also takes the noise into account. More specifically, the t-test statistic is defined as follows:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{\frac{1}{R_1} + \frac{1}{R_2}}} \qquad \text{(Eq. 4.5)}$$

Here, $\bar{y}_1$ is the average protein-level $\log_2$ intensity in the first treatment group, $\bar{y}_2$ the average $\log_2$ intensity in the second treatment group, R_1 the number of observations (MS runs) corresponding to the first treatment and R_2 the number of MS runs corresponding to the second treatment. s is the pooled variance estimate. It is calculated as follows:

$$s = \frac{1}{R_1 + R_2 - 2}\sum_{t=1}^{2}\sum_{r=1}^{R_t}(y_{rt} - \bar{y}_t)^2 \qquad \text{(Eq. 4.6)}$$

Here, $t = 1{,}2$ is the indicator for each treatment and $r = 1, \ldots, R_t$ the indicator for each MS run in a treatment. If the assumptions are correct, the test statistic follows a t-distribution with $R_1 + R_2 - 2$ degrees of freedom under the null hypothesis. If only the homoscedasticity assumption is not met, it is possible to use the Welch two-sample t-test instead. This option is also foreseen in Perseus. A Welch two-sample t-test omits the pooled variance estimator and instead calculates the t-test statistic as follows:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{J_1} + \frac{s_2^2}{J_2}}} \qquad \text{(Eq. 4.7)}$$

With s_1^2 and s_2^2 the sample variances in both treatments. This test statistic no longer follows a t-distribution under the null hypothesis, but it can be approximated as a t-distribution with an adjusted number of degrees of freedom through the Welch-Satterthwaite approximation.

The null hypothesis of a two-sample t-test states that there is, in reality, no difference in the average $\log_2$ intensities between both treatments. Next, the calculated test statistic t is confronted with the t-distribution (Fig. 4.12).

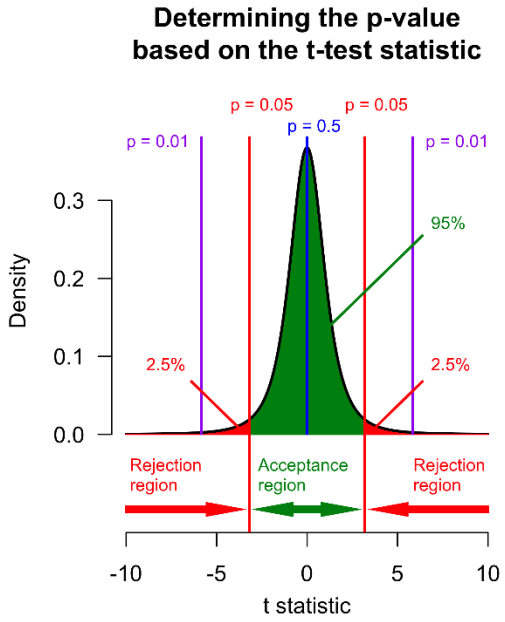


Figure 4.12. Illustration of the determination of p-values based on the t-test statistic. The p-value corresponds to the area under the t-distribution for which the t-statistic is as extreme as or more extreme than the observed t-statistic. The regions for which a t-statistic would be accepted and rejected at the 5% significance level given a t-distribution with three degrees of freedom are also given.

The percentile corresponding to the calculated test statistic can be easily converted into a p-value. This p-value denotes the probability that a new test statistic, calculated based on an independent repeat of the given experiment would be as extreme as, or more extreme than the observed test statistic, given that the null hypothesis is true. If this p-value is very small, it is not very likely to observe the given result under the null hypothesis. One then chooses to reject the null hypothesis and accept the alternative hypothesis, i.e. there is a real difference in the average $\log_2$ intensities between both treatments. The p-value threshold below which one chooses to reject the null hypothesis is called the significance level. Traditionally, this significance level is often set at 5%, but in fact, the choice of the significance level is up to the researcher. Other thresholds (e.g. 1%, 10%) can also be set, depending on the relative impact of falsely reporting non-differentially abundant proteins versus not reporting truly differentially abundant proteins.

In Perseus it is possible to use a moderated t-test statistic. This statistic is calculated as follows:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\tilde{s}}, \qquad \text{(Eq. 4.8)}$$

with:

$$\tilde{s} = s_0 + s\sqrt{\frac{1}{R_1} + \frac{1}{R_2}} \qquad \text{(Eq. 4.9)}$$

Hence, an offset s_0 is provided to the numerator of the t-test statistic. Providing a small offset reduces the impact of a protein's variance estimate s. Indeed, sometimes it happens that, due to random chance, the protein level estimates in the dataset are not very variable. This will

result in a small pooled variance estimate s, and hence a high test statistic t and a small p-value, even if a protein's fold change is rather small. Such proteins, with very small fold changes, but significant p-values due to small variance estimates are often not of interest to the researcher. Adding the offset s_0 stabilizes the test statistic. This procedure is known as significance analysis of microarrays (SAM) [466]. By adding the offset, the test statistic no longer follows a t-distribution and p-values are calculated by permuting the $\log_2$ protein intensities across all proteins over both treatments. The widespread use of the SAM procedure in the proteomics community, whereby s_0 is arbitrarily chosen by the experimenter has been criticized as it may lead to biased quantifications [467].

The use of t-tests and SAM limits Perseus analyses only to two-group comparisons. Consider the CPTAC dataset as an illustration. When comparing condition 6C to condition 6A for example, the independence assumption of the t-test is violated. Indeed, samples that were analyzed in the same lab are more similar than samples that were analyzed in different labs. Therefore, each sample contains some information about the other samples from the same lab, thereby invalidating the independence assumption. Summarizing the data from the sample- to the lab-level seems like a solution, but besides the loss of information, this does not solve the problem that data from the same lab over different conditions are more similar than data from different labs over different conditions.

Therefore, the t-test should be expanded towards a more general framework: linear regression. In the linear regression framework, every observed outcome variable y_r (with index $r = 1, \dots, R$ denoting the MS run) is assumed to originate from a linear combination of covariates x_{rm} and regression coefficients (also termed "parameters" or "effects") β^0 and β_m (with index $m = 1, \dots, M$ denoting the model parameters) summed with a random error term ε_r that covers the deviation of each observation y_r from its expected value under the model:

$$y_r = \beta^0 + \sum_{m=1}^{M} x_{rm}\beta_m + \varepsilon_r \qquad \text{(Eq. 4.10)}$$

The linear regression model has four assumptions:

1. Independence: Independence again denotes that none of the observations holds additional information about any of the other observations after correction for the covariates x_{r1} to x_{rM}.
2. Linearity: Linearity between the response and predictors means that the outcome variable varies linearly in function of the predictors x_{rm} and that there are thus no higher-order trends that cannot be accounted for. Linearity implies that the residuals (i.e. fraction of the data that cannot be explained by the predictors) have a mean of 0 and that they are orthogonal on the predictors.
3. Normality: The errors are assumed to be normally distributed, i.e. $\varepsilon_r \sim \mathrm{N}(0, \sigma^2)$.
4. Homoscedasticity: Homoscedasticity requires that the variance of the residuals is equal for each covariate pattern. This also implicates that are no trends in the spread of the residuals when the residuals are plotted in function of the fitted outcome values.

At a first glance, the linearity assumption seems to be rather restrictive for linear regression modeling. However, linear regression models can be easily adapted to capture higher-order (e.g. quadratic effects) or even non-parametric trends (e.g. splines).

Linear regression models are often written in a compact matrix notation:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad \text{(Eq. 4.11)}$$

For the CPTAC dataset, our aim is to compare the different conditions to each other. Therefore, separate linear regression models can be proposed for every protein i, whereby the protein-level $\log_2$ intensities are modeled in function of the spike-in conditions. By including lab effects, we also account for the blocked experimental design. As the peptide-level intensities are summarized to protein-level intensities in each run, we opt here to write an indicator r for run instead of j and we call R the number of runs for which a protein-level summary could be determined for protein i. Ultimately, the matrices can be specified as follows for each protein i in the CPTAC dataset, whereby the indicator i is suppressed for notational convenience:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \dots \\ y_r \\ \dots \\ y_R \end{bmatrix} \qquad \text{(Eq. 4.12)}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta^0 \\ \beta_2^{\text{condition}} \\ \beta_3^{\text{condition}} \\ \beta_2^{\text{lab}} \\ \beta_3^{\text{lab}} \end{bmatrix} \qquad \text{(Eq. 4.13)}$$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x_1} \\ \dots \\ \boldsymbol{x_r} \\ \dots \\ \boldsymbol{x_R} \end{bmatrix} = \begin{bmatrix} 1 & x_{12}^{\text{condition}} & x_{13}^{\text{condition}} & x_{12}^{\text{lab}} & x_{13}^{\text{lab}} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{r2}^{\text{condition}} & x_{r3}^{\text{condition}} & x_{r2}^{\text{lab}} & x_{r3}^{\text{lab}} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{R2}^{\text{condition}} & x_{R3}^{\text{condition}} & x_{R2}^{\text{lab}} & x_{R3}^{\text{lab}} \end{bmatrix} \qquad \text{(Eq. 4.14)}$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_r \\ \dots \\ \varepsilon_R \end{bmatrix} \qquad \text{(Eq. 4.15)}$$

Here, $\boldsymbol{y}$ is a vector containing all $\log_2$-transformed protein-level intensities y_r. The vector $\boldsymbol{\beta}$ contains the effect sizes: β^0 is a constant intercept, which refers to the average $\log_2$-transformed protein intensity in a certain reference condition 1 (e.g. spike-in condition 6A) in a reference lab 1 (e.g. LTQ-orbitrap at site 86). $\beta_2^{\text{condition}}$ and $\beta_3^{\text{condition}}$ are the effects of the second and the third spike-in conditions relative to the reference condition after correction for lab-effects. They can be directly interpreted in terms of $\log_2$ fold changes between their corresponding spike-in condition and the condition that was chosen as a reference condition. Given a certain spike-in condition, β_2^{lab} and β_3^{lab} denote the effects on the $\log_2$ protein intensity of the second and the third labs, respectively, relative to the reference lab. To model the discrete, non-linear effects of spike-in condition and lab, we make use of so-called "dummy" variables whereby e.g. $x_{r2}^{\text{condition}}$ is equal to 1 if run r corresponds to the second spike-in condition and 0 otherwise. Idem for the other dummies. $\boldsymbol{\varepsilon}$, finally, is a vector that contains the random error terms ε_r. For the CPTAC dataset, the regression model can then be written as follows:

$$y_r = \beta^0 + \sum_{t=2}^{3} x_{rt}^{\text{condition}} \beta_t^{\text{condition}} + \sum_{b=2}^{3} x_{rb}^{\text{lab}} \beta_b^{\text{lab}} + \varepsilon_r \qquad \text{(Eq. 4.16)}$$

Regression models that contain only categorical variables[32] are often presented in the more condense ANOVA notation whereby the predictor variables $\boldsymbol{x_r}$ are not written explicitly, but indices are used instead to denote different levels of the categorical variable. For the CPTAC dataset, the ANOVA notation of the model can be written as follows:

$$y_{tbr} = \beta^0 + \beta_t^{\text{condition}} + \beta_b^{\text{lab}} + \varepsilon_r \quad \text{(Eq. 4.17)}$$

Here, y_{tbr} is the $\log_2$-transformed protein-level intensity for protein i in MS run r, which corresponds to condition (treatment) $t = 2{,}3$ and lab (block) $b = 2{,}3$. β^0 is the constant intercept. $\beta_t^{\text{condition}}$ is the effect of condition t relative to the reference condition, β_b^{lab} the effect of lab b relative to the reference lab and ε_r the random error term. The estimated condition effects $\hat{\beta}_t^{\text{condition}}$ are the effects of interest.

The most common way to estimate the parameters is by least squares, i.e. by minimizing the sum of the squared distances of the observations to the model fit:

$$\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 = (\boldsymbol{y} - \boldsymbol{X\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X\beta}) \quad \text{(Eq. 4.18)}$$

This results in the following estimator for $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathbf{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathbf{T}}\boldsymbol{y} \quad \text{(Eq. 4.19)}$$

In the least-squares context, the estimator for the variance of $\widehat{\boldsymbol{\beta}}$ is given by:

$$\widehat{\mathrm{Var}}\left(\widehat{\boldsymbol{\beta}}\right) = \hat{\sigma}^2\left(\boldsymbol{X}^{\mathbf{T}}\boldsymbol{X}\right)^{-1} \quad \text{(Eq. 4.20)}$$

Herein, $\hat{\sigma}^2$ is an unbiased estimator for the error variance:

$$\hat{\sigma}^2 = \frac{\left\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right\|^2}{R - M}, \quad \text{(Eq. 4.21)}$$

with R the number of runs and M the number of estimated parameters in the mean model. The estimator for the standard error on the mth parameter estimator $\hat{\beta}_m$ can also be written as:

$$\hat{\sigma}_{\hat{\beta}_m} = \hat{\sigma}\sqrt{v_m}, \quad \text{(Eq. 4.22)}$$

with $v_m = \left(\boldsymbol{X}^{\mathbf{T}}\boldsymbol{X}\right)^{-1}_{m,m}$ the mth diagonal element of $\left(\boldsymbol{X}^{\mathbf{T}}\boldsymbol{X}\right)^{-1}$. Under the null hypothesis of no differential abundance, the following test statistics follows a t-distribution with $R - M$ degrees of freedom:

$$\frac{\hat{\beta}_m - a}{\hat{\sigma}_{\hat{\beta}_m}} \sim t_{R-M} \quad \text{(Eq. 4.23)}$$

Here, $\hat{\beta}_m$ is the estimated value for the mth model parameter β_m (typically the effect of a certain treatment; in the case of the CPTAC study: the spike-in condition), a the value of β_m under the null hypothesis (typically zero) and $\hat{\sigma}_{\hat{\beta}_m}$ the estimated variance on the estimate of β_m. Given the null distribution, a p-value can be calculated for each parameter, denoting its statistical significance. Note that some research questions (e.g. the difference in protein abundance between two non-reference conditions) require statistical inference on a linear combination of

[32] Categorical variables are variables that do not correspond to any measurable quantities. Examples include gender, different compounds, different treatments, etc. This is opposed to numerical variables that correspond to measurable quantities (e.g. doses of a certain compound, time after treatment, blood pressure,).

multiple model parameters, so-called statistical contrasts. Just as for single parameters, t-statistics and p-values can also be calculated for contrasts.

The current version (3.12.2) of the popular Bioconductor package MSstats makes use of linear regression at the protein level. These protein-level summaries are obtained after imputation at the PSM-level data under a MNAR assumption followed by a median polish summarization [446].

The popular Bioconductor package limma was originally developed for the analysis of microarray data [468], but has also become popular for differential proteomics analyses [469]. Limma makes use of the huge amounts of data in high-throughput omics datasets to borrow strength across proteins. More specifically, limma assumes that the residual variances from each regression model are composed of a common variance shared by all proteins (models) and a protein-specific variance. This allows to obtain a more stable estimate of the error variance, which is especially beneficial for proteins identified by only a few peptides. Indeed, their variances are stabilized by relying on the variances estimated for proteins with much more data.

Limma uses an empirical Bayes framework to provide a statistically sound alternative to SAM for linear regression models. More specifically, limma proposes a Bayesian model which assumes the following prior distribution on the error variance σ_i^2 for each protein i ($i = 1, \dots, I$):

$$\frac{1}{\sigma_i^2} \sim \frac{1}{d_0 \sigma_0^2} \chi^2_{d_0} \qquad \text{(Eq. 4.24)}$$

In this formula, σ_0^2 is a prior variance and $\chi^2_{d_0}$ denotes a χ^2 distribution with d_0 degrees of freedom. In limma, the user does not define the value of the prior variance σ_0^2 and the prior degrees of freedom d_0, but estimates σ_0^2 and d_0 based on all protein error variance estimates $\hat{\sigma}_i^2$ and the degrees of freedom $d_i = M - R$ of all proteins in the dataset. Statistical inference methods whereby the priors are estimated based on the data are termed empirical Bayesian methods.

Limma's empirical Bayes estimators for the prior variance σ_0^2 and the prior degrees of freedom d_0 have closed-form solutions that are computationally very fast. Furthermore, instead of estimating full posterior distributions, limma calculates a maximum a posteriori point estimate $\tilde{s}_i$ for the residual standard deviations:

$$\tilde{s}_i = \sqrt{\frac{d_i \hat{\sigma}_i^2 + d_0 \hat{\sigma}_0^2}{d_i + d_0}} \qquad \text{(Eq. 4.25)}$$

Herein, $\hat{\sigma}_i$ is the residual standard error for protein i and $\hat{\sigma}_0^2$ the estimated common variance over all proteins. Substituting the residual standard deviation by its maximum a posteriori estimator results in a moderated t-test statistic:

$$\tilde{t}_{im} = \frac{\hat{\beta}_{im} - a}{\tilde{s}_i \sqrt{v_{im}}} \qquad \text{(Eq. 4.26)}$$

It can be shown that the moderated t-test statistic $\tilde{t}_{im}$ follows a t-distribution with $d_i + d_0$ degrees of freedom, with d_i equal to $R - M$, as indicated before. Hence, not only are the variances being stabilized as in SAM, but, contrary to SAM, the null distribution follows an analytical t-distribution. Note that the degrees of freedom of this t-distribution are augmented with d_0 as compared to the degrees of freedom of the null distribution of the ordinary t-test.

This reflects the increased power of the moderated t-test due to the borrowing of strength across proteins.

4.2.3. Peptide-based models

Summarization-based approaches, especially the naive ones such as those based on mean and median summarization, use summary values that are based on different peptides and different numbers of peptides. When such summaries are compared to each other, a bias will be introduced due to the comparison of different peptides with non-negligible differences in ionization efficiency. Also, differences in precision due to the different numbers of peptides for each summary are ignored. Instead of summarizing PSMs directly to the protein level, it is however also possible to keep the data at the PSM or peptide-level and to include the hierarchical nature of the data directly into the statistical model. We call these models peptide-based models (see chapter 8). They have the advantage that they correct for differences in ionization efficiencies between different peptides and for differences in precision due to different numbers of identified peptides in each MS run. Compared to summarization-based methods, two additional parameters need to be added to the model. A peptide-based linear regression model for each protein i in the CPTAC experiment then looks as follows (whereby the indicator i is again suppressed for notational convenience):

$$y_{pr} = \beta^0 + \beta_t^{\text{condition}} + \beta_b^{\text{lab}} + \beta_p^{\text{peptide}} + u_r^{\text{run}} + \varepsilon_{pr} \qquad \text{(Eq. 4.27)}$$

The response variable y_{pr} is now the $\log_2$-transformed intensity of ***peptide*** p in run r. β_p^{peptide} is added to account for the effect of the p^{th} peptide. The effect for MS run, u_r^{run}, is added because for protein i, there can be multiple peptides identified in the same run. This is an important point: the run effect needs to be included because peptide intensities within the same run are expected to be positively correlated. Indeed, due to the run-specific effects described in 3.4, peptide intensities from a protein within the same run will behave more similar compared to peptide intensities that were measured across different runs. However, if the run effect were modeled as a standard fixed effect, statistical inference would only be valid for within-run comparisons because the run-to-run variability would be removed from the model. When the run effect is modeled as a random effect, whereby it is assumed that $u_r^{\text{run}} \sim \text{N}(0, \sigma_u^2)$, both within- and between-run variability are taken into account. This is important in label-free proteomics experiments because the treatment will vary between runs, but not within runs. The run effect thus both accounts for the correlation of all peptides of protein i identified within run r and enables a correct statistical inference for between-run comparisons. Statistical models that contain both fixed and random effects are referred to as mixed models and allow to model correlation structures in the data. This mixed model structure for peptide-level data was first proposed by Daly ***et al.*** (2008) [470].

Clough ***et al.*** (2009) [471] propose a specific parameterization for protein-wise mixed models:

$$y_{fr} = \beta^0 + \beta_f^{\text{feature}} + \beta_t^{\text{condition}} + \beta_{ft}^{\text{feature:condition}} + \beta_b^{\text{biorep}} + \varepsilon_{fr} \qquad \text{(Eq. 4.28)}$$

These authors later implemented this model in the proteomics quantification package MSstats prior to version 3, in which MSstats used to model the data directly at the feature (PSM) level [409, 472].

Herein, β_f^{feature} is the effect of the fth feature (PSM). $\beta_{ft}^{\text{feature:condition}}$ is an interaction effect between feature and condition. Such an interaction allows the effect of interest (condition) to affect each feature differently. β_b^{biorep} is then the effect of the bth biological repeat. In MSstats, the experimenter can opt to encode β_b^{biorep} either as a fixed effect, which is useful when β_b^{biorep}

is a blocking factor, or as a random effect, which is useful in the case of biological replication. Indeed, biological replication caused by e.g. multiple measurements on the same animals also leads to correlation in the data and should therefore be modeled as random. The disadvantage of the former MSstats framework is that it does not allow to correctly model the within-sample correlation unless the samples coincide with the biological repeats. Moreover, just like the present MSstats implementation, it only allows to model experiments that fit into this specific model framework.

It has to be noted that, if the summarization step is performed correctly, the estimated differences in ionization efficiencies can in fact be removed from the data. Hence, the only clear advantage of peptide-based models is that they account for difference in precision due to different numbers in peptides. However, this advantage seems to be rather small in practice, which may be one of the reasons why MSstats in their most recent version, reverted to a faster, summarization-based workflow.

The very first peptide-based model for label-free shotgun proteomics was proposed in 2008, when Bukhman *et al.* proposed the following model [473]:

$$y_{pr} = \gamma_p + \psi_p \sum_i \delta_{ip}\theta_{ir} + \varepsilon_{pr} \qquad \text{(Eq. 4.29)}$$

Herein, y_{pr} is the log-transformed intensity of peptide p in sample (or MS run, assuming each sample was only run once) r, γ_p the background log-transformed intensity of peptide p, ψ_p the peptide-specific effect of peptide p, δ_{ip} an indicator whether peptide p maps to protein i (1 if the peptide maps and 0 if the peptide does not map), θ_{ir} the abundance of protein i in sample r and ε_{pr} a random error term. The inclusion of the ψ_p term allows the sample effect θ_{ir} to be different from peptide to peptide. Note that in this model, all proteins are modeled together and that this model allows to use shared peptides, although the authors exclude peptides that are shared by three or more proteins.

Another peptide-based model was proposed by Henao *et al.* (2012) [474]:

$$y_{psb} = m_{pb} + \sum_l a_{lp}z_{ls} + \sum_i b_{ip}w_{is} + \varepsilon_{ps} \qquad \text{(Eq. 4.30)}$$

Herein, y_{psb} is the average log-transformed intensity of peptide p in sample s in batch b. m_{pb} is the average intensity of peptide p in batch b. This model also accounts for the correlation of certain peptides within the same sample (e.g. peptides that behave similarly due to physicochemical similarities). To capture this correlation, z_{ls} represents the lth intra-sample effect for sample s, while a_{lp} are the peptide-specific effects that correspond to each of these intra-sample effects. w_{is} represents the effect of interest: the effect of protein i in sample s, while b_{ip} models the impact of peptide p on the effect of protein i. ε_{ps} is a random error term. From the model specification, it is clear that this model will be strongly over-parameterized for most, if not all proteins. These authors, however, tried to tackle the quantification problem from a Bayesian perspective. In classic frequentist statistics, the aim is to estimate the true value of one or more unknown population parameters and provide estimates on the uncertainty of these parameter estimates. Contrary, in Bayesian statistics, the population parameters are *believed* upfront to follow certain distributions, the prior distributions. When experiments are performed, evidence (data) is collected that might confirm or challenge this prior belief. The prior distributions are then updated based on the data by making use of Bayes' theorem and result in posterior distributions that reflect the statistician's new beliefs after confronting his beliefs with the data.

The idea of including a researcher's beliefs into a statistical method is very sensible because experiments are rarely, if ever, performed without any prior knowledge. Indeed, even if nothing is known about a protein in the literature, a protein's true $\log_2$ fold change will either be 0 (unregulated), or a positive or negative value rather close to zero (up- or downregulated). Very extreme $\log_2$ fold changes such as +1000 or -1000 are highly unlikely and can therefore be given a very low prior probability. And, even if absolutely nothing is known or can be assumed about an experiment and all values are equally likely, a so-called uninformative prior can be used. In this respect, the framework of Bayesian statistics is very elegant because it allows the posterior distribution of a previous experiment to be used as a prior distribution in a follow-up experiment and thus to organically update our beliefs based on the data.

Henao *et al.* (2012) indeed place Gaussian (normal) priors on the average abundance μ_{im} and the noise component ε_{in} [474]. Gaussian priors are also assigned to a_{il}, while a Laplace prior is set on z_{ln} to allow these intra-sample effects to shrink to 0 if necessary. b_{ik} is also given a normal prior, but hyperpriors are set in such a way that proteins can also be correlated with each other in a hierarchical tree structure. Bayesian models have also been proposed to include the effects of shared peptides [475].

Disadvantages of the Bayesian framework are that the choice of the prior is always somewhat arbitrary and based on the beliefs of the researcher. Moreover, Bayesian inference models mostly do not have a closed-form expression for the posterior distributions. Therefore, the posterior distributions need to be approximated by repeated sampling, e.g. by making use of Markov Chain Monte Carlo (MCMC) methods, which are computationally very intensive.

Another persistent issue in the proteomics field are missing values, hence the usual custom of including an imputation step in a typical workflow. Instead of imputing missing values, it is also possible to handle missing values within the framework of the statistical model. For shotgun proteomics data, this approach was pioneered by Karpievitch *et al.* (2009) [425]. In their censored regression model, peptide intensities are assumed to be either missing completely at random, or missing not at random if a peptide's intensity falls below a certain censoring threshold c_{ip} for each peptide p corresponding to protein i. These authors model all proteins together in one model. It is assumed that all $\log_2$-transformed intensities originate from a normal distribution with mean μ_{ipt} and standard deviation σ_{ip}. The expected intensity for a peptide p of a protein i in treatment condition t can then be described as follows:

$$\mu_{ipt} = \beta^0 + \beta_i^{\text{protein}} + \beta_{ip}^{\text{peptide}} + \beta_{it}^{\text{condition}} \qquad \text{(Eq. 4.31)}$$

In each MS run r, the probability of a peak to be missing at random is assumed to be equal to π_r. If W_{iptr} is the probability that a $\log_2$-transformed intensity y_{iptr} is observed (0 if observed and 1 if unobserved), the probability that intensity y_{iptr} will be missing can be written as follows:

$$P(W_{iptr} = 1) = \pi_r + (1 - \pi_r)\Phi\left(\frac{c_{ip} - \mu_{ipt}}{\sigma_{ip}}\right) \qquad \text{(Eq. 4.32)}$$

In this expression, Φ is the cumulative distribution of the normal distribution with mean equal to 0 and standard deviation equal to 1. This expression shows that peptides are missing at random (MAR, conditionally on run r) with a probability π_r. Any peptide that is not MAR will be MNAR if its $\log_2$-transformed intensity is lower than the peptide-specific censoring threshold c_{ip}. The current implementation of MSstats uses a very similar model to impute missing values prior to summarization to the protein level, albeit without the random missingness component [446]. In 2012, Koopmans *et al.* proposed an empirical Bayesian random censoring threshold model to cope with missing values in a summarization-based context, but just like most

Bayesian inference models, the posterior does not have a closed-form solution and needs to be constructed by repeated MCMC sampling, which makes it computationally intensive [476].

4.2.4. Ridge regression

Due to low protein abundances, limited numbers of tryptic peptides per protein and the data-dependent nature of the acquisition, the number of observed peptides is relatively small for most of the proteins in a typical label-free shotgun proteomics dataset. This makes protein-wise statistical modeling challenging because even relatively simple models are prone to over-fitting for such proteins. Over-fitting occurs when a model is too complex with respect to the amount of data that is available: the model is fit too closely to the observed data but will not generalize towards new data.

Ridge regression is a way to reduce over-fitting. Recall that for ordinary least squares, the following loss function is minimized:

$$\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 = (\boldsymbol{y} - \boldsymbol{X\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X\beta}) \qquad \text{(Eq. 4.33)}$$

Ridge regression adds a penalty term to this loss function (indicated in red):

$$(\boldsymbol{y} - \boldsymbol{X\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{D\beta}, \qquad \text{(Eq. 4.34)}$$

with

$$D = \begin{bmatrix} 0 & \mathbf{0}_{1\times(M-1)} \\ \mathbf{0}_{(M-1)\times 1} & \boldsymbol{I}_{(M-1)\times(M-1)} \end{bmatrix} \qquad \text{(Eq. 4.35)}$$

Herein, $\mathbf{0}$ is a matrix with only zeros and $\boldsymbol{I}$ is the unit matrix[33]. Matrix D allows certain parameters to be unpenalized by setting their corresponding diagonal elements to 0. In the given example, only the intercept β^0 remains unpenalized. The penalty term increases if the absolute values of the parameters $\boldsymbol{\beta}$ increase. This prevents over-fitting by shrinking model parameters towards 0 (Fig. 4.14).

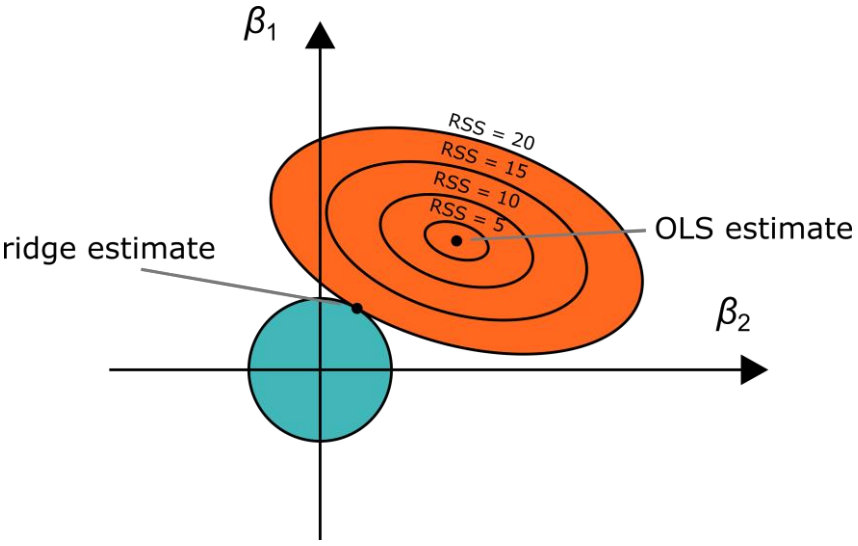


Figure 4.14. Graphical representation of the ridge estimate and the ordinary least squares (OLS) estimate in an example case where two model parameters β_1 and β_2 need to be estimated. The OLS

[33] The unit matrix is a matrix with 1 on its diagonal elements and 0 on its off-diagonal elements. The model thus implies equal variances for all covariates and no correlation between the covariates.

estimate minimizes the residual sum of squares (RSS), while the ridge estimates are shrunken towards 0.

When fitting this regression model, the aim is to minimize the mean squared error (MSE):

$$\mathrm{MSE}(\widehat{\boldsymbol{\beta}}) \stackrel{\text{def}}{=} \mathrm{E}\left[\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^2\right] \qquad \text{(Eq. 4.36)}$$

It can be shown that the MSE can also be written as [477]:

$$\mathrm{MSE}(\widehat{\boldsymbol{\beta}}) = \mathrm{E}\left[\left(\widehat{\boldsymbol{\beta}} - \mathrm{E}[\widehat{\boldsymbol{\beta}}]\right)^2\right] + \mathrm{E}\left[\left(\mathrm{E}[\widehat{\boldsymbol{\beta}}] - \boldsymbol{\beta}\right)\right]^2 \qquad \text{(Eq. 4.37)}$$

$$= \mathrm{Var}(\widehat{\boldsymbol{\beta}}) + \mathrm{Bias}(\widehat{\boldsymbol{\beta}})^2 \qquad \text{(Eq. 4.38)}$$

In 1956, Stein showed that for models with 3 or more parameters, certain shrinkage estimators outperform the least-squares estimator in terms of MSE [478]. Shrinkage estimators introduce a small bias but reduce the overall MSE due to a strong reduction in the variance of the estimator. Therefore, such shrinkage estimators are more stable overall. Leave-one-out cross-validation is one way to tune the penalty parameter. Cross-validation enables to assess how accurate a model is on new, unobserved data. In leave-one-out cross-validation, the data is fitted to the dataset from which one observation j is removed. This allows to estimate the MSE: the model's estimate $\hat{y}_j$ for observation j can be seen as a "new" data point. With cross-validation, the following estimator for the overall mean squared error is minimized towards λ [479]:

$$f(\lambda) = \frac{1}{J}\sum_{j=1}^{J}\left(\boldsymbol{x_j}\widehat{\boldsymbol{\beta}}^{(j)}(\lambda) - y_j\right)^2 \qquad \text{(Eq. 4.39)}$$

Herein $\widehat{\boldsymbol{\beta}}^{(j)}(\lambda)$ is a vector of ridge parameter estimates based on the data from which the jth observation is removed and $\left[\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(j)}(\lambda)\right]_j$ the leave-one-out model estimate for the jth observation y_j. Leave-one-out cross validation requires iteratively fitting ridge models without the jth observation until convergence. It is also possible to repeatedly leave out K observations and reduce the squared distances of the leave-K-out model fit to the K observations that were left out.

As explained in 4.2.3, peptide-based models require the inclusion of a random sample effect to allow correct statistical inference. Leave-one-out cross-validation would require iterative fitting of a mixed model. There is however a link between mixed models and ridge regression, which is tempting to exploit when introducing ridge regression in a mixed model context as this would give a big computational advantage. As a demonstration of this link, assume the following mixed model:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon} \qquad \text{(Eq. 4.40)}$$

With $\boldsymbol{X}$ the design matrix for the fixed effects $\boldsymbol{\beta} = \begin{bmatrix}\beta^0 \\ \beta_1 \\ \dots \\ \beta_m \\ \dots \\ \beta_M\end{bmatrix}$, $\boldsymbol{Z}$ the design matrix for the random effects $\boldsymbol{u} = \begin{bmatrix}u_1 \\ \dots \\ u_n \\ \dots \\ u_N\end{bmatrix}$ with $\boldsymbol{u} \sim \mathrm{MVN}(\boldsymbol{0}, \sigma_u^2\boldsymbol{I})$. Herein, MVN denotes the multivariate normal

distribution, $\mathbf{0}$ a $1 \times N$ column vector with zeros, σ_u^2 the variance on the random effects and $\boldsymbol{I}$ the unit matrix. $\boldsymbol{\varepsilon} \sim \mathrm{MVN}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ denotes the random error terms. We maximize the joint likelihood of y, $\boldsymbol{\beta}$ and $\boldsymbol{u}$ towards $\boldsymbol{\beta}$ and $\boldsymbol{u}$:

$$\mathrm{L}(y, \boldsymbol{\beta}, \boldsymbol{u}) = \mathrm{L}(y, \boldsymbol{\beta} | \boldsymbol{u}) \mathrm{L}(\boldsymbol{u}) \quad \text{(Eq. 4.41)}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu})/2\sigma^2} \frac{1}{\sqrt{2\pi\sigma_u^2}} e^{-(-\boldsymbol{u})^{\mathrm{T}}(-\boldsymbol{u})/2\sigma_u^2} \quad \text{(Eq. 4.42)}$$

Log-transformation results in:

$$\mathrm{l}(y, \boldsymbol{\beta}, \boldsymbol{u}) = -\frac{J}{2}\log(2\pi) - \frac{J}{2}\log(\sigma^2) - \frac{(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu})}{2\sigma^2}$$
$$-\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma_u^2) - \frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}}{2\sigma_u^2} \quad \text{(Eq. 4.43)}$$

With N the number or random effect parameters. After replacing σ_u^2 by σ^2/λ and multiplying by -2, this is equivalent with minimizing the following expression:

$$J\log(2\pi) + J\log(\sigma^2) + \frac{(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu})}{\sigma^2} + N\log(2\pi) + N\log\left(\frac{\sigma^2}{\lambda}\right) + \frac{\lambda\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}}{\sigma^2} \quad \text{(Eq. 4.44)}$$

Set $\boldsymbol{\theta} = \begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{u}\end{bmatrix}$ and $\boldsymbol{C} = [\boldsymbol{X} \quad \boldsymbol{Z}]$. Minimization to $\boldsymbol{\theta}$ only involves:

$$(\boldsymbol{y}-\boldsymbol{C\theta})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{C\theta}) + \lambda\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{D\theta}, \quad \text{(Eq. 4.45)}$$

which is exactly the ridge regression loss function. Hence, parameters with a ridge penalty can be estimated by parameterizing them as random effects in a mixed model. In peptide-based models, where a random run effect needs to be included to account for within-run correlation, this link between mixed models and ridge regression can be exploited to estimate parameters with a ridge penalty, as we will see in Chapter 9.1. After optimizing the loss function, we obtain the following estimator:

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix}\hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}}\end{bmatrix} = \left(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C} + \hat{\lambda}\boldsymbol{D}\right)^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{y} \quad \text{(Eq. 4.46)}$$

This estimator for $\hat{\boldsymbol{u}}$ is termed the best linear unbiased predictor (BLUP). An estimate for σ^2 can be obtained by plugging in the estimates $\hat{\boldsymbol{\theta}}$ into the log likelihood and solving this profile log-likelihood towards σ^2. In practice, we make use of the restricted maximum likelihood (REML) criterion that also accounts for the degrees of freedom due to the fixed effects in the model. Conditional on $\boldsymbol{u}$, the variance on $\hat{\boldsymbol{\theta}}$ can be estimated as follows:

$$\widehat{\mathrm{Var}}\left(\begin{bmatrix}\hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} - \boldsymbol{u}\end{bmatrix} | \boldsymbol{u}\right) = \hat{\sigma}^2\left(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C} + \hat{\lambda}\boldsymbol{D}\right)^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}\left(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C} + \hat{\lambda}\boldsymbol{D}\right)^{-1} \quad \text{(Eq. 4.47)}$$

This variance estimator, however, does not account for the bias in the estimator $\hat{\boldsymbol{u}}$. Hence, inference based on this estimator will only be correct when the bias is negligible. However, since $\mathrm{E}(\boldsymbol{u}) = 0$, the BLUP estimator $\hat{\boldsymbol{u}}$ is unbiased on average over the distribution of u. Unconditional on $\boldsymbol{u}$, the variance estimator on $\hat{\boldsymbol{\theta}}$ is given by:

$$\widehat{\mathrm{Var}}\left(\begin{bmatrix}\hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} - \boldsymbol{u}\end{bmatrix}\right) = \hat{\sigma}^2\left(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C} + \hat{\lambda}\boldsymbol{D}\right)^{-1} \quad \text{(Eq. 4.48)}$$

The estimator also accounts for the bias introduced by the penalized regression and is therefore somewhat larger than the conditional variance estimator. Ridge regression has been used in other omics fields to predict the effects of various molecular markers on organismal phenotypes [480, 481].

4.2.5. Robust regression with M estimation

Outliers are observations with extreme response values. If such observations also have a strong leverage (i.e. if they have a large distance to the average predictor values), the observation will have a strong influence on the model fit: the model will be fit close to the influential observation due to a lack of neighboring observations. Outliers are more likely to correspond to less reliable measurements. Indeed, an extremely high intensity value could for example have originated from a spike in electrospray voltage, while a very low intensity value is very likely to be missing in a repeated run due to intensity-dependent missingness. Moreover, even if the outlier is a valid measurement, it is often undesirable that a single observation has a very strong impact on the model.

Robust regression with M estimation aims to minimize the maximal bias of the estimators. With robust regression, statistical tests are only asymptotically valid. However, if the errors are normally distributed, M estimators have a high efficiency. Recall the OLS loss function:

$$\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 = (\boldsymbol{y} - \boldsymbol{X\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X\beta}) \qquad (\mathrm{Eq.}\ 4.49)$$

This function is also called the L2 loss function since it minimizes the L2 norm of $\boldsymbol{y} - \boldsymbol{X\beta}$. With M estimation, the following loss function is minimized:

$$\Omega(\boldsymbol{y} - \boldsymbol{X\beta}) \qquad (\mathrm{Eq.}\ 4.50)$$

The function $\Omega(x)$ should have the following characteristics:

- $\Omega(x)$ is symmetric
- $\Omega(x)$ has a minimum at $\Omega(0) = 0$
- $\Omega(x)$ is positive for all $x \neq 0$
- $\Omega(x)$ increases as x increases

Given these conditions, the estimator $\widehat{\boldsymbol{\beta}}$ is the solution to the equation:

$$\omega(\boldsymbol{y} - \boldsymbol{X\beta}) = 0 \qquad (\mathrm{Eq.}\ 4.51)$$

Where ω is the derivative of Ω. For $\widehat{\boldsymbol{\beta}}$ to possess the robustness property, ω should be bounded (i.e. there exists a real number M such that $|\omega(x)| \leq M$ for all values of x). However, robust ω functions are non-linear in $\boldsymbol{\beta}$ and typically do not have a closed-form solution. We will therefore recast the problem.

When location parameters $\boldsymbol{\beta}$ and a scale parameter σ have to be estimated simultaneously, we minimize:

$$\Omega\left(\frac{\boldsymbol{y} - \boldsymbol{X\beta}}{\sigma}\right) \qquad (\mathrm{Eq.}\ 4.52)$$

Whereby:

$$\omega\left(\frac{\boldsymbol{y} - \boldsymbol{X\beta}}{\sigma}\right) = 0 \qquad (\mathrm{Eq.}\ 4.53)$$

Define $\boldsymbol{\eta} = \frac{\boldsymbol{y}-\boldsymbol{X\beta}}{\sigma}$ and weight function $w(\boldsymbol{\eta}) = \omega(\boldsymbol{\eta})/\boldsymbol{\eta}$. The last estimation equation can then be rewritten as:

$$w(\boldsymbol{\eta})\boldsymbol{\eta} = 0 \qquad \text{(Eq. 4.54)}$$

This expression can be solved as an iteratively reweighted least-squares (IRWLS) problem. Herein, the weights $w(\boldsymbol{\eta})$ are kept constant in $\boldsymbol{\eta}$ and the expression is solved to $\boldsymbol{\beta}$. Then, the weights are recalculated based on the new $\widehat{\boldsymbol{\beta}}$ and the procedure is repeated until convergence.

Examples of robust loss functions Ω include:

- Huber: $\Omega^{\text{Huber}} = \begin{cases} \mathrm{x}^2/2 & \text{if } |x| \leq k \\ k(|x| - k/2) & \text{if } |x| > k' \end{cases}$ (Eq. 4.55)
  with $k = 1.345$ the default tuning constant
- "Fair": $\Omega^{\text{Fair}} = c^2\left(\frac{|x|}{c} - \log\left(1 + \frac{|x|}{c}\right)\right)$, (Eq. 4.56)
  with $c = 1.3998$ the default tuning constant
- Cauchy: $\Omega^{\text{Cauchy}} = \frac{c^2}{2}(1 + (x/c)^2)$, (Eq. 4.57)
  with $c = 1.3849$ the default tuning constant

The default tuning constants are chosen in such a way that the methods have a 95% asymptotic efficiency when applied to standard normal data. Fig. 4.15 demonstrates the impact of robust regression with M estimation with Huber weights.

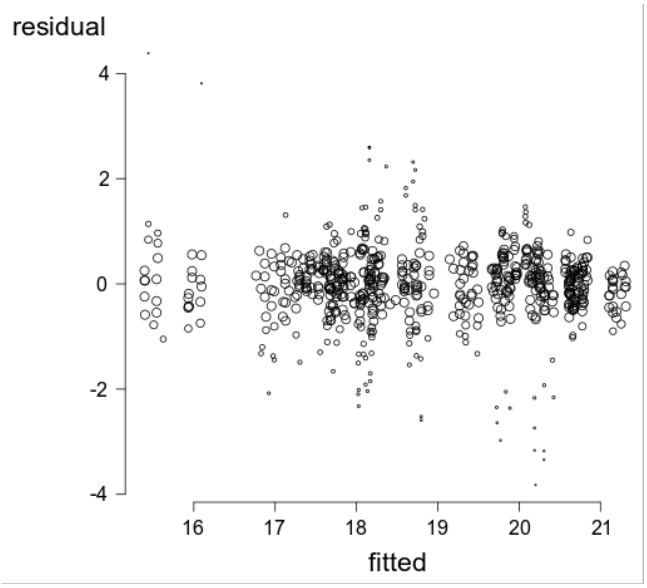


Figure 4.15. Plot of the residuals $y_j - \hat{y}_j$ in function of the model-fitted values for all $\log_2$-tranformed peptide intensities $\hat{y}_j$ $(j = 1, \ldots, 637)$ of yeast protein SYKC in the CPTAC dataset after fitting regression model $(\text{Eq. } 4.27)$ robustly with Huber weights. The sizes of the datapoints are proportional to the Huber weights in the IRWLS procedure. Note that observations with large residuals have small weights, which is indicative of the robustness property.

Examples of common loss functions and their corresponding weight functions are plotted in Fig. 4.16 and 4.17, respectively. An extensive overview of robust loss functions can be found in Bolstad (2004) [482].

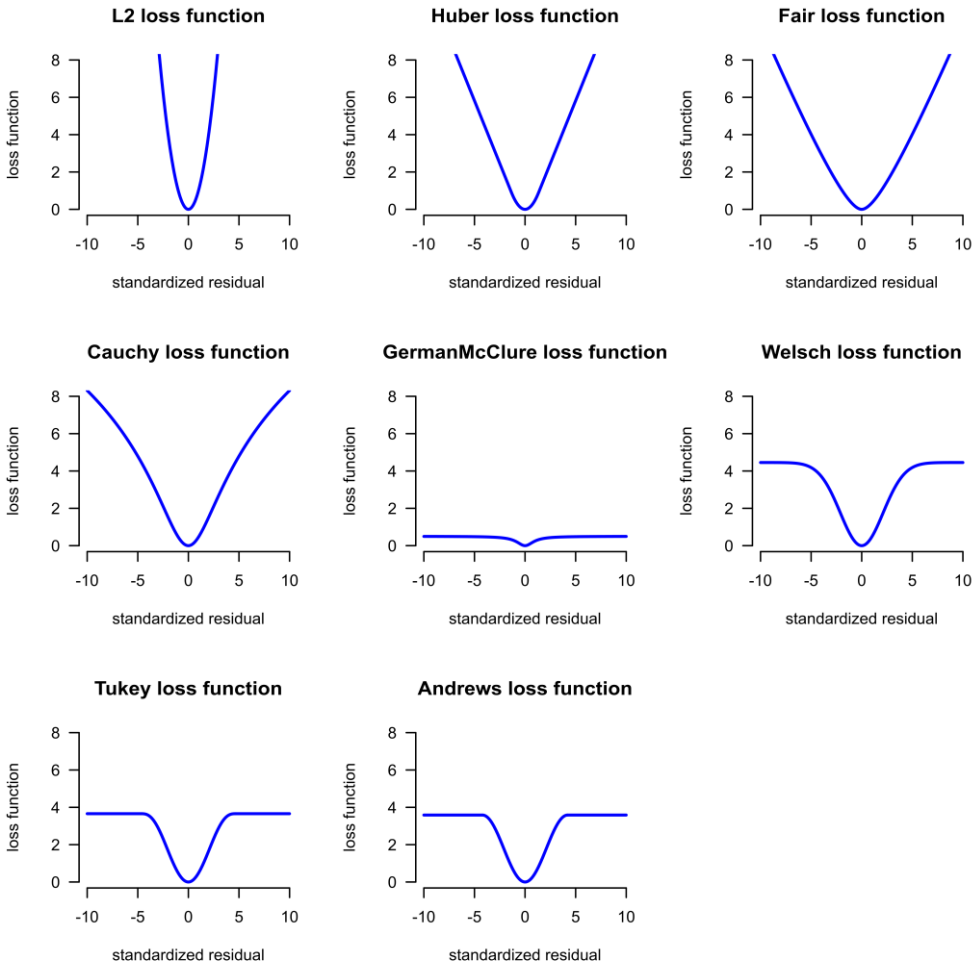


Figure 4.16. The default L2 loss function and examples of loss functions Ω that are commonly used in robust M estimation. Figure adapted from Bolstad (2004) [482].

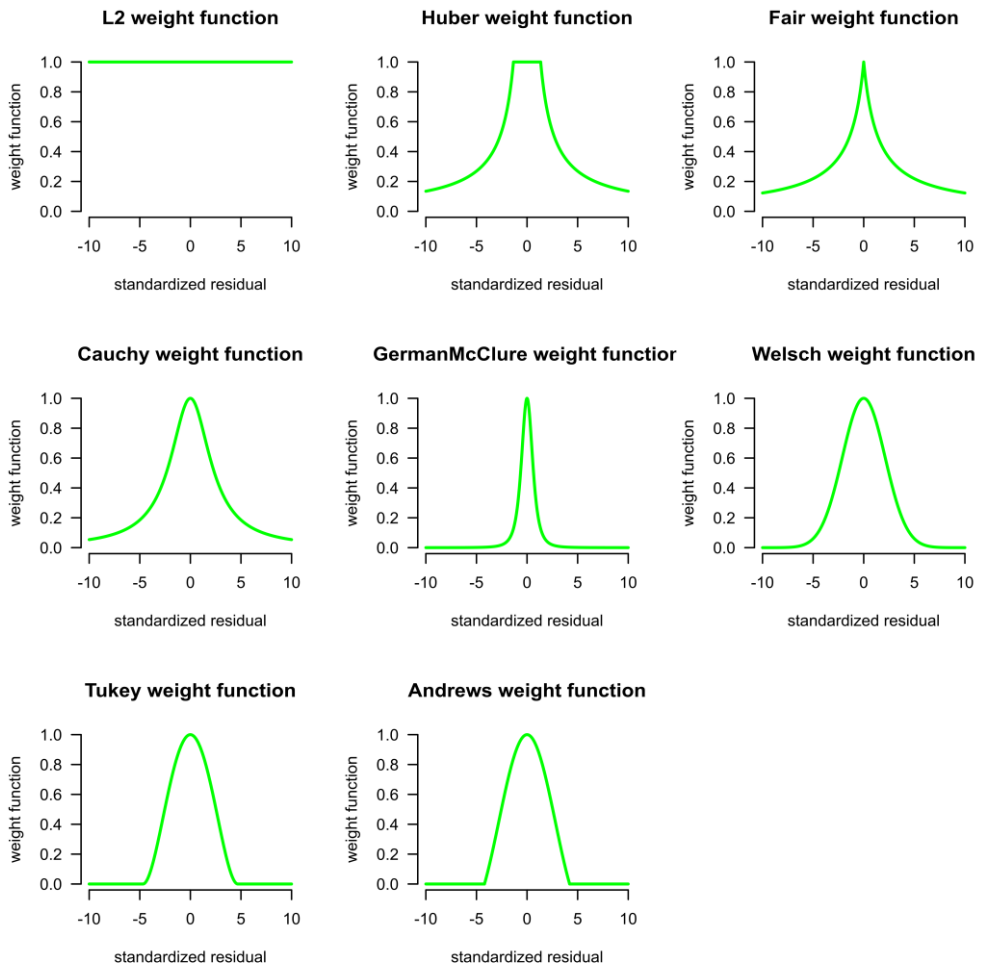


Figure 4.17. Examples of the weight functions w that are used in the IRWL procedure to obtain the corresponding loss functions in Fig. 4.15. Figure adapted from Bolstad (2004) [482].

4.2.6. Counting-based methods

To quantify proteins with any of the above-described methods, it is necessary to extract ion intensities, which generally requires, as shown in section 3.3, rather advanced algorithms that are implemented in specialized software. Moreover, reliable protein quantification with these methods requires, depending on the study design, basic to advanced statistical knowledge. It was soon noticed that simply counting the number of MS² spectra that map to a certain protein provides a reasonably good approximation of a protein's abundance [483, 484]. This makes sense because the more abundant a protein, the more of its peptide ions that can be expected to be detectable above noise levels. Moreover, the more abundant a peptide, the longer its elution time and hence the higher the chance that it will be targeted for fragmentation more than once, thus generating more MS² spectra. Spectral counting also deals more naturally with missing values: as a zero count [90]. Spectral counting became very appealing to many researchers, not because of its reliability, but mainly because of its ease-of-use. Indeed, researchers could now simply count the number of MS² spectra mapping to a protein and directly divide these numbers in order to obtain a fold change estimate. However, when the significance of such fold changes needs to be assessed, statistics are again needed. A natural framework for handling count data is Poisson regression. Poisson regression assumes that the spectral counts x_{ir} for each protein i in each run r follow a Poisson distribution:

$$x_{ir} \sim \text{Poisson}(\mu_{itb}) \quad \text{(Eq. 4.58)}$$

This makes sense because a Poisson distribution is typically used to model the number of times an event occurs (detecting a spectrum that maps to protein i) during a fixed time interval (an MS run). For the CPTAC example, we can make use of the generalized linear model framework, to model the first two moments (mean and variance) of the spectral counts. As counts always have a lower bound of 0 and negative means are not meaningful for count data, we make use of a log-link function to allow unbounded estimation of the model parameters. For the CPTAC dataset, the model can be specified as follows:

$$\log(\mu_{itb}) = \beta_i^0 + \beta_{it}^{\text{condition}} + \beta_{ib}^{\text{lab}} \quad \text{(Eq. 4.59)}$$

Note that for a Poisson distribution the mean μ_{itb} and the variance v_{itb} are equal.

Here, β_i^0 is the intercept, $\beta_{it}^{\text{condition}}$ is the effect of spike-in condition t and β_{ib}^{lab} is the effect of lab b. By using a Poisson distribution, it is implied that the variance in the data v_{itb} is equal to the mean μ_{itb}. Such a mean-variance relationship is very restrictive. In reality, the residual variance is often larger (over-dispersion) or sometimes even smaller (under-dispersion) than what would be expected under the Poisson distribution. The mean-variance relationship can however be relaxed by making use of a quasi-Poisson regression model [485].

Note that quasi-Poisson regression does not model the full distribution, but only the first two moments of the distribution: the mean μ_{itb} and the variance v_{itb}. The specification of the mean model is identical to Poisson regression. However, the variance is more flexible:

$$v_{itb} = \varphi_i \mu_{itb} \quad \text{(Eq. 4.60)}$$

The factor φ_i allows to to correct for over- or under-dispersion. Count data can also be proposed to follow a negative binomial distribution, which assumes a quadratic mean-variance relationship:

$$v_{itb} = \mu_{itb} + \varphi_i \mu_{itb}^2 \quad \text{(Eq. 4.61)}$$

A negative binomial distribution reduces to a Poisson distribution if φ_i is zero. Negative binomial generalized linear models are implemented in the popular RNA sequencing quantification packages EdgeR [486, 487] and DESeq2 [488], which have also been applied in proteomics contexts [489].

Other statistical models have been proposed as well to deal with count data in proteomics, including a generalized linear mixed effects Poisson regression model in which all proteins are modeled together [490], a beta-binomial model [491] and Bayesian models [492].

Peptide counting is an alternative to spectral counting. In peptide counting, the number of unique peptides instead of the number of unique PSMs that match to each protein are counted. Some authors reported that spectral counting is more accurate and more reproducible than peptide counting which is in turn more reproducible than sequence coverage-based approaches [483, 493, 494]. This is probably because spectral counting is more fine-grained than peptide counting (there at least as much PSMs as peptides per protein), which might make it more feasible to quantify smaller differences in abundance.

A very simple peptide counting method is the Exponentially Modified Protein Abundance Index (emPAI). For each protein i, the emPAI is calculated as follows [174]:

$$\text{emPAI}_i = 10^{n_i^{\text{pep}}/n_i^{\text{predpep}}} - 1 \quad \text{(Eq. 4.62)}$$

Hereby, n_i^{pep} is the number of observed unique peptides mapping to protein i and n_i^{predpep} is the number of tryptic peptides that can theoretically map to protein i. emPAI is an example of so-called “absolute protein quantification” method. By normalizing the peptide count of each protein by its number of predicted unique tryptic peptides n_i^{predpep}, emPAI claims to be able to compare the abundances of different proteins to each other (as opposed to comparing the abundances of the same proteins over different conditions, so-called “relative quantification”). The content of protein i in mol % is then calculated as follows:

$$\text{Protein } i \text{ content (mol\%)} = \frac{\text{emPAI}_i}{\sum_{k=1}^{n^{\text{protein}}} \text{emPAI}_k}, \qquad \text{(Eq. 4.63)}$$

with n^{protein} the total number of proteins in the dataset. Of course, for quantitative protein inference, peptide counting methods also require some statistical modeling. The models for peptide counting are very similar to those of spectral counting and many statistical models for spectral counting and peptide counting can be used interchangeably.

Absolute Protein Expression (APEX) is an example of an absolute quantification method based on spectral counting that has gained quite some traction [175, 495]. An APEX score for protein i is calculated as follows:

$$\text{APEX}_i = \frac{C n_i^{\text{MS2}} \pi_i^{\text{ID}}}{n_i^{\text{predPSM}} \sum_{k=1}^{n^{\text{protein}}} \frac{n_k^{\text{MS2}} \pi_k^{\text{ID}}}{n_k^{\text{predPSM}}}}, \qquad \text{(Eq. 4.64)}$$

with n_i^{MS2} the total number of MS² spectra mapping to protein i, π_i^{ID} the probability that protein i is correctly identified, n_i^{predPSM} the number of computationally predicted PSMs for protein i and C an estimate of the total concentration of protein molecules in the cell.

Do note that although absolute quantification methods might give some indication about a protein’s abundance, these estimates remain generally very crude because normalizing protein abundances to each other based on e.g. the theoretical number of tryptic peptides they can generate is quite inaccurate. Also, count-based approaches have become largely obsolete in present-day proteomics given that quantification based on continuous intensity-based signals is clearly superior over discrete counts [317]. It has indeed been shown that count-based methods have a lower linear response to various protein loading amounts, a lower reproducibility, a lower quantitative accuracy, a lower precision, lower sensitivity, and a higher ratio of false positives to false negatives compared to MS intensity-based quantification [307]. This is because peptide counting disregards the inherent abundance-intensity relationship (within a certain dynamic range [496]) for each peptide. Dynamic exclusion, during which identified ions are excluded from being re-targeted for fragmentation for a certain amount of time, further obscures the relationship between spectral counts and protein abundances [497]. Counting-based approaches perform especially poorly for low-abundant proteins [228, 483, 498]. Indeed, it is not possible to calculate an accurate protein ratio between two conditions if only one or two spectra per condition are mapped to the protein of interest [497]. And, at higher levels of protein abundances, saturation effects come into play for relatively low total protein concentrations when all peptides that can theoretically be detected are in fact detected [228]. Spectral counting is reviewed extensively in Lundgren *et al.* (2010) [494].

4.2.7. Controlling the false discovery rate

Whether statistical inference is done with t-tests, linear regression models, or other statistical approaches, the fact that statistical inference is done for each protein in the dataset creates a

huge multiple testing problem. Imagine testing 1,000 proteins, all of which are not differentially abundant (i.e. the null hypothesis is true). When using the traditional 5% cut-off at the p-value level, on average 5%, i.e. 50 proteins, will be erroneously declared differentially abundant (false positives). Therefore, it is clear that, in the case of multiple testing, a significance threshold based on p-values will be way too liberal.

Solutions to this problem have been proposed in the form of controlling the family-wise error rate (FWER). The aim of FWER procedures is to control the probability of detecting at least a single false positive at a given level, typically 5%. An example of an FWER procedure is the simple, but somewhat conservative Bonferroni correction [499]. Here, a protein is only declared significantly differentially abundant if its p-value is smaller than $\alpha/n^{\text{protein}}$, with α the significant threshold (e.g. 5%) and n^{protein} the number of proteins that are being tested. It turns out that FWER procedures are often too conservative for high-throughput applications. Indeed, the number of biological replicates in such a context is often rather low, which limits the statistical power of each individual test.

To cope with the specific context of high-throughput experiments, false discovery rate (FDR) procedures were developed. The false discovery rate aims to control the expected fraction of false positive proteins in a list of differentially abundant proteins at a certain level, again, typically 5% [500]. In practice, FDR-controlled lists are much more interesting for practitioners: a researcher will prefer a list of 20 significant proteins of which on average 1 is a false positive (i.e. the FDR is controlled at 5%) over a list of maybe 2 or 3 proteins that are not in error according to the FWER criterion. The most well-known and most widely used FDR procedure is the Benjamini-Hochberg FDR [501]. The procedure works as follows:

In a first step, the p-values are sorted from large to small. If I p-values need to be FDR corrected, let $i = 1, \dots, I$ be the rank of the ith p-value. Then, the q-value q_i for the ith p-value is calculated as follows:

$$q_i = \min_k \left(1, \frac{I p_k}{I - k + 1}\right), \qquad \text{(Eq. 4.65)}$$

for $1 \leq k \leq i$. All proteins with a q-value smaller than the proposed threshold, e.g. 5%, are then considered statistically significant.

5. RESEARCH HYPOTHESIS

5.1. Setting the stage

Many biological processes strongly depend on balanced levels of protein expression and the perturbation of a single protein can already lead to organismal malfunctioning (e.g. abnormal hemoglobin production in thalassemia [47] or extensive cellular remodeling towards a cancerous phenotype [502]). In this respect, quantitative knowledge of a proteome is very important for the unraveling of the development and progression of diseases and the identification of biomarkers, amongst others. It is fair to state that the transcriptome and the translatome only reflect the proteins that can be or are being expressed, but largely fail to provide information on a protein's activity, function, interaction partners, localization and modification state (see 1.2.3). As proteins and their modified variants are expressed over a large concentration range, accurate quantitative information is a must to distinguish different cellular and organismal conditions [503].

Label-free shotgun proteomics leads to the identification and quantification of thousands of peptides and proteins in a single experiment. Here, analysis of differential abundance of proteins is based on ratios derived from reconstructed elution profiles based on MS intensities for all PSMs pointing to the same protein or the same protein group in a sample (see section 3.2). However, the data are highly hierarchical, and intensities can be strongly influenced by variations of peptide-specific properties. In addition, the number of peptides identified in all samples is usually limited, which leads to large numbers of missing values (Fig. 5.1), which do not occur at random. Some peptides are better detected than others and high-abundant peptide ions are more likely to yield higher numbers of fragmentation spectra. "Match between runs" algorithms can only partly compensate for this effect by reducing the number of missing values (see section 3.4).

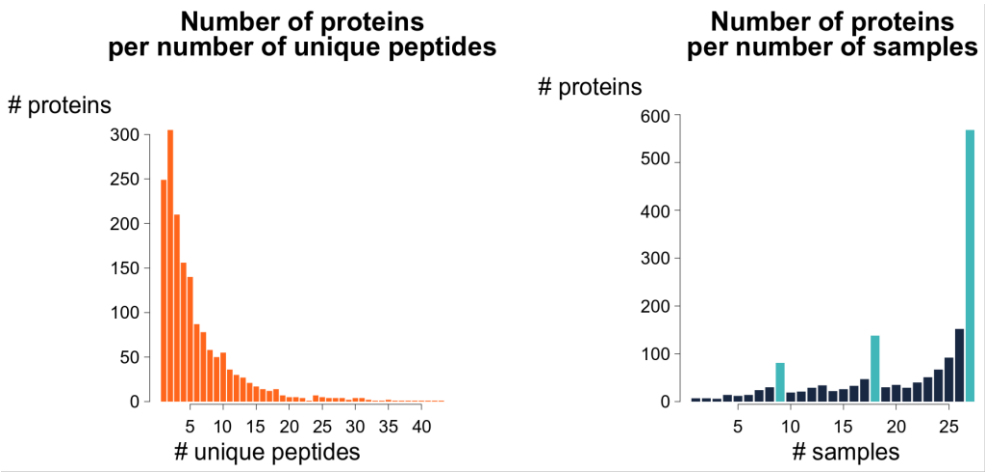


Figure 5.12. The missing value problem. Left: the number of unique peptides identified per protein in the CPTAC dataset [426]. Right: the number of samples in which each protein in the CPTAT dataset is identified. The number of proteins identified in 9, 18 and all 27 samples are markedly elevated (cyan bars). This is because of the higher numbers of proteins that are exclusively identified in all MS runs from 1, 2 or 3 labs respectively.

Researchers have used a plethora of pipelines for the analysis of label-free shotgun proteomics data (see chapter 4). Indeed, since most mass spectrometry researchers do not have a background in statistics, different preprocessing and differential analysis methods are often combined *ad hoc* without a proper motivation. Many methods analyze the data protein

by protein by (a) filtering out proteins as soon as they are missing in a specified number of samples (see 4.1.2), (b) summarize the peptides for each protein in a sample and ignore missing values (see 4.1.5) or (c) impute missing values based on the observed peptides for the corresponding protein in other samples (see 4.1.4). Filtering out all proteins (a) leads to substantial information loss. Indeed, as demonstrated in Fig. 5.1, removing proteins that are only identified with more than a given number of peptides or requiring proteins to be identified in all of the samples would remove a substantial amount of proteins from the dataset. Summarization (b) needs to be executed such that it correctly takes the peptide-specific effects into account. A particular challenge here are the limited numbers of peptides that are identified across different samples. If summarization does not correctly take the differences in ionization efficiencies into account, it will introduce a bias because the final data will be based on different peptides. Moreover, summarization will always ignore differences in accuracy due to the different numbers of peptides on which each summary is based. Imputation (c) is even more tedious because missing values in proteomics are a combination of missing completely at random and intensity- and even context-dependent missingness. Since it is impossible to know the exact contributions of these different types of missingness, imputation according to incorrect assumptions will lead to biased quantifications. However, if missing values are simply ignored, low-abundant proteins will be over-estimated due to the limited linear dynamic range, which will reduce the power to detect differential abundance.

In addition, a proteomics experiment typically consists of a relatively small number of biological repeats compared to the large number of proteins analyzed. This gives rise to unstable variance estimates for certain proteins, especially if these proteins are identified with only a few peptides. Indeed, a small sample drawn from a given population might display a variance that is much smaller (or much bigger) than the true population variance just by random chance. Therefore, some observations might be flagged as differentially abundant solely because of their low observed variances while other truly differentially abundant proteins might be missed due to their large observed variances. This data sparsity also leads to unstable fold change estimates: one or two outlying intensities, which can for example be caused by misidentifications or co-eluting peptides, can strongly influence a protein's fold change.

These unstable fold change and variance estimates may cause a significant increase in the number of false positives and false negatives (see e.g. the supplementary material of Doll *et al.* (2017) [504], where SERINC3 and PNMA1 were declared significantly altered with very weak evidence based on only a single peptide). It is therefore not surprising that a recent power calculation study on four biological repeats of *Arabidopsis thaliana* Col-0 samples analyzed in technical triplicate estimated that for MaxLFQ summaries, a minimal fold change of 1.4 is required to detect a statistically significant difference with 95% confidence and a power of 80%[463].

State-of-the-art proteomics quantification methods only focus on some of the sub-problems mentioned above. They generally remain fairly sensitive to outliers and correct for missing observations only to a limited extent. In addition, they usually prune the number of potential hits by filtering out proteins which have few peptides in common across the majority of the samples. Methods that model peptide-level data are often more sensitive because they naturally correct for the correlation present within the same samples, as well as for the peptide effects and the number of detected peptides for a given protein in each sample [425, 471].

Like summarization-based methods, many peptide-based models still produce unstable estimates of differential abundance and variance components due to over-parameterization. This especially occurs when few peptides are identified per protein: many low-abundant proteins are then falsely labeled as differentially abundant. Filtering on the basis of the number

of identified peptide spectra can provide a solution, but there is a risk that real hits will also be filtered out. Moreover, most peptide-based models remain very sensitive to outliers.

In addition, many methods are only suitable to analyze specific types of experimental designs. For example, in Perseus, it is possible to compare multiple groups to each other, but it is not possible to accurately analyze blocked designs, such as CPTAC. In MSstats, the user should annotate each experiment with a “Run”, “Condition” and “BioReplicate”. Hereby, “Run” refers to the MS run, “Condition” to the treatment of interest and “BioReplicate” to a grouping factor that is encoded as a random effect. While this set-up allows most simple experimental designs to be analyzed in a correct way, it is insufficiently flexible to correctly analyze more complicated designs with multiple confounding effects.

Finally, state-of-the-art data analysis methods, such as published models at the peptide level, do not always find their way to proteomic labs [474, 475]. This is a non-negligible problem! Indeed, many groups only demonstrated proof-of-concept, but never developed their method into a usable software package. And, those who did often lack a convenient graphical user interface that is appealing to less-experienced users. Availability, user-friendliness, documentation, support and maintenance of software tools are very important for new methods to be effective for end users.

5.2. Aims of my PhD work

Based on the previous section, it is clear that state-of-the-art proteomics quantification methods do not yet make optimal use of the data, resulting in suboptimal protein quantifications. The development of robust data analysis tools for quantitative proteomics is therefore essential for the further development of the proteomics research field because, with the current data-analytical methods, many proteins remain under the radar [505]. The overall aim of my work was to develop a more robust, easy-to use proteomics quantification method that is also usable for the analysis of proteins with a limited overlap in identified peptides.

More specifically, this method should:

- account for the hierarchical nature of the data,
- handle missing peptides in a more correct way,
- derive strength from the massive parallel availability of peptides to estimate variance components more correctly,
- be robust to outliers,
- and be able to handle complex experimental designs.

In such a method, it is important that a random effect for run is included in order to allow to correct for within-run correlation. In this respect, it is tempting to build upon the link between mixed models and ridge regression, to make use of robust regression with M estimation and to allow for empirical Bayes variance estimation. Implementing this method is expected to increase the number of truly differentially abundant proteins identified in screening experiments. The method should also be implemented and distributed in a user-friendly software tool for differential proteomics with the possibility of a graphical user interface to maximize the impact of the research.

To develop such a method, it is first necessary to thoroughly benchmark the most promising and most commonly used quantification methods. This will shed light on how preprocessing, standardization and differential analysis in a label-free proteomics quantification workflow influence a method’s performance.

6. OUTLINE

The remainder of my thesis is constructed as follows: first, I will present each of my published papers. In my first paper, I demonstrate how I compared different data analysis methods for differential quantification in label-free shotgun proteomics. My second paper presents the rationale behind MSqRob, the algorithm I developed to improve quantification in label-free shotgun proteomics. In my third paper, I provide a tutorial on experimental design and data analysis with MSqRob. In my fourth, unpublished paper, I make use of the additional information of peptide counts to boost MSqRob's power and to indicate whether a protein's significance is mainly driven by differential abundance, differential detection or both.

In the discussion, I explore the significance of my work and place it in a broader context. In the future research perspectives, I give some indications on how my research could go on from here.

7. REFERENCES PART I

1. Gladyshev, V.N. and G.V. Kryukov, *Evolution of selenocysteine-containing proteins: Significance of identification and functional characterization of selenoproteins.* BioFactors, 2001. **14**(1-4): p. 87-92.
2. Prat, L. *et al.*, *Carbon source-dependent expansion of the genetic code in bacteria.* Proceedings of the National Academy of Sciences, 2012. **109**(51): p. 21070-21075.
3. Koch, A. *et al.*, *A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites.* Proteomics, 2014. **14**(23-24): p. 2688-2698.
4. Ruiz-Orera, J. *et al.*, *Long non-coding RNAs as a source of new peptides.* eLife, 2014. **3**: p. e03523-e03523.
5. Rion, N. and M.A. Rüegg, *LncRNA-encoded peptides: More than translational noise?* Cell Research, 2017. **27**: p. 604.
6. Choi, S.-W., H.-W. Kim, and J.-W. Nam, *The small peptide world in long noncoding RNAs.* Briefings in Bioinformatics, 2018: p. bby055-bby055.
7. Gebert, L.F.R. and I.J. MacRae, *Regulation of microRNA function in animals.* Nature Reviews Molecular Cell Biology, 2019. **20**(1): p. 21-37.
8. Bhat, S.A. *et al.*, *Long non-coding RNAs: Mechanism of action and functional utility.* Non-coding RNA Research, 2016. **1**(1): p. 43-50.
9. Brimacombe, R. and W. Stiege, *Structure and function of ribosomal RNA.* The Biochemical journal, 1985. **229**(1): p. 1-17.
10. Walter, N.G. and D.R. Engelke, *Ribozymes: catalytic RNAs that cut things, make things, and do odd and useful jobs.* Biologist (London, England), 2002. **49**(5): p. 199-203.
11. Müller, S. *et al.*, *Thirty-five years of research into ribozymes and nucleic acid catalysis: where do we stand today?* F1000Research, 2016. **5**: p. F1000 Faculty Rev-1511.
12. Mühlhausen, S. *et al.*, *Endogenous Stochastic Decoding of the CUG Codon by Competing Ser- and Leu-tRNAs in Ascoidea asiatica.* Current Biology, 2018. **28**(13): p. 2046-2057.e5.
13. Hofhuis, J. *et al.*, *The functional readthrough extension of malate dehydrogenase reveals a modification of the genetic code.* Open Biology, 2016. **6**(11): p. 160246.
14. Inamine, J.M. *et al.*, *Evidence that UGA is read as a tryptophan codon rather than as a stop codon by Mycoplasma pneumoniae, Mycoplasma genitalium, and Mycoplasma gallisepticum.* Journal of Bacteriology, 1990. **172**(1): p. 504-506.
15. Piatkov, K.I. *et al.*, *Formyl-methionine as a degradation signal at the N-termini of bacterial proteins.* Microbial Cell, 2015. **2**(10): p. 376-393.
16. Wingfield, P., *N-Terminal Methionine Processing.* Current Protocols in Protein Science, 2017. **88**: p. 6.14.1-6.14.3.
17. Falb, M. *et al.*, *Archaeal N-terminal Protein Maturation Commonly Involves N-terminal Acetylation: A Large-scale Proteomics Survey.* Journal of Molecular Biology, 2006. **362**(5): p. 915-924.
18. Jonckheere, V., D. Fijałkowska, and P. Van Damme, *Omics Assisted N-terminal Proteoform and Protein Expression Profiling On Methionine Aminopeptidase 1 (MetAP1) Deletion.* Molecular & Cellular Proteomics, 2018. **17**(4): p. 694-708.
19. Belinky, F., I.B. Rogozin, and E.V. Koonin, *Selection on start codons in prokaryotes and potential compensatory nucleotide substitutions.* Scientific Reports, 2017. **7**(1): p. 12422.
20. Kearse, M.G. and J.E. Wilusz, *Non-AUG translation: a new start for protein synthesis in eukaryotes.* Genes & Development, 2017. **31**(17): p. 1717-1731.
21. Hecht, A. *et al.*, *Measurements of translation initiation from all 64 codons in E. coli.* Nucleic Acids Research, 2017. **45**(7): p. 3615-3626.
22. Jungreis, I. *et al.*, *Evidence of abundant stop codon readthrough in Drosophila and other metazoa.* Genome Research, 2011. **21**(12): p. 2096-2113.

23. Atkins, J.F. *et al.*, *Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use.* Nucleic Acids Research, 2016. **44**(15): p. 7007-7078.
24. Dinman, J.D., *Programmed Ribosomal Frameshifting Goes Beyond Viruses: Organisms from all three kingdoms use frameshifting to regulate gene expression, perhaps signaling a paradigm shift.* Microbe (Washington, D.C.), 2006. **1**(11): p. 521-527.
25. Baranov, P.V. *et al.*, *Transcriptional slippage in bacteria: distribution in sequenced genomes and utilization in IS element gene expression.* Genome Biology, 2005. **6**(3): p. R25-R25.
26. Freeman, M.F. *et al.*, *Seven enzymes create extraordinary molecular complexity in an uncultivated bacterium.* Nature Chemistry, 2016. **9**: p. 387.
27. Morinaka, B.I. *et al.*, *Natural noncanonical protein splicing yields products with diverse β-amino acid residues.* Science, 2018. **359**(6377): p. 779-782.
28. Berk, A.J., *Discovery of RNA splicing and genes in pieces.* Proceedings of the National Academy of Sciences, 2016. **113**(4): p. 801-805.
29. Vila-Perelló, M. and T.W. Muir, *Biological Applications of Protein Splicing.* Cell, 2010. **143**(2): p. 191-200.
30. Smith, L.M., N.L. Kelleher, and P. The Consortium for Top Down, *Proteoform: a single term describing protein complexity.* Nature Methods, 2013. **10**(3): p. 186-187.
31. Pace, C.N., J.M. Scholtz, and G.R. Grimsley, *Forces stabilizing proteins.* FEBS Letters, 2014. **588**(14): p. 2177-2184.
32. Saibil, H., *Chaperone machines for protein folding, unfolding and disaggregation.* Nature Reviews Molecular Cell Biology, 2013. **14**(10): p. 630-642.
33. Hartl, F.U., A. Bracher, and M. Hayer-Hartl, *Molecular chaperones in protein folding and proteostasis.* Nature, 2011. **475**: p. 324.
34. Fuhs, S.R. and T. Hunter, *pHisphorylation; The Emergence of Histidine Phosphorylation as a Reversible Regulatory Modification.* Current Opinion in Cell Biology, 2017. **45**: p. 8-16.
35. Potel, C.M. *et al.*, *Widespread bacterial protein histidine phosphorylation revealed by mass spectrometry-based proteomics.* Nature Methods, 2018. **15**: p. 187.
36. Besant, P.G., P.V. Attwood, and M.J. Piggott, *Focus on Phosphoarginine and Phospholysine.* Current Protein & Peptide Science, 2009. **10**(6): p. 536-550.
37. Attwood, P.V., P.G. Besant, and M.J. Piggott, *Focus on phosphoaspartate and phosphoglutamate.* Amino Acids, 2011. **40**(4): p. 1035-1051.
38. Hardman, G. *et al.*, *Extensive non-canonical phosphorylation in human cells revealed using strong-anion exchange-mediated phosphoproteomics.* bioRxiv, 2017.
39. Mijakovic, I., C. Grangeasse, and K. Turgay, *Exploring the diversity of protein modifications: special bacterial phosphorylation systems.* FEMS Microbiology Reviews, 2016. **40**(3): p. 398-417.
40. Jean Beltran, P.M. *et al.*, *Proteomics and integrative omic approaches for understanding host-pathogen interactions and infectious diseases.* Molecular Systems Biology, 2017. **13**(3): p. 922-922.
41. Doyle, H.A. and M.J. Mamula, *Autoantigenesis: the evolution of protein modifications in autoimmune disease.* Current Opinion in Immunology, 2012. **24**(1): p. 112-118.
42. Chung, K.K.K. *et al.*, *S-Nitrosylation of Parkin Regulates Ubiquitination and Compromises Parkin's Protective Function.* Science, 2004. **304**(5675): p. 1328-1331.
43. Ren, R.-J. *et al.*, *Proteomics of protein post-translational modifications implicated in neurodegeneration.* Translational Neurodegeneration, 2014. **3**(1): p. 23-23.
44. Creasy, D.M. and J.S. Cottrell, *Unimod: Protein modifications for mass spectrometry.* Proteomics, 2004. **4**(6): p. 1534-1536.
45. Ideker, T. and R. Sharan, *Protein networks in disease.* Genome Research, 2008. **18**(4): p. 644-652.
46. Marengo-Rowe, A.J., *The thalassemias and related disorders.* Proceedings (Baylor University. Medical Center), 2007. **20**(1): p. 27-31.

47. Thein, S.L., *The Molecular Basis of β-Thalassemia.* Cold Spring Harbor Perspectives in Medicine, 2013. **3**(5): p. a011700.
48. Svartman, M., G. Stone, and R. Stanyon, *Molecular cytogenetics discards polyploidy in mammals.* Genomics, 2005. **85**(4): p. 425-430.
49. Gatz, M. *et al.*, *Role of genes and environments for explaining alzheimer disease.* Archives of General Psychiatry, 2006. **63**(2): p. 168-174.
50. Diaz-Espinoza, R. *et al.*, *Treatment with a Non-toxic, Self-replicating Anti-prion Delays or Prevents Prion Disease In vivo.* Molecular psychiatry, 2018. **23**(3): p. 777-788.
51. Füzéry, A.K. *et al.*, *Translation of proteomic biomarkers into FDA approved cancer diagnostics: issues and challenges.* Clinical Proteomics, 2013. **10**(1): p. 13-13.
52. Zhang, Q.C. *et al.*, *Structure-based prediction of protein–protein interactions on a genome-wide scale.* Nature, 2012. **490**: p. 556.
53. Makley, L.N. and J.E. Gestwicki, *Expanding the Number of "Druggable" Targets: Non-Enzymes and Protein-Protein Interactions.* Chemical Biology & Drug Design, 2013. **81**(1): p. 22-32.
54. Śledź, P. and A. Caflisch, *Protein structure-based drug design: from docking to molecular dynamics.* Current Opinion in Structural Biology, 2018. **48**: p. 93-102.
55. Murata, K. and M. Wolf, *Cryo-electron microscopy for structural analysis of dynamic biological macromolecules.* Biochimica et Biophysica Acta (BBA) - General Subjects, 2018. **1862**(2): p. 324-334.
56. Pearson, W.R. and M.L. Sierk, *The limits of protein sequence comparison?* Current Opinion in Structural Biology, 2005. **15**(3): p. 254-260.
57. Lagassé, H.A.D. *et al.*, *Recent advances in (therapeutic protein) drug development.* F1000Research, 2017. **6**: p. 113.
58. Fala, L., *Nucala (Mepolizumab): First IL-5 Antagonist Monoclonal Antibody FDA Approved for Maintenance Treatment of Patients with Severe Asthma.* American Health & Drug Benefits, 2016. **9**(Spec Feature): p. 106-110.
59. Raedler, L.A., *Empliciti (Elotuzumab): First SLAMF7 Antibody Therapy Approved for the Treatment of Patients with Previously Treated Multiple Myeloma.* American Health & Drug Benefits, 2016. **9**(Spec Feature): p. 74-77.
60. Singh, A.D. and S. Parmar, *Ramucirumab (Cyramza): A Breakthrough Treatment for Gastric Cancer.* Pharmacy and Therapeutics, 2015. **40**(7): p. 430-468.
61. Khan, M.A. and J.A. Haller, *Ocriplasmin for Treatment of Vitreomacular Traction: An Update.* Ophthalmology and Therapy, 2016. **5**(2): p. 147-159.
62. Ramsey, L.B. *et al.*, *Consensus Guideline for Use of Glucarpidase in Patients with High-Dose Methotrexate Induced Acute Kidney Injury and Delayed Methotrexate Clearance.* The Oncologist, 2017.
63. Franchini, M. and P.M. Mannucci, *Von Willebrand factor (Vonvendi®): the first recombinant product licensed for the treatment of von Willebrand disease.* Expert Review of Hematology, 2016. **9**(9): p. 825-830.
64. Zimmer, M., *Green Fluorescent Protein (GFP): Applications, Structure, and Related Photophysical Behavior.* Chemical Reviews, 2002. **102**(3): p. 759-782.
65. Hsu, P.D., E.S. Lander, and F. Zhang, *Development and Applications of CRISPR-Cas9 for Genome Engineering.* Cell, 2014. **157**(6): p. 1262-1278.
66. Robinson, R., *What Governs Enzyme Activity? For One Enzyme, Charge Contributes Only Weakly.* PLoS Biology, 2006. **4**(4): p. e133.
67. Singh, R. *et al.*, *Microbial enzymes: industrial progress in 21st century.* 3 Biotech, 2016. **6**(2): p. 174.
68. Wells, A.S. *et al.*, *Use of Enzymes in the Manufacture of Active Pharmaceutical Ingredients—A Science and Safety-Based Approach To Ensure Patient Safety and Drug Quality.* Organic Process Research & Development, 2012. **16**(12): p. 1986-1993.
69. de Souza, P.M. and P. de Oliveira Magalhães, *Application of microbial α-amylase in industry – A review.* Brazilian Journal of Microbiology, 2010. **41**(4): p. 850-861.
70. Garg, G. *et al.*, *Microbial pectinases: an ecofriendly tool of nature for industries.* 3 Biotech, 2016. **6**(1): p. 47.

71. Saqib, S. *et al.*, *Sources of β-galactosidase and its applications in food industry.* 3 Biotech, 2017. **7**(1): p. 79.
72. Przybysz Buzała, K. *et al.*, *Effect of Cellulases and Xylanases on Refining Process and Kraft Pulp Properties.* PLOS ONE, 2016. **11**(8): p. e0161575.
73. Olsen, H.S. and P. Falholt, *The role of enzymes in modern detergency.* Journal of Surfactants and Detergents, 1998. **1**(4): p. 555-567.
74. Noraini, M.Y. *et al.*, *A review on potential enzymatic reaction for biofuel production from algae.* Renewable and Sustainable Energy Reviews, 2014. **39**: p. 24-34.
75. Ye, X. *et al.*, *Engineering the Provitamin A (β-Carotene) Biosynthetic Pathway into (Carotenoid-Free) Rice Endosperm.* Science, 2000. **287**(5451): p. 303-305.
76. Paine, J.A. *et al.*, *Improving the nutritional value of Golden Rice through increased pro-vitamin A content.* Nature Biotechnology, 2005. **23**: p. 482.
77. Dawe, D., R. Robertson, and L. Unnevehr, *Golden rice: what role could it play in alleviation of vitamin A deficiency?* Food Policy, 2002. **27**(5): p. 541-560.
78. Tang, G. *et al.*, *Golden Rice is an effective source of vitamin A.* The American Journal of Clinical Nutrition, 2009. **89**(6): p. 1776-1783.
79. Tang, G. *et al.*, *β-Carotene in Golden Rice is as good as β-carotene in oil at providing vitamin A to children.* The American Journal of Clinical Nutrition, 2012. **96**(3): p. 658-664.
80. *Micronutrient deficiencies*. World Health Organization. Acessed on: Available from: http://www.who.int/nutrition/topics/vad/en (cited October 15th 2018).
81. Braun, P. *et al.*, *Plant Protein Interactomes.* Annual Review of Plant Biology, 2013. **64**(1): p. 161-187.
82. Pandey, P. *et al.*, *Impact of Combined Abiotic and Biotic Stresses on Plant Growth and Avenues for Crop Improvement by Exploiting Physio-morphological Traits.* Frontiers in Plant Science, 2017. **8**: p. 537.
83. Luo, M. *et al.*, *Comparative Proteomics of Contrasting Maize Genotypes Provides Insights into Salt-Stress Tolerance Mechanisms.* Journal of Proteome Research, 2018. **17**(1): p. 141-153.
84. Michaletti, A. *et al.*, *Metabolomics and proteomics reveal drought-stress responses of leaf tissues from spring-wheat.* Scientific Reports, 2018. **8**(1): p. 5710.
85. Brun, G. *et al.*, *Seed germination in parasitic plants: what insights can we expect from strigolactone research?* Journal of Experimental Botany, 2018. **69**(9): p. 2265-2280.
86. Vékey, K., A. Telekes, and A. Vertes, *Medical Applications of Mass Spectrometry*. 2008, Amsterdam: Elsevier. 561-581.
87. Winter, D. and H. Steen, *Optimization of cell lysis and protein digestion protocols for the analysis of HeLa S3 cells by LC-MS/MS.* Proteomics, 2011. **11**(24): p. 4726-4730.
88. Moore, S.M., S.M. Hess, and J.W. Jorgenson, *Extraction, Enrichment, Solubilization, and Digestion Techniques for Membrane Proteomics.* Journal of Proteome Research, 2016. **15**(4): p. 1243-1252.
89. Compton, P.D. *et al.*, *Native Proteomics: A New Approach to Protein Complex Discovery and Characterization.* The FASEB Journal, 2017. **31**(1_supplement): p. 760.2-760.2.
90. Karpievitch, Y.V. *et al.*, *Liquid Chromatography Mass Spectrometry-Based Proteomics: Biological and Technological Aspects.* The annals of applied statistics, 2010. **4**(4): p. 1797-1823.
91. Zhang, Y. *et al.*, *Protein Analysis by Shotgun/Bottom-up Proteomics.* Chemical Reviews, 2013. **113**(4): p. 2343-2394.
92. Cristobal, A. *et al.*, *Toward an Optimized Workflow for Middle-Down Proteomics.* Analytical Chemistry, 2017. **89**(6): p. 3318-3325.
93. Choudhary, G. *et al.*, *Multiple Enzymatic Digestion for Enhanced Sequence Coverage of Proteins in Complex Proteomic Mixtures Using Capillary LC with Ion Trap MS/MS.* Journal of Proteome Research, 2003. **2**(1): p. 59-67.

94. Swaney, D.L., C.D. Wenger, and J.J. Coon, *Value of using multiple proteases for large-scale mass spectrometry-based proteomics.* Journal of Proteome Research, 2010. **9**(3): p. 1323-1329.
95. López-Ferrer, D. *et al.*, *Pressurized Pepsin Digestion in Proteomics.* AN AUTOMATABLE ALTERNATIVE TO TRYPSIN FOR INTEGRATED TOP-DOWN BOTTOM-UP PROTEOMICS*, 2011. **10**(2): p. M110.001479.
96. Peng, M. *et al.*, *Protease bias in absolute protein quantitation.* Nature Methods, 2012. **9**(6): p. 524-525.
97. Meyer, J.G. *et al.*, *Expanding proteome coverage with orthogonal-specificity α-lytic proteases.* Molecular & Cellular Proteomics, 2014. **13**(3): p. 823-835.
98. Guo, X. *et al.*, *Confetti: A Multiprotease Map of the HeLa Proteome for Comprehensive Proteomics.* Molecular & Cellular Proteomics, 2014. **13**(6): p. 1573-1584.
99. Giansanti, P. *et al.*, *Six alternative proteases for mass spectrometry–based proteomics beyond trypsin.* Nature Protocols, 2016. **11**: p. 993.
100. Bian, Y. *et al.*, *Improve the Coverage for the Analysis of Phosphoproteome of HeLa Cells by a Tandem Digestion Approach.* Journal of Proteome Research, 2012. **11**(5): p. 2828-2837.
101. Huesgen, P.F. *et al.*, *LysargiNase mirrors trypsin for protein C-terminal and methylation-site identification.* Nature Methods, 2014. **12**: p. 55.
102. Tsiatsiani, L. and A.J.R. Heck, *Proteomics beyond trypsin.* The FEBS Journal, 2015. **282**(14): p. 2612-2626.
103. Wu, C. *et al.*, *A protease for 'middle-down' proteomics.* Nature Methods, 2012. **9**: p. 822.
104. Zhang, X., *Less is More: Membrane Protein Digestion Beyond Urea-Trypsin Solution for Next-level Proteomics.* Molecular & cellular proteomics, 2015. **14**(9): p. 2441-2453.
105. Chen, E.I. *et al.*, *Optimization of mass spectrometry-compatible surfactants for shotgun proteomics.* Journal of Proteome Research, 2007. **6**(7): p. 2529-2538.
106. Proc, J.L. *et al.*, *A quantitative study of the effects of chaotropic agents, surfactants, and solvents on the digestion efficiency of human plasma proteins by trypsin.* Journal of Proteome Research, 2010. **9**(10): p. 5422-5437.
107. Rundlett, K.L. and D.W. Armstrong, *Mechanism of Signal Suppression by Anionic Surfactants in Capillary Electrophoresis−Electrospray Ionization Mass Spectrometry.* Analytical Chemistry, 1996. **68**(19): p. 3493-3497.
108. Botelho, D. *et al.*, *Top-Down and Bottom-Up Proteomics of SDS-Containing Solutions Following Mass-Based Separation.* Journal of Proteome Research, 2010. **9**(6): p. 2863-2870.
109. Ilavenil, S. *et al.*, *Removal of SDS from biological protein digests for proteomic analysis by mass spectrometry.* Proteome Science, 2016. **14**(1): p. 11.
110. HaileMariam, M. *et al.*, *S-Trap, an Ultrafast Sample-Preparation Approach for Shotgun Proteomics.* Journal of Proteome Research, 2018. **17**(9): p. 2917-2924.
111. Kim, S.C. *et al.*, *A Clean, More Efficient Method for In-Solution Digestion of Protein Mixtures without Detergent or Urea.* Journal of Proteome Research, 2006. **5**(12): p. 3446-3452.
112. Hodge, K. *et al.*, *Cleaning up the masses: Exclusion lists to reduce contamination with HPLC-MS/MS.* Journal of Proteomics, 2013. **88**: p. 92-103.
113. Gunawardena, H.P., J.F. Emory, and S.A. McLuckey, *Phosphopeptide Anion Characterization via Sequential Charge Inversion and Electron-Transfer Dissociation.* Analytical Chemistry, 2006. **78**(11): p. 3788-3793.
114. Chouchani, E.T. *et al.*, *Proteomic approaches to the characterization of protein thiol modification.* Current Opinion in Chemical Biology, 2011. **15**(1): p. 120-128.
115. Riley, N.M. and J.J. Coon, *Phosphoproteomics in the Age of Rapid and Deep Proteome Profiling.* Analytical Chemistry, 2016. **88**(1): p. 74-94.
116. Swaney, D.L. and J. Villén, *Proteomic Analysis of Protein Posttranslational Modifications by Mass Spectrometry.* Cold Spring Harbor Protocols, 2016. **2016**(3): p. pdb.top077743.

117. Doll, S. and A.L. Burlingame, *Mass Spectrometry-Based Detection and Assignment of Protein Posttranslational Modifications.* ACS Chemical Biology, 2015. **10**(1): p. 63-71.
118. Nagaraj, N. *et al.*, *System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap.* Molecular & Cellular Proteomics, 2012. **11**(3): p. M111.013722-M111.013722.
119. Washburn, M.P., D. Wolters, and J.R. Yates, 3rd, *Large-scale analysis of the yeast proteome by multidimensional protein identification technology.* Nature Biotechnology, 2001. **19**(3): p. 242-7.
120. Panchaud, A. *et al.*, *Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean.* Analytical Chemistry, 2009. **81**(15): p. 6481-6488.
121. Szájli, E., T. Fehér, and K.F. Medzihradszky, *Investigating the Quantitative Nature of MALDI-TOF MS.* Molecular & Cellular Proteomics, 2008. **7**(12): p. 2410-2418.
122. Wilm, M., *Principles of electrospray ionization.* Molecular & Cellular Proteomics, 2011. **10**(7): p. M111.009407.
123. Riley, N.M. *et al.*, *The Negative Mode Proteome with Activated Ion Negative Electron Transfer Dissociation (AI-NETD).* Molecular & Cellular Proteomics, 2015. **14**(10): p. 2644-2660.
124. Wang, N. and L. Li, *Exploring the Precursor Ion Exclusion Feature of Liquid Chromatography−Electrospray Ionization Quadrupole Time-of-Flight Mass Spectrometry for Improving Protein Identification in Shotgun Proteome Analysis.* Analytical Chemistry, 2008. **80**(12): p. 4696-4710.
125. Mitchell Wells, J. and S.A. McLuckey, *Collision-Induced Dissociation (CID) of Peptides and Proteins*, in *Methods in Enzymology*. 2005, Academic Press. p. 148-185.
126. Olsen, J.V. *et al.*, *Higher-energy C-trap dissociation for peptide modification analysis.* Nature Methods, 2007. **4**(9): p. 709-712.
127. Molina, H. *et al.*, *Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(7): p. 2199-2204.
128. Chi, A. *et al.*, *Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(7): p. 2193-2198.
129. Smith, S.A. *et al.*, *Enhanced Characterization of Singly Protonated Phosphopeptide Ions by Femtosecond Laser-induced Ionization/Dissociation Tandem Mass Spectrometry (fs-LID-MS/MS).* Journal of the American Society for Mass Spectrometry, 2010. **21**(12): p. 2031-2040.
130. Fort, K.L. *et al.*, *Implementation of Ultraviolet Photodissociation on a Benchtop Q Exactive Mass Spectrometer and Its Application to Phosphoproteomics.* Analytical Chemistry, 2016. **88**(4): p. 2303-2310.
131. Mayfield, J.E. *et al.*, *Mapping the Phosphorylation Pattern of Drosophila melanogaster RNA Polymerase II Carboxyl-Terminal Domain Using Ultraviolet Photodissociation Mass Spectrometry.* ACS Chemical Biology, 2017. **12**(1): p. 153-162.
132. Robinson, M.R. *et al.*, *193 nm Ultraviolet Photodissociation Mass Spectrometry for Phosphopeptide Characterization in the Positive and Negative Ion Modes.* Journal of Proteome Research, 2016. **15**(8): p. 2739-2748.
133. Pejchinovski, M. *et al.*, *Comparison of higher energy collisional dissociation and collision-induced dissociation MS/MS sequencing methods for identification of naturally occurring peptides in human urine.* PROTEOMICS – Clinical Applications, 2015. **9**(5-6): p. 531-542.
134. Murray Kermit, K. *et al.*, *Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013)*, in *Pure and Applied Chemistry*. 2013. p. 1515.
135. Geiger, T., J. Cox, and M. Mann, *Proteomics on an Orbitrap Benchtop Mass Spectrometer Using All-ion Fragmentation.* Molecular & Cellular Proteomics, 2010. **9**(10): p. 2252-2261.

136. Steen, H. and M. Mann, *The abc's (and xyz's) of peptide sequencing.* Nature Reviews Molecular Cell Biology, 2004. **5**: p. 699.
137. Michalski, A. *et al.*, *A Systematic Investigation into the Nature of Tryptic HCD Spectra.* Journal of Proteome Research, 2012. **11**(11): p. 5479-5491.
138. Frese, C.K. *et al.*, *Improved Peptide Identification by Targeted Fragmentation Using CID, HCD and ETD on an LTQ-Orbitrap Velos.* Journal of Proteome Research, 2011. **10**(5): p. 2377-2388.
139. Shao, C., Y. Zhang, and W. Sun, *Statistical characterization of HCD fragmentation patterns of tryptic peptides on an LTQ Orbitrap Velos mass spectrometer.* Journal of Proteomics, 2014. **109**: p. 26-37.
140. Bekker-Jensen, D.B. *et al.*, *An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes.* Cell Systems, 2017. **4**(6): p. 587-599.e4.
141. Shishkova, E., A.S. Hebert, and J.J. Coon, *Now, More Than Ever, Proteomics Needs Better Chromatography.* Cell Systems, 2016. **3**(4): p. 321-324.
142. Mann, M. *et al.*, *The Coming Age of Complete, Accurate, and Ubiquitous Proteomes.* Molecular Cell, 2013. **49**(4): p. 583-590.
143. Martens, L. and J.A. Vizcaíno, *A Golden Age for Working with Public Proteomics Data.* Trends in Biochemical Sciences, 2017. **42**(5): p. 333-341.
144. Matsumoto, A. *et al.*, *mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide.* Nature, 2016. **541**: p. 228.
145. Feigin, C.Y. *et al.*, *Genome of the Tasmanian tiger provides insights into the evolution and demography of an extinct marsupial carnivore.* Nature Ecology & Evolution, 2018. **2**(1): p. 182-192.
146. Nowoshilow, S. *et al.*, *The axolotl genome and the evolution of key tissue formation regulators.* Nature, 2018. **554**: p. 50.
147. Gutekunst, J. *et al.*, *Clonal genome evolution and rapid invasive spread of the marbled crayfish.* Nature Ecology & Evolution, 2018. **2**(3): p. 567-573.
148. Jaiswal, S.K. *et al.*, *Genome Sequence of Indian Peacock Reveals the Peculiar Case of a Glittering Bird.* bioRxiv, 2018.
149. Edwards, R.J. *et al.*, *Draft genome assembly of the invasive cane toad, Rhinella marina.* GigaScience, 2018. **7**(9): p. giy095-giy095.
150. Stricker, S.H., A. Köferle, and S. Beck, *From profiles to function in epigenomics.* Nature Reviews Genetics, 2016. **18**: p. 51.
151. Lowe, R. *et al.*, *Transcriptomics technologies.* PLOS Computational Biology, 2017. **13**(5): p. e1005457.
152. Hershey, J.W.B., N. Sonenberg, and M.B. Mathews, *Principles of Translational Control: An Overview.* Cold Spring Harbor Perspectives in Biology, 2012. **4**(12).
153. Brar, G.A. and J.S. Weissman, *Ribosome profiling reveals the what, when, where and how of protein synthesis.* Nature Reviews Molecular Cell Biology, 2015. **16**(11): p. 651-664.
154. Acharjee, A. *et al.*, *Integration of metabolomics, lipidomics and clinical data using a machine learning method.* BMC Bioinformatics, 2016. **17**(15): p. 440.
155. Coman, C. *et al.*, *Simultaneous Metabolite, Protein, Lipid Extraction (SIMPLEX): A Combinatorial Multimolecular Omics Approach for Systems Biology.* Molecular & Cellular Proteomics, 2016. **15**(4): p. 1453-1466.
156. Gingras, A.-C. and B. Raught, *Beyond hairballs: The use of quantitative mass spectrometry data to understand protein–protein interactions.* FEBS Letters, 2012. **586**(17): p. 2723-2731.
157. Wohlgemuth, I., C. Lenz, and H. Urlaub, *Studying macromolecular complex stoichiometries by peptide-based mass spectrometry.* Proteomics, 2015. **15**(5-6): p. 862-879.
158. Bauer, A. and B. Kuster, *Affinity purification-mass spectrometry.* European Journal of Biochemistry, 2003. **270**(4): p. 570-578.
159. Hein, Marco Y. *et al.*, *A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances.* Cell, 2015. **163**(3): p. 712-723.

160. Schopper, S. *et al.*, *Measuring protein structural changes on a proteome-wide scale using limited proteolysis-coupled mass spectrometry.* Nature Protocols, 2017. **12**: p. 2391.
161. Dearmond, P.D. *et al.*, *Discovery of novel cyclophilin A ligands using an H/D exchange- and mass spectrometry-based strategy.* Journal of Biomolecular Screening, 2010. **15**(9): p. 1051-1062.
162. Strickland, E.C. *et al.*, *Thermodynamic analysis of protein-ligand binding interactions in complex biological mixtures using the stability of proteins from rates of oxidation.* Nature Protocols, 2012. **8**: p. 148.
163. Ong, S.-E. *et al.*, *Identifying the proteins to which small-molecule probes and drugs bind in cells.* Proceedings of the National Academy of Sciences, 2009. **106**(12): p. 4617-4622.
164. Huber, K. *et al.*, *Approaching cellular resolution and reliable identification in mass spectrometry imaging of tryptic peptides.* Analytical and Bioanalytical Chemistry, 2018. **410**(23): p. 5825-5837.
165. Bandura, D.R. *et al.*, *Mass Cytometry: Technique for Real Time Single Cell Multitarget Immunoassay Based on Inductively Coupled Plasma Time-of-Flight Mass Spectrometry.* Analytical Chemistry, 2009. **81**(16): p. 6813-6822.
166. Itzhak, D.N. *et al.*, *A Mass Spectrometry-Based Approach for Mapping Protein Subcellular Localization Reveals the Spatial Proteome of Mouse Primary Neurons.* Cell Reports, 2017. **20**(11): p. 2706-2718.
167. Mulvey, C.M. *et al.*, *Using hyperLOPIT to perform high-resolution mapping of the spatial proteome.* Nature Protocols, 2017. **12**: p. 1110.
168. Sharon, M., *How Far Can We Go with Structural Mass Spectrometry of Protein Complexes?* Journal of the American Society for Mass Spectrometry, 2010. **21**(4): p. 487-500.
169. Painter, A.J. *et al.*, *Real-Time Monitoring of Protein Complexes Reveals their Quaternary Organization and Dynamics.* Chemistry & Biology, 2008. **15**(3): p. 246-253.
170. Skinner, O.S. *et al.*, *Top-down characterization of endogenous protein complexes with native proteomics.* Nature chemical biology, 2018. **14**(1): p. 36-41.
171. Hall, Z., A. Politis, and Carol V. Robinson, *Structural Modeling of Heteromeric Protein Complexes from Disassembly Pathways and Ion Mobility-Mass Spectrometry.* Structure, 2012. **20**(9): p. 1596-1609.
172. Kostyukevich, Y. *et al.*, *Hydrogen/deuterium exchange in mass spectrometry.* Mass Spectrometry Reviews, 2018. **37**(6): p. 811-853.
173. Leitner, A. *et al.*, *Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines.* Trends in Biochemical Sciences, 2016. **41**(1): p. 20-32.
174. Ishihama, Y. *et al.*, *Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein.* Molecular & Cellular Proteomics, 2005. **4**(9): p. 1265-1272.
175. Braisted, J.C. *et al.*, *The APEX Quantitative Proteomics Tool: generating protein quantitation estimates from LC-MS/MS proteomics results.* BMC Bioinformatics, 2008. **9**: p. 529-529.
176. Collins, M.O., L. Yu, and J.S. Choudhary, *Analysis of protein phosphorylation on a proteome-scale.* Proteomics, 2007. **7**(16): p. 2751-2768.
177. Li, X. *et al.*, *Elucidating Human Phosphatase-Substrate Networks.* Science Signaling, 2013. **6**(275): p. rs10-rs10.
178. Klaeger, S. *et al.*, *The target landscape of clinical kinase drugs.* Science, 2017. **358**(6367): p. eaan4368.
179. Zagorac, I. *et al.*, *In vivo phosphoproteomics reveals kinase activity profiles that predict treatment outcome in triple-negative breast cancer.* Nature Communications, 2018. **9**(1): p. 3501.
180. Fíla, J. and D. Honys, *Enrichment techniques employed in phosphoproteomics.* Amino Acids, 2012. **43**(3): p. 1025-1047.

181. Phillips, D.M., *The presence of acetyl groups of histones.* The Biochemical journal, 1963. **87**(2): p. 258-263.
182. Allfrey, V.G., R. Faulkner, and A.E. Mirsky, *ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS.* Proceedings of the National Academy of Sciences of the United States of America, 1964. **51**(5): p. 786-794.
183. Drazic, A. *et al.*, *The world of protein acetylation.* Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 2016. **1864**(10): p. 1372-1401.
184. Kouzarides, T., *Chromatin Modifications and Their Function.* Cell, 2007. **128**(4): p. 693-705.
185. Behnia, R. *et al.*, *Targeting of the Arf-like GTPase Arl3p to the Golgi requires N-terminal acetylation and the membrane protein Sys1p.* Nature Cell Biology, 2004. **6**: p. 405.
186. Setty, S.R.G. *et al.*, *Golgi targeting of ARF-like GTPase Arl3p requires its Nα-acetylation and the integral membrane protein Sys1p.* Nature Cell Biology, 2004. **6**: p. 414.
187. Behnia, R. *et al.*, *The yeast orthologue of GRASP65 forms a complex with a coiled-coil protein that contributes to ER to Golgi traffic.* The Journal of Cell Biology, 2007. **176**(3): p. 255-261.
188. Forte, G.M.A., M.R. Pool, and C.J. Stirling, *N-Terminal Acetylation Inhibits Protein Targeting to the Endoplasmic Reticulum.* PLOS Biology, 2011. **9**(5): p. e1001073.
189. Holmes, W.M. *et al.*, *Loss of amino-terminal acetylation suppresses a prion phenotype by modulating global protein folding.* Nature Communications, 2014. **5**: p. 4383.
190. Kuo, H.-P. *et al.*, *ARD1 Stabilization of TSC2 Suppresses Tumorigenesis Through the mTOR Signaling Pathway.* Science Signaling, 2010. **3**(108): p. ra9-ra9.
191. Zhang, X. *et al.*, *HDAC6 Modulates Cell Motility by Altering the Acetylation Level of Cortactin.* Molecular Cell, 2007. **27**(2): p. 197-213.
192. Biggar, K.K. and S.S.C. Li, *Non-histone protein methylation as a regulator of cellular signalling and function.* Nature Reviews Molecular Cell Biology, 2014. **16**: p. 5.
193. Murn, J. and Y. Shi, *The winding path of protein methylation research: milestones and new frontiers.* Nature Reviews Molecular Cell Biology, 2017. **18**: p. 517.
194. Van Damme, P. *et al.*, *A review of COFRADIC techniques targeting protein N-terminal acetylation.* BMC proceedings, 2009. **3 Suppl 6**(Suppl 6): p. S6-S6.
195. Kori, Y. *et al.*, *Proteome-wide acetylation dynamics in human cells.* Scientific Reports, 2017. **7**(1): p. 10296.
196. Carlson, S.M. *et al.*, *Proteome-wide enrichment of proteins modified by lysine methylation.* Nature Protocols, 2014. **9**(1): p. 37-50.
197. Lu, H., Y. Zhang, and P. Yang, *Advancements in mass spectrometry-based glycoproteomics and glycomics.* National Science Review, 2016. **3**(3): p. 345-364.
198. Kopitz, J., *Lipid glycosylation: a primer for histochemists and cell biologists.* Histochemistry and Cell Biology, 2017. **147**(2): p. 175-198.
199. Shental-Bechor, D. and Y. Levy, *Effect of glycosylation on protein folding: a close look at thermodynamic stabilization.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(24): p. 8256-8261.
200. Solá, R.J. and K. Griebenow, *Effects of glycosylation on the stability of protein pharmaceuticals.* Journal of Pharmaceutical Sciences, 2009. **98**(4): p. 1223-1245.
201. Lee, H.S., Y. Qi, and W. Im, *Effects of N-glycosylation on protein conformation and dynamics: Protein Data Bank analysis and molecular dynamics simulation study.* Scientific Reports, 2015. **5**: p. 8926-8926.
202. Ahmad, I. *et al.*, *Phosphorylation and glycosylation interplay: Protein modifications at hydroxy amino acids and prediction of signaling functions of the human β3 integrin family.* Journal of Cellular Biochemistry, 2006. **99**(3): p. 706-718.
203. Pinho, S.S. and C.A. Reis, *Glycosylation in cancer: mechanisms and clinical implications.* Nature Reviews Cancer, 2015. **15**: p. 540.

204. Schnaar, R.L., *Glycans and glycan-binding proteins in immune regulation: A concise introduction to glycobiology for the allergist.* The Journal of allergy and clinical immunology, 2015. **135**(3): p. 609-615.
205. Razaghi, A. *et al.*, *Improved therapeutic efficacy of mammalian expressed-recombinant interferon gamma against ovarian cancer cells.* Experimental Cell Research, 2017. **359**(1): p. 20-29.
206. Pan, S. *et al.*, *Mass spectrometry based glycoproteomics--from a proteomics perspective.* Molecular & Cellular Proteomics, 2011. **10**(1): p. R110.003251-R110.003251.
207. McDowell, G.S. and A. Philpott, *Non-canonical ubiquitylation: Mechanisms and consequences.* The International Journal of Biochemistry & Cell Biology, 2013. **45**(8): p. 1833-1842.
208. Kresge, N., R.D. Simoni, and R.L. Hill, *The Discovery of Ubiquitin-mediated Proteolysis by Aaron Ciechanover, Avram Hershko, and Irwin Rose.* Journal of Biological Chemistry, 2006. **281**(40): p. e32.
209. Jin, L. *et al.*, *Mechanism of Ubiquitin-Chain Formation by the Human Anaphase-Promoting Complex.* Cell, 2008. **133**(4): p. 653-665.
210. Swatek, K.N. and D. Komander, *Ubiquitin modifications.* Cell Research, 2016. **26**: p. 399.
211. Xu, G. and S.R. Jaffrey, *Proteomic identification of protein ubiquitination events.* Biotechnology and Genetic Engineering Reviews, 2013. **29**(1): p. 73-109.
212. Esteban Warren, M.R. *et al.*, *Electrospray ionization tandem mass spectrometry of model peptides reveals diagnostic fragment ions for protein ubiquitination.* Rapid Communications in Mass Spectrometry, 2005. **19**(4): p. 429-437.
213. Denis, N.J. *et al.*, *Tryptic digestion of ubiquitin standards reveals an improved strategy for identifying ubiquitinated proteins by mass spectrometry.* Proteomics, 2007. **7**(6): p. 868-874.
214. Xu, G., J.S. Paige, and S.R. Jaffrey, *Global analysis of lysine ubiquitination by ubiquitin remnant immunoaffinity profiling.* Nature Biotechnology, 2010. **28**: p. 868.
215. Danielsen, J.M.R. *et al.*, *Mass Spectrometric Analysis of Lysine Ubiquitylation Reveals Promiscuity at Site Level.* Molecular & Cellular Proteomics, 2011. **10**(3): p. M110.003590.
216. Kim, W. *et al.*, *Systematic and Quantitative Assessment of the Ubiquitin-Modified Proteome.* Molecular Cell, 2011. **44**(2): p. 325-340.
217. Stes, E. *et al.*, *A COFRADIC Protocol To Study Protein Ubiquitination.* Journal of Proteome Research, 2014. **13**(6): p. 3107-3113.
218. Impens, F. *et al.*, *Mapping of SUMO sites and analysis of SUMOylation changes induced by external stimuli.* Proceedings of the National Academy of Sciences, 2014. **111**(34): p. 12432-12437.
219. Hendriks, I.A. and A.C.O. Vertegaal, *A comprehensive compilation of SUMO proteomics.* Nature Reviews Molecular Cell Biology, 2016. **17**: p. 581.
220. Jones, J. *et al.*, *A targeted proteomic analysis of the ubiquitin-like modifier nedd8 and associated proteins.* Journal of Proteome Research, 2008. **7**(3): p. 1274-1287.
221. Maghames, C.M. *et al.*, *NEDDylation promotes nuclear protein aggregation and protects the Ubiquitin Proteasome System upon proteotoxic stress.* Nature Communications, 2018. **9**(1): p. 4376.
222. Giannakopoulos, N.V. *et al.*, *Proteomic identification of proteins conjugated to ISG15 in mouse and human cells.* Biochemical and Biophysical Research Communications, 2005. **336**(2): p. 496-506.
223. Radoshevich, L. *et al.*, *ISG15 counteracts Listeria monocytogenes infection.* eLife, 2015. **4**: p. e06848.
224. Anderle, M. *et al.*, *Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum.* Bioinformatics, 2004. **20**(18): p. 3575-3582.

225. Marginean, I. *et al.*, *Analytical characterization of the electrospray ion source in the nanoflow regime.* Analytical Chemistry, 2008. **80**(17): p. 6573-6579.
226. Li, Z. *et al.*, *Systematic Comparison of Label-Free, Metabolic Labeling, and Isobaric Chemical Labeling for Quantitative Proteomics on LTQ Orbitrap Velos.* Journal of Proteome Research, 2012. **11**(3): p. 1582-1590.
227. Thompson, A. *et al.*, *Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS.* Analytical Chemistry, 2003. **75**(8): p. 1895-1904.
228. Bantscheff, M. *et al.*, *Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present.* Analytical and Bioanalytical Chemistry, 2012. **404**(4): p. 939-965.
229. Pipkin, J.L. *et al.*, *Analysis of protein incorporation of radioactive isotopes in the chinese hamster ovary cell cycle by electronic sorting and gel microelectrophoresis.* Cytometry, 1986. **7**(2): p. 147-156.
230. Oda, Y. *et al.*, *Accurate quantitation of protein expression and site-specific phosphorylation.* Proceedings of the National Academy of Sciences, 1999. **96**(12): p. 6591-6596.
231. Farquhar, G.D., J.R. Ehleringer, and K.T. Hubick, *Carbon Isotope Discrimination and Photosynthesis.* Annual Review of Plant Physiology and Plant Molecular Biology, 1989. **40**(1): p. 503-537.
232. Conen, F. and A. Neftel, *Do increasingly depleted δ15N values of atmospheric N2O indicate a decline in soil N2O reduction?* Biogeochemistry, 2007. **82**(3): p. 321-326.
233. Zubarev, R.A., *Role of Stable Isotopes in Life—Testing Isotopic Resonance Hypothesis.* Genomics, Proteomics & Bioinformatics, 2011. **9**(1): p. 15-20.
234. Li, X. and M.P. Snyder, *Can heavy isotopes increase lifespan? Studies of relative abundance in various organisms reveal chemical perspectives on aging.* BioEssays : news and reviews in molecular, cellular and developmental biology, 2016. **38**(11): p. 1093-1101.
235. Conrads, T.P. *et al.*, *Quantitative Analysis of Bacterial and Mammalian Proteomes Using a Combination of Cysteine Affinity Tags and ^{15}N-Metabolic Labeling.* Analytical Chemistry, 2001. **73**(9): p. 2132-2139.
236. Wu, C.C. *et al.*, *Metabolic Labeling of Mammalian Organisms with Stable Isotopes for Quantitative Proteomic Analysis.* Analytical Chemistry, 2004. **76**(17): p. 4951-4959.
237. McClatchy, D.B. *et al.*, *^{15}N metabolic labeling of mammalian tissue with slow protein turnover.* Journal of Proteome Research, 2007. **6**(5): p. 2005-2010.
238. Krijgsveld, J. *et al.*, *Metabolic labeling of C. elegans and D. melanogaster for quantitative proteomics.* Nature Biotechnology, 2003. **21**: p. 927.
239. Sonzogni, A. National Nuclear Data Center, Brookhaven National Laboratory. Acessed on: 28-11-2018. Available from: https://www.nndc.bnl.gov/chart/.
240. Beynon, R.J. and J.M. Pratt, *Metabolic Labeling of Proteins for Proteomics.* Molecular & Cellular Proteomics, 2005. **4**(7): p. 857-872.
241. Ong, S.-E. *et al.*, *Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics.* Molecular & Cellular Proteomics, 2002. **1**(5): p. 376-386.
242. Ong, S.-E. and M. Mann, *A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC).* Nature Protocols, 2007. **1**(6): p. 2650-2660.
243. Blagoev, B. *et al.*, *Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics.* Nature Biotechnology, 2004. **22**: p. 1139.
244. Molina, H. *et al.*, *Temporal profiling of the adipocyte proteome during differentiation using a five-plex SILAC based strategy.* Journal of Proteome Research, 2009. **8**(1): p. 48-58.
245. Geiger, T. *et al.*, *Super-SILAC mix for quantitative proteomics of human tumor tissue.* Nature Methods, 2010. **7**: p. 383.
246. Shenoy, A. and T. Geiger, *Super-SILAC: current trends and future perspectives.* Expert Review of Proteomics, 2015. **12**(1): p. 13-19.

247. Larance, M. *et al.*, *Stable-isotope labeling with amino acids in nematodes.* Nature Methods, 2011. **8**(10): p. 849-851.
248. Sury, M.D., J.-X. Chen, and M. Selbach, *The SILAC fly allows for accurate protein quantification in vivo.* Molecular & Cellular Proteomics, 2010. **9**(10): p. 2173-2183.
249. Krüger, M. *et al.*, *SILAC Mouse for Quantitative Proteomics Uncovers Kindlin-3 as an Essential Factor for Red Blood Cell Function.* Cell, 2008. **134**(2): p. 353-364.
250. Lewandowska, D. *et al.*, *Plant SILAC: Stable-Isotope Labelling with Amino Acids of Arabidopsis Seedlings for Quantitative Proteomics.* PLoS ONE, 2013. **8**(8): p. e72207.
251. Geiger, T. *et al.*, *Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics.* Nature Protocols, 2011. **6**: p. 147.
252. Scheerlinck, E. *et al.*, *Assessing the impact of minimizing arginine conversion in fully defined SILAC culture medium in human embryonic stem cells.* Proteomics, 2016. **16**(20): p. 2605-2614.
253. Bendall, S.C. *et al.*, *Prevention of amino acid conversion in SILAC experiments with embryonic stem cells.* Molecular & Cellular Proteomics, 2008. **7**(9): p. 1587-1597.
254. Lößner, C. *et al.*, *Preventing arginine-to-proline conversion in a cell-line-independent manner during cell cultivation under stable isotope labeling by amino acids in cell culture (SILAC) conditions.* Analytical Biochemistry, 2011. **412**(1): p. 123-125.
255. Blagoev, B. and M. Mann, *Quantitative proteomics to study mitogen-activated protein kinases.* Methods, 2006. **40**(3): p. 243-250.
256. Ong, S.-E. and M. Mann, *Mass spectrometry–based proteomics turns quantitative.* Nature Chemical Biology, 2005. **1**: p. 252.
257. Tebbe, A. *et al.*, *Systematic evaluation of label-free and super-SILAC quantification for proteome expression analysis.* Rapid Communications in Mass Spectrometry, 2015. **29**(9): p. 795-801.
258. Liu, N.Q. *et al.*, *Quantitative Proteomic Analysis of Microdissected Breast Cancer Tissues: Comparison of Label-Free and SILAC-based Quantification with Shotgun, Directed, and Targeted MS Approaches.* Journal of Proteome Research, 2013. **12**(10): p. 4627-4641.
259. Merl, J. *et al.*, *Direct comparison of MS-based label-free and SILAC quantitative proteome profiling strategies in primary retinal Müller cells.* Proteomics, 2012. **12**(12): p. 1902-1911.
260. Houel, S. *et al.*, *Quantifying the Impact of Chimera MS/MS Spectra on Peptide Identification in Large-Scale Proteomics Studies.* Journal of Proteome Research, 2010. **9**(8): p. 4152-4160.
261. Overmyer, K.A. *et al.*, *Multiplexed proteome analysis with neutron-encoded stable isotope labeling in cells and mice.* Nature Protocols, 2018. **13**(1): p. 293-306.
262. Turck, C.W. *et al.*, *The Association of Biomolecular Resource Facilities Proteomics Research Group 2006 Study.* Relative Protein Quantitation, 2007. **6**(8): p. 1291-1298.
263. Bantscheff, M. *et al.*, *Quantitative mass spectrometry in proteomics: a critical review.* Analytical and Bioanalytical Chemistry, 2007. **389**(4): p. 1017-31.
264. Asara, J.M. *et al.*, *A label-free quantification method by MS/MS TIC compared to SILAC and spectral counting in a proteomics screen.* Proteomics, 2008. **8**(5): p. 994-999.
265. Altelaar, A.F.M. *et al.*, *Benchmarking stable isotope labeling based quantitative proteomics.* Journal of Proteomics, 2013. **88**: p. 14-26.
266. Hebert, A.S. *et al.*, *Neutron-encoded mass signatures for multiplexed proteome quantification.* Nature Methods, 2013. **10**(4): p. 332-334.
267. Hsu, J.-L. and S.-H. Chen, *Stable isotope dimethyl labelling for quantitative proteomics and beyond.* Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 2016. **374**(2079): p. 20150364.
268. Boersema, P.J. *et al.*, *Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics.* Nature Protocols, 2009. **4**: p. 484.
269. Hsu, J.L. *et al.*, *Beyond quantitative proteomics: signal enhancement of the a1 ion as a mass tag for peptide sequencing using dimethyl labeling.* Journal of Proteome Research, 2005. **4**(1): p. 101-8.

270. Hsu, J.-L. *et al.*, *Stable-Isotope Dimethyl Labeling for Quantitative Proteomics.* Analytical Chemistry, 2003. **75**(24): p. 6843-6852.
271. Boersema, P.J. *et al.*, *Triplex protein quantification based on stable isotope labeling by peptide dimethylation applied to cell and tissue lysates.* Proteomics, 2008. **8**(22): p. 4624-4632.
272. Hsu, J.-L., S.-Y. Huang, and S.-H. Chen, *Dimethyl multiplexed labeling combined with microcolumn separation and MS analysis for time course study in proteomics.* Electrophoresis, 2006. **27**(18): p. 3652-3660.
273. Wu, Y. *et al.*, *Five-plex isotope dimethyl labeling for quantitative proteomics.* Chemical Communications, 2014. **50**(14): p. 1708-1710.
274. Wang, F. *et al.*, *A six-plex proteome quantification strategy reveals the dynamics of protein turnover.* Scientific Reports, 2013. **3**: p. 1827-1827.
275. Lau, H.-T. *et al.*, *Comparing SILAC- and stable isotope dimethyl-labeling approaches for quantitative proteomics.* Journal of Proteome Research, 2014. **13**(9): p. 4164-4174.
276. Turowski, M. *et al.*, *Deuterium Isotope Effects on Hydrophobic Interactions: The Importance of Dispersion Interactions in the Hydrophobic Phase.* Journal of the American Chemical Society, 2003. **125**(45): p. 13836-13849.
277. Zhang, R. *et al.*, *Fractionation of Isotopically Labeled Peptides in Quantitative Proteomics.* Analytical Chemistry, 2001. **73**(21): p. 5142-5149.
278. Zhang, R. *et al.*, *Controlling Deuterium Isotope Effects in Comparative Proteomics.* Analytical Chemistry, 2002. **74**(15): p. 3662-3669.
279. Zhang, R. and F.E. Regnier, *Minimizing Resolution of Isotopically Coded Peptides in Comparative Proteomics.* Journal of Proteome Research, 2002. **1**(2): p. 139-147.
280. Schnölzer, M., P. Jedrzejewski, and W.D. Lehmann, *Protease-catalyzed incorporation of 18O into peptide fragments and its application for protein sequencing by electrospray and matrix-assisted laser desorption/ionization mass spectrometry.* ELECTROPHORESIS, 1996. **17**(5): p. 945-953.
281. Shevchenko, A. *et al.*, *Rapid 'de novo' peptide sequencing by a combination of nanoelectrospray, isotopic labeling and a quadrupole/time-of-flight mass spectrometer.* Rapid Communications in Mass Spectrometry, 1997. **11**(9): p. 1015-1024.
282. Uttenweiler-Joseph, S. *et al.*, *Automated de novo sequencing of proteins using the differential scanning technique.* Proteomics, 2001. **1**(5): p. 668-682.
283. Mirgorodskaya, O.A. *et al.*, *Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using 18O-labeled internal standards.* Rapid Communications in Mass Spectrometry, 2000. **14**(14): p. 1226-1232.
284. Yao, X. *et al.*, *Proteolytic 18O Labeling for Comparative Proteomics: Model Studies with Two Serotypes of Adenovirus.* Analytical Chemistry, 2001. **73**(13): p. 2836-2842.
285. Larsen, M.R. *et al.*, *Characterization of differently processed forms of enolase 2 from Saccharomyces cerevisiae by two-dimensional gel electrophoresis and mass spectrometry.* ELECTROPHORESIS, 2001. **22**(3): p. 566-575.
286. Stewart, II, T. Thomson, and D. Figeys, *18O labeling: a tool for proteomics.* Rapid Communications in Mass Spectrometry, 2001. **15**(24): p. 2456-65.
287. Ye, X. *et al.*, *^{18}O stable isotope labeling in MS-based proteomics.* Briefings in Functional Genomics & Proteomics, 2009. **8**(2): p. 136-144.
288. Ross, P.L. *et al.*, *Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents.* Molecular & Cellular Proteomics, 2004. **3**(12): p. 1154-1169.
289. Sonnett, M., E. Yeung, and M. Wühr, *Accurate, Sensitive, and Precise Multiplexed Proteomics Using the Complement Reporter Ion Cluster.* Analytical Chemistry, 2018. **90**(8): p. 5032-5039.
290. Dayon, L. *et al.*, *Relative Quantification of Proteins in Human Cerebrospinal Fluids by MS/MS Using 6-Plex Isobaric Tags.* Analytical Chemistry, 2008. **80**(8): p. 2921-2931.
291. Pierce, A. *et al.*, *Eight-channel iTRAQ Enables Comparison of the Activity of Six Leukemogenic Tyrosine Kinases.* Molecular & Cellular Proteomics, 2008. **7**(5): p. 853-863.

292. Werner, T. *et al.*, *Ion Coalescence of Neutron Encoded TMT 10-Plex Reporter Ions.* Analytical Chemistry, 2014. **86**(7): p. 3594-3601.
293. Stepanova, E., S.P. Gygi, and J.A. Paulo, *Filter-Based Protein Digestion (FPD): A Detergent-Free and Scaffold-Based Strategy for TMT Workflows.* Journal of Proteome Research, 2018. **17**(3): p. 1227-1234.
294. Zhang, J., Y. Wang, and S. Li, *Deuterium Isobaric Amine-Reactive Tags for Quantitative Proteomics.* Analytical Chemistry, 2010. **82**(18): p. 7588-7595.
295. Xiang, F. *et al.*, *N,N-Dimethyl Leucines as Novel Isobaric Tandem Mass Tags for Quantitative Proteomics and Peptidomics.* Analytical Chemistry, 2010. **82**(7): p. 2817-2825.
296. Frost, D.C., T. Greer, and L. Li, *High-Resolution Enabled 12-Plex DiLeu Isobaric Tags for Quantitative Proteomics.* Analytical Chemistry, 2015. **87**(3): p. 1646-1654.
297. Savitski, M.M. *et al.*, *Multiplexed Proteome Dynamics Profiling Reveals Mechanisms Controlling Protein Homeostasis.* Cell, 2018. **173**(1): p. 260-274.e25.
298. Ow, S.Y. *et al.*, *iTRAQ Underestimation in Simple and Complex Mixtures: "The Good, the Bad and the Ugly".* Journal of Proteome Research, 2009. **8**(11): p. 5347-5355.
299. Karp, N.A. *et al.*, *Addressing Accuracy and Precision Issues in iTRAQ Quantitation.* Molecular & Cellular Proteomics, 2010. **9**(9): p. 1885-1897.
300. Hogrebe, A. *et al.*, *Benchmarking common quantification strategies for large-scale phosphoproteomics.* Nature Communications, 2018. **9**(1): p. 1045.
301. Wenger, C.D. *et al.*, *Gas-phase purification enables accurate, multiplexed proteome quantification with isobaric tagging.* Nature Methods, 2011. **8**(11): p. 933-935.
302. Sturm, R.M., C.B. Lietz, and L. Li, *Improved isobaric tandem mass tag quantification by ion mobility mass spectrometry.* Rapid Communications in Mass Spectrometry, 2014. **28**(9): p. 1051-1060.
303. Ting, L. *et al.*, *MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics.* Nature Methods, 2011. **8**(11): p. 937-940.
304. McAlister, G.C. *et al.*, *MultiNotch MS3 Enables Accurate, Sensitive, and Multiplexed Detection of Differential Expression across Cancer Cell Line Proteomes.* Analytical Chemistry, 2014. **86**(14): p. 7150-7158.
305. Wühr, M. *et al.*, *Accurate Multiplexed Proteomics at the MS2 Level Using the Complement Reporter Ion Cluster.* Analytical Chemistry, 2012. **84**(21): p. 9214-9221.
306. Virreira Winter, S. *et al.*, *EASI-tag enables accurate multiplexed and interference-free MS2-based proteome quantification.* Nature Methods, 2018. **15**(7): p. 527-530.
307. Tu, C. *et al.*, *Systematic Assessment of Survey Scan and MS2-Based Abundance Strategies for Label-Free Quantitative Proteomics Using High-Resolution MS Data.* Journal of Proteome Research, 2014. **13**(4): p. 2069-2079.
308. Patel, V.J. *et al.*, *A Comparison of Labeling and Label-Free Mass Spectrometry-Based Proteomics Approaches.* Journal of Proteome Research, 2009. **8**(7): p. 3752-3759.
309. Latosinska, A. *et al.*, *Comparative Analysis of Label-Free and 8-Plex iTRAQ Approach for Quantitative Tissue Proteomic Analysis.* PLoS One, 2015. **10**(9): p. e0137048.
310. Piehowski, P.D. *et al.*, *Sources of Technical Variability in Quantitative LC–MS Proteomics: Human Brain Tissue Sample Analysis.* Journal of Proteome Research, 2013. **12**(5): p. 2128-2137.
311. Rodriguez, J. *et al.*, *Does Trypsin Cut Before Proline?* Journal of Proteome Research, 2008. **7**(1): p. 300-305.
312. Burkhart, J.M. *et al.*, *Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics.* Journal of Proteomics, 2012. **75**(4): p. 1454-1462.
313. Clemmer, D.E. *et al.*, *Fast and accurate identification of semi-tryptic peptides in shotgun proteomics.* Bioinformatics, 2007. **24**(1): p. 102-109.
314. Lowenthal, M.S. *et al.*, *Quantitative Bottom-Up Proteomics Depends on Digestion Conditions.* Analytical Chemistry, 2014. **86**(1): p. 551-558.
315. Moruz, L. and L. Käll, *Peptide retention time prediction.* Mass Spectrometry Reviews, 2017. **36**(5): p. 615-623.

316. Moruz, L. *et al.*, *Chromatographic retention time prediction for posttranslationally modified peptides.* Proteomics, 2012. **12**(8): p. 1151-1159.
317. Cox, J. *et al.*, *Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ.* Molecular & Cellular Proteomics, 2014. **13**(9): p. 2513-2526.
318. Lai, X. *et al.*, *A Novel Alignment Method and Multiple Filters for Exclusion of Unqualified Peptides To Enhance Label-Free Quantification Using Peptide Intensity in LC–MS/MS.* Journal of Proteome Research, 2011. **10**(10): p. 4799-4812.
319. Staes, A. *et al.*, *Asn3, a Reliable, Robust, and Universal Lock Mass for Improved Accuracy in LC–MS and LC–MS/MS.* Analytical Chemistry, 2013. **85**(22): p. 11054-11060.
320. Holman, S.W., L. McLean, and C.E. Eyers, *RePLiCal: A QconCAT Protein for Retention Time Standardization in Proteomics Studies.* Journal of Proteome Research, 2016. **15**(3): p. 1090-1102.
321. Krokhin, O.V. and V. Spicer, *Predicting Peptide Retention Times for Proteomics.* Current Protocols in Bioinformatics, 2010. **31**(1): p. 13.14.1-13.14.15.
322. Moruz, L., D. Tomazela, and L. Käll, *Training, Selection, and Robust Calibration of Retention Time Models for Targeted Proteomics.* Journal of Proteome Research, 2010. **9**(10): p. 5209-5216.
323. Maboudi Afkham, H. *et al.*, *Uncertainty estimation of predictions of peptides' chromatographic retention times in shotgun proteomics.* Bioinformatics, 2017. **33**(4): p. 508-513.
324. Mitulović, G. *et al.*, *Preventing Carryover of Peptides and Proteins in Nano LC-MS Separations.* Analytical Chemistry, 2009. **81**(14): p. 5955-5960.
325. Hsieh, E.J. *et al.*, *Effects of Column and Gradient Lengths on Peak Capacity and Peptide Identification in Nanoflow LC-MS/MS of Complex Proteomic Samples.* Journal of The American Society for Mass Spectrometry, 2013. **24**(1): p. 148-153.
326. Young, C., A.V. Podtelejnikov, and M.L. Nielsen, *Improved Reversed Phase Chromatography of Hydrophilic Peptides from Spatial and Temporal Changes in Column Temperature.* Journal of Proteome Research, 2017. **16**(6): p. 2307-2317.
327. Antignac, J.-P. *et al.*, *The ion suppression phenomenon in liquid chromatography–mass spectrometry and its consequences in the field of residue analysis.* Analytica Chimica Acta, 2005. **529**(1): p. 129-136.
328. Lu, W. *et al.*, *Response of peptide intensity to concentration in ESI-MS-based proteome.* Science China Chemistry, 2014. **57**(5): p. 686-694.
329. Nilsson, L.B. and P. Skansen, *Investigation of absolute and relative response for three different liquid chromatography/tandem mass spectrometry systems; the impact of ionization and detection saturation.* Rapid Communications in Mass Spectrometry, 2012. **26**(12): p. 1399-1406.
330. Jarnuczak, A.F. *et al.*, *Analysis of Intrinsic Peptide Detectability via Integrated Label-Free and SRM-Based Absolute Quantitative Proteomics.* Journal of Proteome Research, 2016. **15**(9): p. 2945-2959.
331. Liu, H. *et al.*, *The Prediction of Peptide Charge States for Electrospray Ionization in Mass Spectrometry.* Procedia Environmental Sciences, 2011. **8**: p. 483-491.
332. Morand, K., G. Talbo, and M. Mann, *Oxidation of peptides during electrospray ionization.* Rapid Communications in Mass Spectrometry, 1993. **7**(8): p. 738-743.
333. Godugu, B. *et al.*, *Effect of N-Terminal Glutamic Acid and Glutamine on Fragmentation of Peptide Ions.* Journal of the American Society for Mass Spectrometry, 2010. **21**(7): p. 1169-1176.
334. Gorshkov, V., T. Verano-Braga, and F. Kjeldsen, *SuperQuant: A Data Processing Approach to Increase Quantitative Proteome Coverage.* Analytical Chemistry, 2015. **87**(12): p. 6319-6327.
335. Plubell, D.L. *et al.*, *Extended Multiplexing of Tandem Mass Tags (TMT) Labeling Reveals Age and High Fat Diet Specific Proteome Changes in Mouse Epididymal Adipose Tissue.* Molecular & Cellular Proteomics, 2017. **16**(5): p. 873-890.

336. Wang, H., S. Alvarez, and L.M. Hicks, *Comprehensive Comparison of iTRAQ and Label-free LC-Based Quantitative Proteomics Approaches Using Two Chlamydomonas reinhardtii Strains of Interest for Biofuels Engineering.* Journal of Proteome Research, 2012. **11**(1): p. 487-501.
337. Boschetti, E. *et al.*, *Romancing the "hidden proteome", Anno Domini two zero zero seven.* Journal of Chromatography A, 2007. **1153**(1): p. 277-290.
338. Rabilloud, T., *Membrane proteins and proteomics: Love is possible, but so difficult.* ELECTROPHORESIS, 2009. **30**(S1): p. S174-S180.
339. Perdigão, N. *et al.*, *Unexpected features of the dark proteome.* Proceedings of the National Academy of Sciences of the United States of America, 2015. **112**(52): p. 15898-15903.
340. Perdigão, N., A.C. Rosa, and S.I. O'Donoghue, *The Dark Proteome Database.* BioData Mining, 2017. **10**: p. 24-24.
341. Doerr, A., *Mass spectrometry–based targeted proteomics.* Nature Methods, 2012. **10**: p. 23.
342. Lange, V. *et al.*, *Selected reaction monitoring for quantitative proteomics: a tutorial.* Molecular Systems Biology, 2008. **4**: p. 222-222.
343. de Graaf, E.L. *et al.*, *Improving SRM Assay Development: A Global Comparison between Triple Quadrupole, Ion Trap, and Higher Energy CID Peptide Fragmentation Spectra.* Journal of Proteome Research, 2011. **10**(9): p. 4334-4341.
344. Peterson, A.C. *et al.*, *Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics.* Molecular & Cellular Proteomics, 2012. **11**(11): p. 1475-1488.
345. Ronsein, G.E. *et al.*, *Parallel reaction monitoring (PRM) and selected reaction monitoring (SRM) exhibit comparable linearity, dynamic range and precision for targeted quantitative HDL proteomics.* Journal of Proteomics, 2015. **113**: p. 388-399.
346. Purvine, S. *et al.*, *Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer.* Proteomics, 2003. **3**(6): p. 847-850.
347. Plumb, R.S. *et al.*, *UPLC/MSE; a new approach for generating molecular fragment information for biomarker structure elucidation.* Rapid Communications in Mass Spectrometry, 2006. **20**(13): p. 1989-1994.
348. Venable, J.D. *et al.*, *Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra.* Nature Methods, 2004. **1**(1): p. 39-45.
349. Panchaud, A. *et al.*, *Faster, quantitative, and accurate precursor acquisition independent from ion count.* Analytical Chemistry, 2011. **83**(6): p. 2250-2257.
350. Bern, M. *et al.*, *Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry.* Analytical Chemistry, 2010. **82**(3): p. 833-841.
351. Gillet, L.C. *et al.*, *Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis.* Molecular & Cellular Proteomics, 2012. **11**(6).
352. Vowinckel, J. *et al.*, *The beauty of being (label)-free: sample preparation methods for SWATH-MS and next-generation targeted proteomics.* F1000Research, 2014. **2**: p. 272-272.
353. Röst, H.L. *et al.*, *OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data.* Nature Biotechnology, 2014. **32**: p. 219.
354. Kelstrup, C.D. *et al.*, *Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics.* Journal of Proteome Research, 2018. **17**(1): p. 727-738.
355. Bruderer, R. *et al.*, *Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues.* Molecular & Cellular Proteomics, 2015. **14**(5): p. 1400-1410.
356. Collins, B.C. *et al.*, *Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry.* Nature Communications, 2017. **8**(1): p. 291.

357. Liu, Y. *et al.*, *Quantitative measurements of N-linked glycoproteins in human plasma by SWATH-MS.* Proteomics, 2013. **13**(8): p. 1247-1256.
358. Schmidlin, T. *et al.*, *Assessment of SRM, MRM3, and DIA for the targeted analysis of phosphorylation dynamics in non-small cell lung cancer.* Proteomics, 2016. **16**(15-16): p. 2193-2205.
359. Ludwig, C. *et al.*, *Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial.* Molecular Systems Biology, 2018. **14**(8): p. e8126.
360. Keller, A. *et al.*, *Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search.* Analytical Chemistry, 2002. **74**(20): p. 5383-5392.
361. Eng, J.K., A.L. McCormack, and J.R. Yates, *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, 1994. **5**(11): p. 976-89.
362. Huala, E. *et al.*, *The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant.* Nucleic Acids Research, 2001. **29**(1): p. 102-105.
363. Frank, A.M., *A ranking-based scoring function for peptide-spectrum matches.* Journal of Proteome Research, 2009. **8**(5): p. 2241-2252.
364. Shortreed, M.R. *et al.*, *Global Identification of Protein Post-translational Modifications in a Single-Pass Database Search.* Journal of Proteome Research, 2015. **14**(11): p. 4714-4720.
365. David, M. *et al.*, *SpecOMS: A Full Open Modification Search Method Performing All-to-All Spectra Comparisons within Minutes.* Journal of Proteome Research, 2017. **16**(8): p. 3030-3038.
366. Craig, R. and R.C. Beavis, *A method for reducing the time required to match protein sequences with tandem mass spectra.* Rapid Communications in Mass Spectrometry, 2003. **17**(20): p. 2310-2316.
367. Bogdanow, B., H. Zauber, and M. Selbach, *Systematic Errors in Peptide and Protein Identification and Quantification by Modified Peptides.* Molecular & cellular proteomics, 2016. **15**(8): p. 2791-2801.
368. Perkins, D.N. *et al.*, *Probability-based protein identification by searching sequence databases using mass spectrometry data.* Electrophoresis, 1999. **20**(18): p. 3551-67.
369. Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.* Nature Methods, 2007. **4**: p. 207.
370. Käll, L. *et al.*, *Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases.* Journal of Proteome Research, 2008. **7**(1): p. 29-34.
371. Jeong, K., S. Kim, and N. Bandeira, *False discovery rates in spectral identification.* BMC Bioinformatics, 2012. **13 Suppl 16**(Suppl 16): p. S2-S2.
372. Diament, B.J. and W.S. Noble, *Faster SEQUEST Searching for Peptide Identification from Tandem Mass Spectra.* Journal of Proteome Research, 2011. **10**(9): p. 3871-3879.
373. Fenyö, D. and R.C. Beavis, *A Method for Assessing the Statistical Significance of Mass Spectrometry-Based Protein Identifications Using General Scoring Schemes.* Analytical Chemistry, 2003. **75**(4): p. 768-774.
374. Kim, S. and P.A. Pevzner, *MS-GF+ makes progress towards a universal database search tool for proteomics.* Nature Communications, 2014. **5**: p. 5277.
375. Dorfer, V. *et al.*, *MS Amanda, a Universal Identification Algorithm Optimized for High Accuracy Tandem Mass Spectra.* Journal of Proteome Research, 2014. **13**(8): p. 3679-3684.
376. Tabb, D.L., C.G. Fernando, and M.C. Chambers, *MyriMatch: Highly Accurate Tandem Mass Spectral Peptide Identification by Multivariate Hypergeometric Analysis.* Journal of Proteome Research, 2007. **6**(2): p. 654-661.

377. Eng, J.K., T.A. Jahan, and M.R. Hoopmann, *Comet: An open-source MS/MS sequence database search tool.* Proteomics, 2013. **13**(1): p. 22-24.
378. Cox, J. *et al.*, *Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment.* Journal of Proteome Research, 2011. **10**(4): p. 1794-1805.
379. Geer, L.Y. *et al.*, *Open Mass Spectrometry Search Algorithm.* Journal of Proteome Research, 2004. **3**(5): p. 958-964.
380. Ma, B., *Novor: Real-Time Peptide de Novo Sequencing Software.* Journal of The American Society for Mass Spectrometry, 2015. **26**(11): p. 1885-1894.
381. Tabb, D.L. *et al.*, *DirecTag: Accurate Sequence Tags from Peptide MS/MS through Statistical Scoring.* Journal of Proteome Research, 2008. **7**(9): p. 3838-3846.
382. Vaudel, M. *et al.*, *SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches.* Proteomics, 2011. **11**(5): p. 996-999.
383. Vaudel, M. *et al.*, *PeptideShaker enables reanalysis of MS-derived proteomics data sets.* Nature Biotechnology, 2015. **33**(1): p. 22-24.
384. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nature Biotechnology, 2008. **26**(12): p. 1367-1372.
385. Nesvizhskii, A.I. and R. Aebersold, *Interpretation of Shotgun Proteomic Data.* The Protein Inference Problem, 2005. **4**(10): p. 1419-1440.
386. Nesvizhskii, A.I. *et al.*, *A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry.* Analytical Chemistry, 2003. **75**(17): p. 4646-4658.
387. Shen, C. *et al.*, *A hierarchical statistical model to assess the confidence of peptides and proteins inferred from tandem mass spectrometry.* Bioinformatics, 2008. **24**(2): p. 202-208.
388. Gerster, S. *et al.*, *Protein and gene model inference based on statistical modeling in k-partite graphs.* Proceedings of the National Academy of Sciences, 2010. **107**(27): p. 12101-12106.
389. Higdon, R. and E. Kolker, *A predictive model for identifying proteins by a single peptide match.* Bioinformatics, 2007. **23**(3): p. 277-280.
390. Gupta, N. and P.A. Pevzner, *False Discovery Rates of Protein Identifications: A Strike against the Two-Peptide Rule.* Journal of Proteome Research, 2009. **8**(9): p. 4173-4181.
391. Reiter, L. *et al.*, *Protein Identification False Discovery Rates for Very Large Proteomics Data Sets Generated by Tandem Mass Spectrometry.* Molecular & Cellular Proteomics, 2009. **8**(11): p. 2405-2417.
392. Granholm, V. *et al.*, *Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics.* Journal of Proteomics, 2013. **80**: p. 123-131.
393. Adamski, M. *et al.*, *Data management and preliminary data analysis in the pilot phase of the HUPO Plasma Proteome Project.* Proteomics, 2005. **5**(13): p. 3246-3261.
394. Li, Y.F. *et al.*, *A bayesian approach to protein inference problem in shotgun proteomics.* Journal of computational biology, 2009. **16**(8): p. 1183-1193.
395. Serang, O., M.J. MacCoss, and W.S. Noble, *Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data.* Journal of Proteome Research, 2010. **9**(10): p. 5346-5357.
396. Savitski, M.M. *et al.*, *A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets.* Molecular & cellular proteomics, 2015. **14**(9): p. 2394-2404.
397. The, M., A. Tasnim, and L. Käll, *How to talk about protein-level false discovery rates in shotgun proteomics.* Proteomics, 2016. **16**(18): p. 2461-2469.
398. Huang, T. *et al.*, *Protein inference: a review.* Briefings in Bioinformatics, 2012. **13**(5): p. 586-614.
399. Serang, O. and W. Noble, *A review of statistical methods for protein identification using tandem mass spectrometry.* Statistics and Its Interface, 2012. **5**(1): p. 3-20.

400. Tsou, C.-C. *et al.*, *IDEAL-Q, an automated tool for label-free quantitation analysis using an efficient peptide alignment approach and spectral data validation.* Molecular & Cellular Proteomics, 2010. **9**(1): p. 131-144.
401. Valot, B. *et al.*, *MassChroQ: A versatile tool for mass spectrometry quantification.* Proteomics, 2011. **11**(17): p. 3572-3577.
402. Krey, J.F. *et al.*, *Accurate Label-Free Protein Quantitation with High- and Low-Resolution Mass Spectrometers.* Journal of Proteome Research, 2014. **13**(2): p. 1034-1044.
403. Tyanova, S. *et al.*, *Visualization of LC-MS/MS proteomics data in MaxQuant.* Proteomics, 2015. **15**(8): p. 1453-1456.
404. Sinitcyn, P. *et al.*, *MaxQuant goes Linux.* Nature Methods, 2018. **15**(6): p. 401-401.
405. Argentini, A. *et al.*, *moFF: a robust and automated approach to extract peptide ion intensities.* Nature Methods, 2016. **13**(12): p. 964-966.
406. Argentini, A. *et al. Using moFF to Extract Peptide Ion Intensities from LC-MS experiments*. Protocol Exchange, 2016. DOI: doi:10.1038/protex.2016.085.
407. Sandin, M. *et al.*, *Data processing methods and quality control strategies for label-free LC–MS protein quantification.* Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 2014. **1844**(1, Part A): p. 29-41.
408. Schliekelman, P. and S. Liu, *Quantifying the Effect of Competition for Detection between Coeluting Peptides on Detection Probabilities in Mass-Spectrometry-Based Proteomics.* Journal of Proteome Research, 2013. **13**(2): p. 348-361.
409. Choi, M. *et al.*, *MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments.* Bioinformatics, 2014. **30**(17): p. 2524-2526.
410. Arsova, B., H. Zauber, and W.X. Schulze, *Precision, Proteome Coverage, and Dynamic Range of Arabidopsis Proteome Profiling Using 15N Metabolic Labeling and Label-free Approaches.* Molecular & Cellular Proteomics, 2012. **11**(9): p. 619-628.
411. Brazma, A. *et al.*, *The PRIDE database and related tools and resources in 2019: improving support for quantification data.* Nucleic Acids Research, 2018. **47**(D1): p. D442-D450.
412. Sandberg, A. *et al.*, *Quantitative accuracy in mass spectrometry based proteomics of complex samples: The impact of labeling and precursor interference.* Journal of Proteomics, 2014. **96**: p. 133-144.
413. Michalski, A., J. Cox, and M. Mann, *More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority is Inaccessible to Data-Dependent LC−MS/MS.* Journal of Proteome Research, 2011. **10**(4): p. 1785-1793.
414. Griss, J. *et al.*, *Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets.* Nature Methods, 2016. **13**: p. 651.
415. Leitner, A., A. Foettinger, and W. Lindner, *Improving fragmentation of poorly fragmenting peptides and phosphopeptides during collision-induced dissociation by malondialdehyde modification of arginine residues.* Journal of Mass Spectrometry, 2007. **42**(7): p. 950-9.
416. Chick, J.M. *et al.*, *A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides.* Nature Biotechnology, 2015. **33**(7): p. 743-749.
417. Prakash, A. *et al.*, *Signal Maps for Mass Spectrometry-based Comparative Proteomics.* Molecular & Cellular Proteomics, 2006. **5**(3): p. 423-432.
418. Mueller, L.N. *et al.*, *SuperHirn – a novel tool for high resolution LC-MS-based peptide/protein profiling.* Proteomics, 2007. **7**(19): p. 3470-3480.
419. Bielow, C., G. Mastrobuoni, and S. Kempa, *Proteomics Quality Control: Quality Control Software for MaxQuant Results.* Journal of Proteome Research, 2016. **15**(3): p. 777-787.
420. Beer, L.A. *et al.*, *Efficient Quantitative Comparisons of Plasma Proteomes Using Label-Free Analysis with MaxQuant.* Methods in molecular biology (Clifton, N.J.), 2017. **1619**: p. 339-352.

421. Milac, T.I., T.W. Randolph, and P. Wang, *Analyzing LC-MS/MS data by spectral count and ion abundance: two case studies.* Statistics and Its Interface, 2012. **5**(1): p. 75-87.
422. Goeminne, L.J.E. *et al.*, *Summarization vs Peptide-Based Models in Label-Free Quantitative Proteomics: Performance, Pitfalls, and Data Analysis Guidelines.* Journal of Proteome Research, 2015. **14**(6): p. 2457-2465.
423. Ramus, C. *et al. Spiked proteomic standard dataset for testing label-free quantitative software and statistical methods*. Data in Brief, 2016. **6**, 286-294 DOI: 10.1016/j.dib.2015.11.063.
424. Karpievitch, Y.V., A.R. Dabney, and R.D. Smith, *Normalization and missing value imputation for label-free LC-MS analysis.* BMC Bioinformatics, 2012. **13 Suppl 16**: p. S5.
425. Karpievitch, Y. *et al.*, *A statistical framework for protein quantitation in bottom-up MS-based proteomics.* Bioinformatics, 2009. **25**(16): p. 2028-2034.
426. Paulovich, A.G. *et al.*, *Interlaboratory Study Characterizing a Yeast Performance Standard for Benchmarking LC-MS Platform Performance.* Molecular & Cellular Proteomics, 2010. **9**(2): p. 242-254.
427. Bourgon, R., R. Gentleman, and W. Huber, *Independent filtering increases detection power for high-throughput experiments.* Proceedings of the National Academy of Sciences, 2010. **107**(21): p. 9546-9551.
428. Lai, X. *et al.*, *A novel alignment method and multiple filters for exclusion of unqualified peptides to enhance label-free quantification using peptide intensity in LC-MS/MS.* Journal of Proteome Research, 2011. **10**(10): p. 4799-4812.
429. Gentleman, R. *et al.*, *genefilter: genefilter: methods for filtering genes from high-throughput experiments*. 2018, Bioconductor/R package.
430. Belouah, I. *et al.*, *Peptide filtering differently affects the performances of XIC-based quantification methods.* Journal of Proteomics, 2019. **193**: p. 131-141.
431. Yang, Y.H. *et al.*, *Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.* Nucleic Acids Research, 2002. **30**(4): p. e15.
432. Amaratunga, D. and J. Cabrera, *Analysis of Data From Viral DNA Microchips.* Journal of the American Statistical Association, 2001. **96**(456): p. 1161-1170.
433. Bolstad, B.M. *et al.*, *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-193.
434. Park, T. *et al.*, *Evaluation of normalization methods for microarray data.* BMC Bioinformatics, 2003. **4**: p. 33.
435. Valikangas, T., T. Suomi, and L.L. Elo, *A systematic evaluation of normalization methods in quantitative label-free proteomics.* Briefings in Bioinformatics, 2016.
436. Huber, W. *et al.*, *Variance stabilization applied to microarray data calibration and to the quantification of differential expression.* Bioinformatics, 2002. **18 Suppl 1**: p. S96-104.
437. Callister, S.J. *et al.*, *Normalization Approaches for Removing Systematic Biases Associated with Mass Spectrometry and Label-Free Proteomics.* Journal of Proteome Research, 2006. **5**(2): p. 277-286.
438. Kultima, K. *et al.*, *Development and Evaluation of Normalization Methods for Label-free Relative Quantification of Endogenous Peptides.* Molecular & Cellular Proteomics, 2009. **8**(10): p. 2285-2295.
439. Webb-Robertson, B.-J.M. *et al.*, *A statistical selection strategy for normalization procedures in LC-MS proteomics experiments through dataset-dependent ranking of normalization scaling factors.* Proteomics, 2011. **11**(24): p. 4736-4741.
440. Little, R.J.A. and D.B. Rubin, *Statistical Analysis with Missing Data, 2nd Edition*. Wiley Series in Probability and Statistics. 2002: Wiley.
441. Beretta, L. and A. Santaniello, *Nearest neighbor imputation algorithms: a critical evaluation.* BMC Medical Informatics and Decision Making, 2016. **16**(3): p. 74.
442. Troyanskaya, O. *et al.*, *Missing value estimation methods for DNA microarrays.* Bioinformatics, 2001. **17**(6): p. 520-525.

443. Lazar, C., *QRILC: a quantile regression approach for the imputation of left-censored missing data in quantitative proteomics.* to be submitted.
444. Gatto, L. and K.S. Lilley, *MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation.* Bioinformatics, 2012. **28**(2): p. 288-289.
445. Tyanova, S. *et al.*, *The Perseus computational platform for comprehensive analysis of (prote)omics data.* Nature Methods, 2016. **13**: p. 731.
446. Choi, M., *A flexible and versatile framework for statistical design and analysis of quantitative mass spectrometry-based proteomic experiments*. 2016, Purdue University: Open Access Dissertations.
447. Lazar, C. *et al.*, *Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies.* Journal of Proteome Research, 2016. **15**(4): p. 1116-1125.
448. Webb-Robertson, B.-J.M. *et al.*, *Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics.* Journal of Proteome Research, 2015. **14**(5): p. 1993-2001.
449. Välikangas, T., T. Suomi, and L.L. Elo, *A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation.* Briefings in Bioinformatics, 2017: p. bbx054-bbx054.
450. Zhang, X. *et al.*, *Proteome-wide identification of ubiquitin interactions using UbIA-MS.* Nature Protocols, 2018. **13**: p. 530.
451. Xiaoyan, Y. *et al.*, *Multiple imputation and analysis for high-dimensional incomplete proteomics data.* Statistics in Medicine, 2016. **35**(8): p. 1315-1326.
452. Wang, J. *et al.*, *In-depth method assessments of differentially expressed protein detection for shotgun proteomics data with missing values.* Scientific Reports, 2017. **7**(1): p. 3367.
453. Wu, Z. *et al.*, *Quantitative Chemical Proteomics Reveals New Potential Drug Targets in Head and Neck Cancer.* Molecular & Cellular Proteomics, 2011. **10**(12).
454. Breitwieser, F.P. *et al.*, *General Statistical Modeling of Data from Protein Relative Expression Isobaric Tags.* Journal of Proteome Research, 2011. **10**(6): p. 2758-2766.
455. Lin, W.-T. *et al.*, *Multi-Q: A Fully Automated Tool for Multiplexed Protein Quantitation.* Journal of Proteome Research, 2006. **5**(9): p. 2328-2338.
456. Raj, D.A.A. *et al.*, *A multiplex quantitative proteomics strategy for protein biomarker studies in urinary exosomes.* Kidney Int, 2012. **81**(12): p. 1263-1272.
457. Mosteller, F. and J.W. Tukey, *Data Analysis and Regression: A Second Course in Statistics*. 1977: Addison-Wesley Publishing Company.
458. Hoaglin, D.C., F. Mosteller, and J.W. Tukey, *Understanding robust and exploratory data analysis*. 1983: Wiley.
459. Gygi, S.P. *et al.*, *Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.* Nature Biotechnology, 1999. **17**(10): p. 994-999.
460. Najm, F.J. *et al.*, *Drug-based modulation of endogenous stem cells promotes functional remyelination in vivo.* Nature, 2015. **522**: p. 216.
461. Mertins, P. *et al.*, *Investigation of Protein-tyrosine Phosphatase 1B Function by Quantitative Proteomics.* Molecular & Cellular Proteomics, 2008. **7**(9): p. 1763-1777.
462. Margolin, A.A. *et al.*, *Empirical Bayes Analysis of Quantitative Proteomics Experiments.* PLoS ONE, 2009. **4**(10): p. e7454.
463. Al Shweiki, M.H.D.R. *et al.*, *Assessment of Label-Free Quantification in Discovery Proteomics and Impact of Technological Factors and Natural Variability of Protein Abundance.* Journal of Proteome Research, 2017. **16**(4): p. 1410-1424.
464. Blainey, P., M. Krzywinski, and N. Altman, *Points of Significance: Replication.* Nature Methods, 2014. **11**(9): p. 879-880.
465. Oberg, A.L. and O. Vitek, *Statistical Design of Quantitative Mass Spectrometry-Based Proteomic Experiments.* Journal of Proteome Research, 2009. **8**(5): p. 2144-2156.

466. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(9): p. 5116-5121.
467. Giai Gianetto, Q. *et al.*, *Uses and misuses of the fudge factor in quantitative discovery proteomics.* Proteomics, 2016. **16**(14): p. 1955-1960.
468. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.
469. Ting, L. *et al.*, *Normalization and Statistical Analysis of Quantitative Proteomics Data Generated by Metabolic Labeling.* Molecular & Cellular Proteomics, 2009. **8**(10): p. 2227-2242.
470. Daly, D.S. *et al.*, *Mixed-Effects Statistical Model for Comparative LC−MS Proteomics Studies.* Journal of Proteome Research, 2008. **7**(3): p. 1209-1217.
471. Clough, T. *et al.*, *Protein Quantification in Label-Free LC-MS Experiments.* Journal of Proteome Research, 2009. **8**(11): p. 5275-5284.
472. Clough, T. *et al.*, *Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs.* BMC Bioinformatics, 2012. **13**(16): p. S6.
473. Bukhman, Y.V. *et al.*, *Design and analysis of quantitative differential proteomics investigations using LC-MS technology.* Journal of Bioinformatics and Computational Biology, 2008. **6**(1): p. 107-23.
474. Henao, R. *et al. Hierarchical factor modeling of proteomics data*. in *Computational Advances in Bio and Medical Sciences (ICCABS), 2012 IEEE 2nd International Conference on*. 2012.
475. Blein-Nicolas, M. *et al.*, *Including shared peptides for estimating protein abundances: A significant improvement for quantitative proteomics.* Proteomics, 2012. **12**(18): p. 2797-2801.
476. Koopmans, F. *et al.*, *Empirical Bayesian Random Censoring Threshold Model Improves Detection of Differentially Abundant Proteins.* Journal of Proteome Research, 2014.
477. Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. Springer Series in Statistics. 2009, New York: Springer.
478. Stein, C. *Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution*. in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. 1956. Berkeley, Calif.: University of California Press.
479. Golub, G.H., M. Heath, and G. Wahba, *Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter.* Technometrics, 1979. **21**(2): p. 215-223.
480. Azevedo, C.F. *et al.*, *Ridge, Lasso and Bayesian additive-dominance genomic models.* BMC Genetics, 2015. **16**(1): p. 105.
481. Tissier, R., J. Houwing-Duistermaat, and M. Rodríguez-Girondo, *Improving stability of prediction models based on correlated omics data by using network approaches.* PloS one, 2018. **13**(2): p. e0192853-e0192853.
482. Bolstad, B.M., *Low-level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization*. 2004, University of California, Berkeley.
483. Liu, H., R.G. Sadygov, and J.R. Yates, *A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics.* Analytical Chemistry, 2004. **76**(14): p. 4193-4201.
484. Colinge, J. *et al.*, *Differential Proteomics via Probabilistic Peptide Identification Scores.* Analytical Chemistry, 2005. **77**(2): p. 596-606.
485. Li, M. *et al.*, *Comparative shotgun proteomics using spectral count data and quasi-likelihood modeling.* Journal of Proteome Research, 2010. **9**(8): p. 4295-4305.
486. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-140.

487. McCarthy, D.J., Y. Chen, and G.K. Smyth, *Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.* Nucleic Acids Research, 2012. **40**(10): p. 4288-4297.
488. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology, 2014. **15**(12): p. 550.
489. Branson, O.E. and M.A. Freitas, *Tag-Count Analysis of Large-Scale Proteomic Data.* Journal of Proteome Research, 2016. **15**(12): p. 4742-4746.
490. Choi, H., D. Fermin, and A.I. Nesvizhskii, *Significance analysis of spectral count data in label-free shotgun proteomics.* Molecular & Cellular Proteomics, 2008. **7**(12): p. 2373-2385.
491. Pham, T.V. *et al.*, *On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics.* Bioinformatics, 2010. **26**(3): p. 363-369.
492. Richardson, K. *et al.*, *A Probabilistic Framework for Peptide and Protein Quantification from Data-Dependent and Data-Independent LC-MS Proteomics Experiments.* OMICS: A Journal of Integrative Biology, 2012. **16**(9): p. 468-482.
493. Zhang, B. *et al.*, *Detecting Differential and Correlated Protein Expression in Label-Free Shotgun Proteomics.* Journal of Proteome Research, 2006. **5**(11): p. 2909-2918.
494. Lundgren, D.H. *et al.*, *Role of spectral counting in quantitative proteomics.* Expert Review of Proteomics, 2010. **7**(1): p. 39-53.
495. Lu, P. *et al.*, *Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation.* Nature Biotechnology, 2007. **25**(1): p. 117-124.
496. Liu, K. *et al.*, *Relationship between Sample Loading Amount and Peptide Identification and Its Effects on Quantitative Proteomics.* Analytical Chemistry, 2009. **81**(4): p. 1307-1314.
497. Schulze, W.X. and B. Usadel, *Quantitation in Mass-Spectrometry-Based Proteomics.* Annual Review of Plant Biology, 2010. **61**(1): p. 491-516.
498. Old, W.M. *et al.*, *Comparison of Label-free Methods for Quantifying Human Proteins by Shotgun Proteomics.* Molecular & Cellular Proteomics, 2005. **4**(10): p. 1487-1502.
499. Dunn, O.J., *Multiple Comparisons Among Means.* Journal of the American Statistical Association, 1961. **56**(293): p. 52-64.
500. Burger, T., *Gentle Introduction to the Statistical Foundations of False Discovery Rate in Quantitative Proteomics.* Journal of Proteome Research, 2018. **17**(1): p. 12-22.
501. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.
502. Hart, J.R. *et al.*, *The butterfly effect in cancer: a single base mutation can remodel the cell.* Proceedings of the National Academy of Sciences of the United States of America, 2015. **112**(4): p. 1131-1136.
503. Meissner, F. and M. Mann, *Quantitative shotgun proteomics: considerations for a high-quality workflow in immunology.* Nature Immunology, 2014. **15**(2): p. 112-117.
504. Doll, S. *et al.*, *Region and cell-type resolved quantitative proteomic map of the human heart.* Nature Communications, 2017. **8**(1): p. 1469.
505. Smith, R. *et al.*, *Proteomics, lipidomics, metabolomics: a mass spectrometry tutorial from a computer scientist's point of view.* BMC Bioinformatics, 2014. **15**(7): p. 1-14.

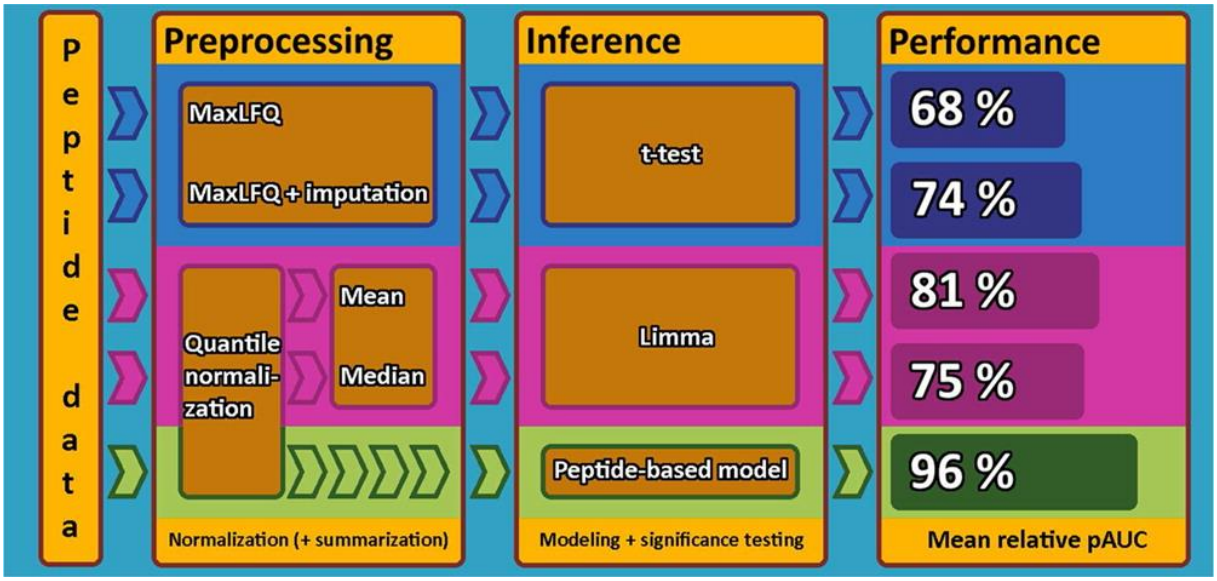
PART II: RESEARCH PAPERS

8. SUMMARIZATION VS PEPTIDE-BASED MODELS IN LABEL-FREE QUANTITATIVE PROTEOMICS: PERFORMANCE, PITFALLS, AND DATA ANALYSIS GUIDELINES

In chapter 8, we show that peptide-based models outperform summarization-based pipelines for the analysis of quantitative proteomics data. We also demonstrate that the predefined false discovery rate cut-offs for the detection of differentially regulated proteins can become problematic when differentially abundant (DA) proteins are highly abundant in one or more samples. We also show that care should be taken when data are interpreted from samples with spiked-in internal controls and from samples that contain a few very highly abundant proteins. For this work, I performed most of the data analysis and wrote the manuscript together with my co-authors.

Goeminne L.J.E.*, Argentini A.*, Martens L. and Clement L. (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

* equal contributions



8.1. Abstract

Quantitative label-free mass spectrometry is increasingly used to analyze the proteomes of complex biological samples. However, the choice of appropriate data analysis methods remains a major challenge. We therefore provide a rigorous comparison between peptide-based models and peptide-summarization-based pipelines. We show that peptide-based models outperform summarization-based pipelines in terms of sensitivity, specificity, accuracy, and precision. We also demonstrate that the predefined FDR cutoffs for the detection of differentially regulated proteins can become problematic when differentially expressed (DE) proteins are highly abundant in one or more samples. Care should therefore be taken when data are interpreted from samples with spiked-in internal controls and from samples that

contain a few very highly abundant proteins. We do, however, show that specific diagnostic plots can be used for assessing differentially expressed proteins and the overall quality of the obtained fold change estimates. Finally, our study also illustrates that imputation under the "missing by low abundance" assumption is beneficial for the detection of differential expression in proteins with low abundance, but it negatively affects moderately to highly abundant proteins. Hence, imputation strategies that are commonly implemented in standard proteomics software should be used with care.

8.2. Keywords

data analysis; differential proteomics; linear model

8.3. Introduction

Current high throughput mass spectrometry (MS) experiments enable the simultaneous identification and quantification of thousands of peptides and proteins in biological samples under various experimental conditions. These methods allow us to extend our understanding of biological processes and are important for the identification of biomarkers for the early detection, diagnosis, and prognosis of disease. Quantitative proteomics workflows broadly fall into two categories: labeled approaches and label-free approaches.(1) Labeled workflows rely on the labeling of proteins or peptides with isobaric or isotopic mass tags and are currently more commonly used. Label-free proteomics workflows, however, do not require these additional labor intensive and expensive sample processing steps.(2) Label-free approaches can perform quantitative proteome comparisons among an unlimited number of samples and can also be applied retroactively to previously acquired data.(3) Although label-free quantifications tend to have slightly higher coefficients of variation compared to SILAC labeling, label-free quantifications are more reproducible and can identify up to 60% more proteins than labeled quantifications.(4, 5)

A typical label-free shotgun MS-based proteomics workflow consists of (a) a protein extraction step followed by enzymatic digestion, (b) reverse phase high performance liquid chromatography (HPLC) separation, (c) mass spectrometry (MS), (d) a data analysis step involving the identification and quantification of peptides and proteins, and (e) a statistical analysis for assessing differential protein abundance.(6, 7) In a typical data-dependent analysis, selected peptides are isolated and fragmented, generating a fragmentation spectrum that is then used for peptide identification.(8) Technological constraints, however, limit the number of peptides in each fraction that can be selected for fragmentation. As the selection criteria typically involve the MS peak intensities in a particular time window, the identifications in MS-based experiments are inherently associated with the abundance of ionized peptides. Moreover, the steric effects of digestion enzymes(9) and differences in ionization efficiency favor particular peptides. Coeluting peptides heavily influence the observed MS intensities.(10) Hence, proteomics data suffer from nonrandom missing values and a large variability, rendering the development of reliable data analysis pipelines for quantitative proteomics a challenging task.(11)

The current data analysis strategies for label-free quantitative proteomics are typically based on spectral counting or peak intensities.(1) In the former approach, the number of peptide-to-spectrum matches (PSMs) for a given peptide are counted, and these are then accumulated over all peptides from a given protein.(12) Even though these methods are very intuitive and easy to apply, they remain controversial.(13, 14) Moreover, these methods necessarily ignore a large part of the information available in high precision mass spectra and are not very efficient in detecting low fold changes.(15) Peak-intensity-based methods, however, use the maximum

intensity or the area under the peak as a proxy for peptide abundance and tend to produce more precise protein abundance estimates.(15) We therefore focus on these latter approaches.

Many peak-based data analysis methods for the preprocessing and differential analysis of quantitative label-free proteomics data have been described in the literature. Modular approaches consisting of a separate normalization, summarization, and data analysis step are commonly used.(16, 17) Peptides originating from the same protein can indeed be considered technical replicates and theoretically should lead to similar abundance estimates. However, the summarization of the peptide intensities into protein expression values is cumbersome, and most summarization-based methods do not correct for differences in peptide characteristics or for the between-sample differences in the number of peptides that are identified per protein. This might introduce bias and differences in uncertainty between the aggregated protein expression values, which are typically ignored in downstream data analysis steps. The aforementioned nonrandom character of missing peptides further exacerbates these issues.

In response, linear regression approaches have been developed that immediately estimate the differential abundance between the proteins from observed peptide intensities, and their authors have made bold claims on their performance.(18, 19) Objective comparisons and general guidelines for the practitioner are, however, still lacking, which impedes the dissemination of more efficient data analysis pipelines into the proteomics community.

In this paper, we therefore present a rigorous comparison among modular and peptide-based regression methods for analyzing label-free quantitative proteomics data. We exploit the availability of the benchmark data sets to provide insight into the performance differences and technological artifacts that often arise in label-free proteomics experiments. It should also be noted that the benchmark data used here present a range of concentration differences, which enables us to analyze the suitability of different methods for different situations (e.g., small abundance differences versus large abundance differences or few missing peptides versus many missing peptides across analyses). In section 2 we present the benchmark data, the different data analysis methods, and the performance criteria that will be used in our comparison. The results are presented and discussed in sections 3 and 4.

8.4. Materials and methods

We used the publicly available data set from Study 6 of the Clinical Proteomic Technology Assessment for Cancer (CPTAC) Network(20) for assessing the performance of different data analytic workflows for quantitative label-free proteomics. In the CPTAC study, a mixture of 48 human proteins from the Sigma-Aldrich Universal Proteomics Standard 1 (UPS) was spiked into a 60 ng of protein/μL resuspended yeast lysate of *Saccharomyces cerevisiae* strain BY4741 (*MATa*, *leu2Δ0*, *met15Δ0*, *ura3Δ0*, and *his3Δ1*). Spike-ins were performed at five different concentrations: 0.25 fmol of UPS protein/μL (A), 0.74 fmol of UPS protein/μL (B), 2.22 fmol of UPS protein/μL (C), 6.67 fmol of UPS protein/μL (D), and 20 fmol of UPS protein/μL (E). The prepared samples were then sent to five different laboratories and analyzed on four different mass spectrometry platforms.

We identified peptides by searching the data using MaxQuant v1.5 against the yeast UniprotKB/Swiss-Prot protein database (v 15.14) to which the 48 UPS protein sequences were added. Detailed search settings can be found in the Supporting Information. A general overview of the number of identified peptides and proteins in our search can be found in Table S1, Supporting Information. Statistical analyses were implemented in RStudio version

0.98.978 (RStudio, Boston, MA) interfacing R 3.1.0 ("Spring Dance"). Standard Perseus analysis workflows were executed in Perseus version 1.5. We introduce two Perseus workflows in section 8.4.1, two different modular pipelines that aggregate peptide intensities into protein expression values in section 8.4.2, and three different peptide-based regression methods in section 8.4.3. The performance criteria used to compare the different methods can be found in section 8.4.4.

8.4.1. Perseus-based workflows

We used a typical workflow implemented in the software package Perseus. The analysis starts from (a) the LFQ intensities given in the MaxQuant's proteinGroups.txt file, which consist of normalized and summarized intensities at protein level. The MaxLFQ procedure then proceeds as follows: for all pairwise comparisons of a protein between samples, the median ratio for the common peptides in both samples is calculated. Next, the abundance protein profile that optimally satisfies these protein ratios is reconstructed with a least-squares regression model. The whole profile is then rescaled to the cumulative intensity across the samples with preservation of the total summed intensity for a protein across the samples. As the resulting LFQ intensities are already normalized by the MaxLFQ procedure,(21) no additional normalization step is required. In (b), the LFQ protein intensities are read into Perseus. The proteins that are only identified by a modification site, the contaminants, and the reversed sequences are removed from the data set, and the remaining intensities are $\log_2$-transformed. Next, (c) involves the imputation of missing values using Perseus' standard settings.(22) Finally, (d) consists of inference by pairwise two-sample *t* tests. The multiple testing problem is addressed using the Benjamini–Hochberg False Discovery Rate (FDR) procedure.(23) The (a)–(d) pipeline is referred to as perseusImp. We also consider a second variant, perseusNoImp, in which the imputation step (c) is omitted.

8.4.2. Summarization-based workflows

A typical modular workflow for quantitative proteomics consists of a normalization, summarization, and statistical analysis step.(6, 7) In our contribution, we assess two customized pipelines that build upon popular mean and median summarization strategies for the summarization of peptide intensities into protein expression values. The following steps are considered in the analysis pipelines: (a) the intensities from MaxQuant's peptides.txt output file are $\log_2$-transformed and normalized using quantile normalization (with the peptides mapping to reversed sequences or mapping to multiple proteins being removed from the data), (b) peptide intensities are aggregated into protein expression values using mean or median summarization, and (c) summarized protein expression values are further analyzed using empirical Bayes moderated *t* tests implemented in the R/Bioconductor package "limma".(24) The Benjamini–Hochberg FDR procedure is used to correct for multiple testing. The two resulting methods are referred to as limmaMean and limmaMedian.

In the limma analysis, the following model is considered for each protein i:

$$y_{ikl} = treat_{ik} + exp_{il} + \varepsilon_{ikl}, (1)$$

with y_{ikl} being the aggregated protein intensity for the k-th treatment (*treat*) and the l-th experiment (*exp*) correcting for (lab × instrument × repeat) batch effects. ε_{ikl} is a random error term that is assumed to be normally distributed with mean 0 and variance σ_i^2. Note that the *treat* effect is the effect of interest. Contrasts between *treat* parameters can be interpreted as $\log_2$ fold changes for protein i. For instance, k = A indicates condition A (spike-in concentration of 0.25 fmol of UPS protein/µL) and k = E indicates condition E (spike-in concentration of 20

fmol of UPS protein/μL). If so, then $treat_{iE} - treat_{iA}$ indicates the expected $\log_2$ difference in concentration for protein i between group E and group A. The statistical significance of the contrasts can be addressed by using *t* tests. The limma analysis exploits the massively parallel nature of quantitative proteomics experiments and allows for the borrowing of strength across proteins to estimate the error variance, i.e., makes use of a moderated empirical Bayes variance estimator $\tilde{s}_i^2$:

$$\tilde{s}_i^2 = \frac{d_0 s_0^2 + d_i s_i^2}{d_0 + d_i}$$

with s_i and d_i being the standard deviation and the residual degrees of freedom for protein i, respectively, s_0 the estimated prior standard deviation, and d_0 the prior degrees of freedom. Both the prior standard deviation and the prior degrees of freedom are estimated using empirical Bayes by pooling information across all proteins. Hence, the protein-based variance s_i^2 is shrunk toward a common variance s_0^2, leading to more stable variance estimates ($\tilde{s}_i^2$). Note that the degrees of freedom from the moderated *t* test also increase to $d_0 + d_i$. Detailed information can be found in the work of Smyth.(24)

8.4.3. Peptide-based models

Peptide-based models use the MaxQuant peptides.txt file as input. In (a), the extracted peptide intensities are $\log_2$-transformed and quantile normalized (Figures S8 and S9, Supporting Information), and peptides mapping to reversed sequences or mapping to multiple proteins are removed from the data. In (b), the peptide data are modeled with three different candidate models. In (c), inference is done by pairwise contrast testing. Multiple testing is addressed using the Benjamini–Hochberg FDR.

Linear Model without Sample Effect

For each protein i, the following model is proposed:

$$y_{ijklm} = pep_{ij} + treat_{ik} + exp_{il} + \varepsilon_{ijklm}, (2)$$

with y_{ijklm} being the $\log_2$-transformed intensity for the j-th peptide sequence pep_{ij} of the k-th treatment $treat_{ik}$ and the l-th experiment exp_{il}. ε_{ijklm} is a normally distributed error term with mean 0 and variance σ_i^2. The index m refers to multiple spectra that are identified for the same peptide in the same experiment and the same treatment. Contrasts in $treat_{ik}$ parameters can again be interpreted as $\log_2$ fold changes for protein i. The model also incorporates a pep_{ij} effect to account for peptide-specific fluctuations around the mean protein intensity, which originate from differences in digestion and ionization efficiency, among others.(9)

Linear Model with Sample Effect

Model 2 is extended by incorporating an additional sample effect, $sample_{ikl}$, to capture deviations specific to each MS run (lab × instrument × treatment × repeat):

$$y_{ijklm} = pep_{ij} + treat_{ik} + exp_{il} + sample_{ikl} + \varepsilon_{ijklm}. (3)$$

Note that all remaining effects are similar to those of model 2.

Mixed Model with Random Sample Effect

The mixed model extends the linear model 3 by putting a normal prior on the sample effect, $sample_{ikl} \sim N(0, \sigma^2_{sample,i})$. This model accounts for the correlation within samples and

incorporates both within- and between-sample variability when inference is performed on contrasts in the $treat_{ik}$ effects. The degrees of freedom of the *t* tests are approximated using the Satterthwaite approximation,(25) and the Benjamini–Hochberg FDR procedure is used to account for multiple testing.(23)

8.4.4. Performance

For each method, *p* values are converted to *q* values using the Benjamini–Hochberg FDR procedure,(23) and a cutoff is set at 5% FDR. At this level, the number of false positives (FP), true positives (TP), false negatives (FN), and true negatives (TN) are recorded, and the nominal FDR level is compared to the observed false discovery rate $\widehat{\mathrm{FDR}}$ = FP/(FP + TP). Note that the observed FDR equals 1 minus the positive predictive value, PPV = TP/(FP + TP).

The ROC curves are constructed on the basis of the ordering of the *p* values. Bias (Figures S4 and S5, Supporting Information), standard deviation (sd), median absolute deviation (mad), and root mean squared error (RMSE) (Figures S6 and S7, Supporting Information) are calculated both for yeast and for UPS proteins. We also calculated the F1 score, which is defined as the harmonic mean of the PPV and the sensitivity. Higher F1 scores indicate that a method provides a good balance between the PPV and the recall (Figures S1 and S2, Supporting Information).

8.5. Results

We investigated the sensitivity, specificity, and F1 score of the test procedure as well as the accuracy and the precision of the fold change (FC) estimates for three peptide-based methods and four summarization-based data analysis pipelines using the CPTAC Study 6 data set.(20) This data set consists of samples with a uniform yeast proteome background in which human UPS peptides are spiked at five different concentrations (0.25, 0.74, 2.22, 6.67, and 20 fmol/μL). All 10 pairwise comparisons are assessed in each analysis. The following peptide-based methods are considered: a linear model without sample effect (lmNoSamp), a linear model with sample effect (lmSamp), and a mixed model with a random sample effect (mixedSamp). The summarization-based approaches consist of mean and median summarizations of peptides into protein expression values followed by limma analyses (limmaMean and limmaMedian) as well as the more advanced MaxLFQ summarization(21) followed by a standard Perseus workflow with and without imputation (perseusImp and perseusNoImp). All peptide identifications and intensities were based on MaxQuant so as to avoid biases due to the search engine or peak intensity calculation algorithm.

Receiver operating characteristic (ROC) curves for the four comparisons with the smallest differences in spiked-in protein abundance (B–A, C–B, D–C, and E–D) are shown in Figure 1. Detecting the differential abundance of UPS proteins is most challenging in these comparisons as they only involve fold changes (FCs) very close to 3. ROC curves for the six remaining comparisons can be found in Figure S1 in the Supporting Information. Figure 1 shows that the lmNoSamp and mixedSamp models clearly outperform the other methods. The lmNoSamp, mixedSamp, and perseusImp workflows do control the FDR at 5% for comparisons B–A, C–A, and C–B, but perseusNoImp could only control the FDR for comparisons B–A and C–B, and both limmaMean and limmaMedian could only control the FDR for comparison B–A (see Tables S2–S8 and S12 in the Supporting Information). The lmSamp method is unable to control the FDR. When differences in spiked-in concentrations increase, however, none of the methods are able to control the FDR correctly (Tables S2–S8 and S12 in the Supporting Information). lmSamp is more conservative but cannot control its FDR at 5%, either. The ROC

curves also show that the mean summarization outperforms the more robust but less efficient median summarization.

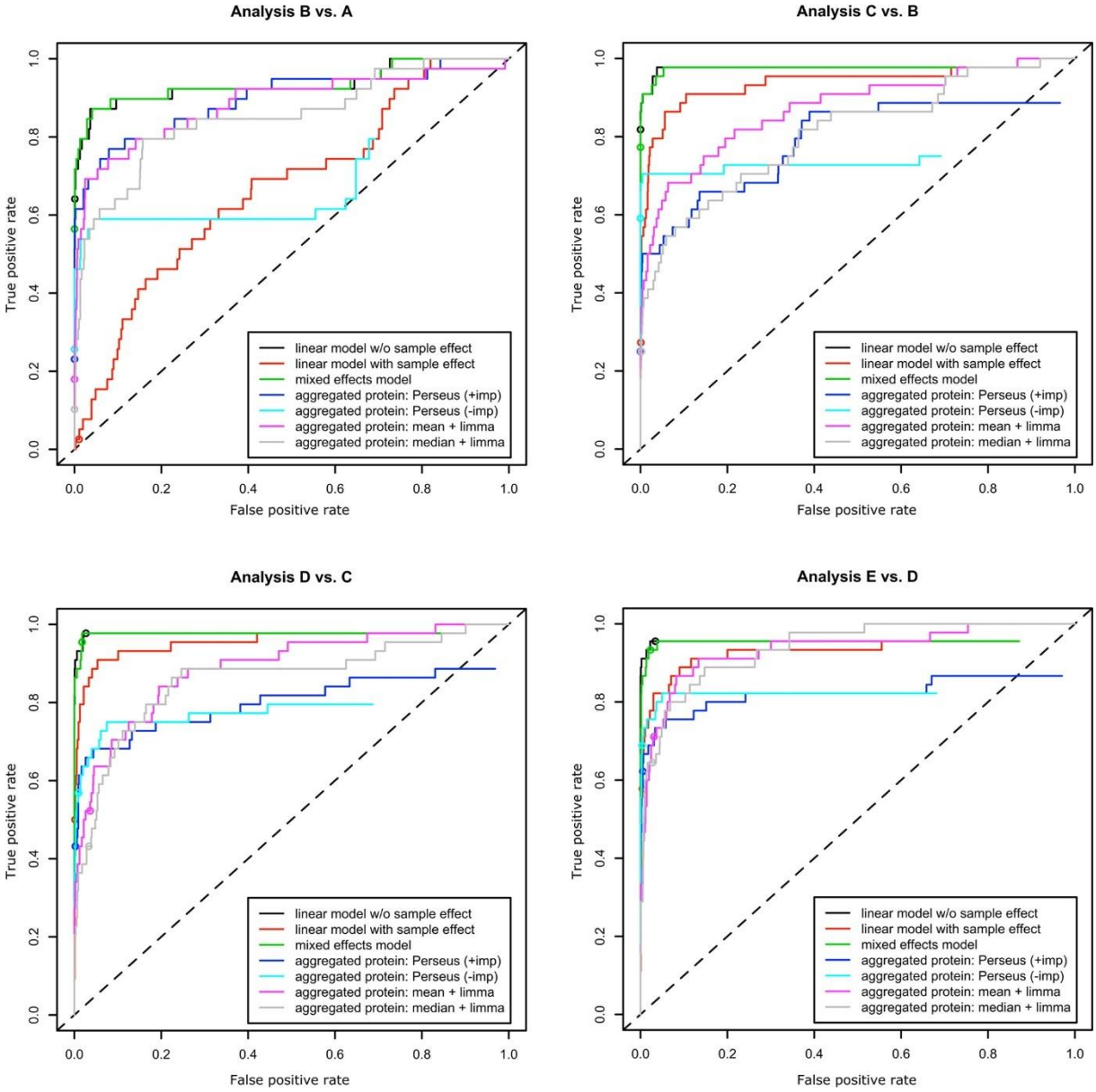


Figure 8.1. Receiver operating characteristic (ROC) curves for the seven analysis methods in comparisons B–A, C–B, D–C, and E–D. The UPS proteins in these comparisons were spiked in at a ratio close to 3:1. Dots denote the estimated cutoff for each method at 5% FDR. The termination of the curve before the point (1, 1) indicates either that proteins are prematurely removed from the analysis (e.g., for the Perseus workflows) or that there is an inability of the models to fit a protein with too few observations (e.g., for peptide-based models).

As only a part of the ROC curve is relevant in practice, i.e., that experimenters typically want to restrict the number of candidate proteins for validation in follow-up experiments, we also compared the relative partial areas under the curve (rpAUC) for FPR <0.1. Relative pAUCs (Table 8.1) are obtained by dividing pAUC values (Table S13, Supporting Information) by the maximum pAUC value of 0.1. Table 8.1 also demonstrates that the lmNoSamp and mixedSamp models are superior to the competing pipelines in terms of pAUC. Their power is higher in spite of the fact that no information is borrowed across proteins for estimating the variance (as compared to the limma workflows) and that there is an absence of imputation (as

compared to the standard Perseus method). perseusImp outperforms limmaMean and limmaMedian in terms of pAUC when differential expression in very-low-abundance proteins needs to be detected (e.g., comparison B–A). In these situations, imputation under the assumption of low abundance strongly boosts the performance of the method. The perseusImp workflow outperforms the perseusNoImp workflow for all comparisons involving A, i.e., when very-low-abundance differentially expressed (DE) proteins are involved in the comparison. But in comparisons with more abundant UPS spikes, the opposite is observed, and perseusImp shows a suboptimal performance compared to that of perseusNoImp.

Table 8.1. Relative Partial Area under the Curve (rpAUC) for FPR <0.1 for All Seven Models for Each of the Ten Comparisons[34].

CP	lm NoSamp	lmSamp	mixed Samp	perseus Imp	perseus NoImp	limma Mean	limma Median
B-A	83.01%	12.41%	83.93%	69.65%	55.70%	68.14%	56.21%
C-A	98.33%	49.31%	98.60%	85.93%	59.76%	87.22%	77.26%
D-A	99.20%	72.65%	99.26%	86.31%	63.42%	96.79%	95.18%
E-A	99.72%	89.06%	99.72%	89.62%	63.79%	97.92%	95.93%
C-B	95.10%	77.81%	94.60%	52.45%	70.08%	61.79%	50.54%
D-B	97.05%	89.60%	96.72%	71.00%	72.06%	89.21%	83.54%
E-B	96.98%	93.50%	95.90%	77.68%	72.41%	90.94%	87.69%
D-C	96.51%	84.85%	96.01%	64.48%	67.09%	59.77%	54.25%
E-C	97.34%	94.48%	96.21%	72.92%	74.85%	81.94%	79.23%
E-D	94.06%	79.45%	92.76%	71.13%	78.02%	74.16%	71.21%
Mean	95.73%	74.31%	95.37%	74.12%	67.72%	80.79%	75.11%

When the F1 score is examined, the lmNoSamp and mixedSamp models show very comparable patterns (Figure S2, Supporting Information). In comparisons E–A, E–B, E–C, and D–B, the lmSamp is superior to the other peptide-based models. This is most likely due to its more conservative nature. lmNoSamp and mixedSamp suffer from many false positives for these comparisons. For the summarization-based models (Figure S3, Supporting Information), we notice that perseusNoImp outperforms the other summarization-based methods for most comparisons. In comparisons D–A, D–B, E–A, E–B, and E–C, perseusImp shows a higher F1 score than Perseus without imputation. Again, the mean summarization method almost consistently outperforms the median summarization in terms of F1 score, although the differences are generally not very large.

The accuracy and precision of the pipelines are assessed by comparing the differential expression estimates to the true $\log_2$ fold changes of the spiked UPS peptides [$\log_2$ FC ≈ $\log_2(3)$, $\log_2(9)$, $\log_2(27)$, and $\log_2(80)$] and the yeast peptides ($\log_2$ FC = 0). Figures 8.2 and 8.3 show boxplots of the different DE estimates of the different methods for UPS and yeast proteins, respectively. The actual $\log_2$ FC is also indicated in the plot.

[34] The UPS proteins were spiked in at concentrations ranging from 0.25–20 fmol/µL (conditions A–E).

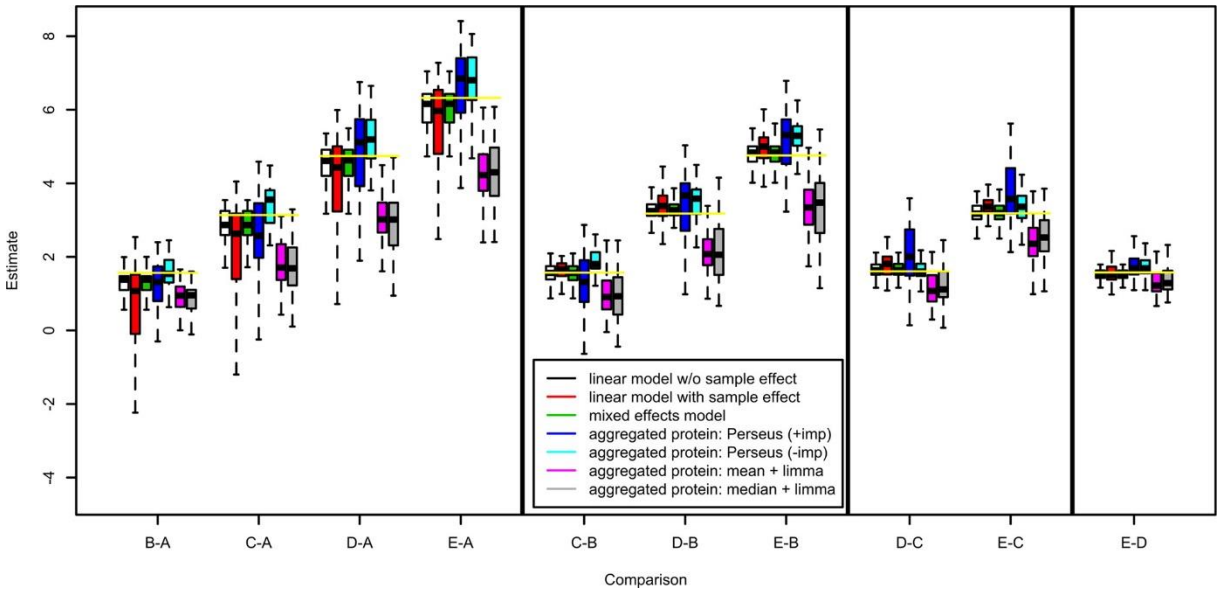


Figure 8.2. Boxplots showing the distributions of the DE estimates of the UPS proteins for each of the seven methods in each of the 10 comparisons. Outliers are not shown. The actual fold changes of the spikes are indicated with the yellow horizontal lines.

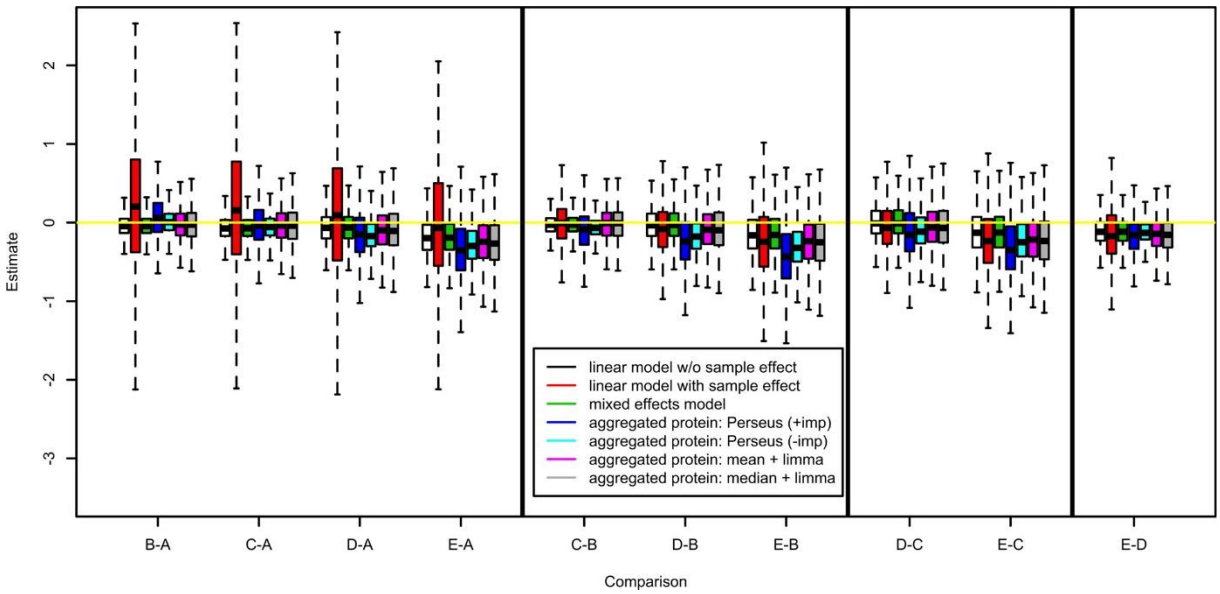


Figure 8.3. Boxplots showing the distributions of the DE estimates of the yeast proteins for each of the seven methods in each of the 10 comparisons. Outliers are not shown. All samples consisted of the same yeast background. Hence, no differential expression should occur for these proteins.

Figure 8.2 illustrates that the lmNoSamp and mixedSamp models are superior to the other methods in terms of both the accuracy and the precision of the FC estimates for the differentially abundant UPS proteins. The mean and median summarization methods systematically show a downward bias (Figure S4, Supporting Information). The bias is more pronounced in comparisons involving condition A. In condition A, the lowest concentration UPS (0.25 fmol/µL) is spiked in and, consequently, fewer UPS peptides are identified. Missingness, however, can be expected to involve peptides with a lower ionization efficiency, which typically display lower peak intensities than other peptides of the same protein. The simple mean and median summarization methods do not correct for differences in peptide characteristics, leading to an overestimation of the expression value for UPS proteins in condition A. This leads to moderation of the $\log_2$ fold change estimates involving condition A. Peptide-based pipelines

correcting for peptide effects also suffer from a slight negative bias in comparisons that involve condition A. Note that imputation has a severe impact on the precision. Figure S7 in the Supporting Information also shows that summarization-based limmaMean and limmaMedian methods give the highest root mean squared error (RMSE = $[bias^2/variance]^{1/2}$) among all methods that were evaluated.

Figure 8.3 confirms that the lmNoSamp and the mixedSamp models are favorable in terms of accuracy and precision. For the yeast proteins (non-DE), the median and mean summarization methods show a bias similar to that of competing methods. An increasing downward bias of the log_2 FC estimates can be observed for the null proteins (yeast) in comparisons involving increasing UPS concentrations. This becomes very apparent for comparisons that involve condition E. In this condition, a very high fraction of the total protein mass in the sample consists of UPS proteins. Hence, yeast peptides are likely to be masked by UPS, leading to an underestimation of the abundance of yeast peptides in the D and E mix. Most false positive yeast proteins had negative log_2 FC estimates as opposed to the spiked UPS proteins, which show positive log_2 FC estimates in each comparison. Therefore, issues involving the FDR are likely to be linked to the extreme sample composition under conditions D and E, which invokes an MS bias.(10) The F1 score masks this artifact, as it combines PPV and sensitivity. The same trend is also visible in MA plots for the linear model without sample effect (Figures S10 and S11, Supporting Information). In these graphs, the average FC is plotted in function of the average protein expression for a particular comparison. These graphs are therefore very helpful for screening for artifacts induced by the technical and data analysis workflows. Figure 8.3 also illustrates that the precision reduces with increasing FC, i.e., for comparisons involving conditions D and E. Finally, the lmSamp method shows a dramatic decrease in precision for comparison B–A. This is a data analysis artifact; the model is overidentified for many proteins, leading to the aliasing of sample and treatment effects. Due to the specific model parametrization, the overidentification has a larger impact on comparisons involving condition A.

We also investigated alternative data analysis strategies to alleviate this problem. For the peptide-based lmNoSamp method, we assessed the impact of testing against the median log_2 FC of all proteins instead of testing against 0. This slightly improves the observed FDR except for comparisons C–A and D–B and improves the rpAUC except for comparisons D–A, D–B, D–C, and E–C (Table S14, Supporting Information). However, the method still returns too many false positives for the comparisons involving high concentrations (Table S9, Supporting Information). For the summarization-based methods, we assessed the impact of switching the order of the normalization and summarization steps. When the quantile normalization is performed after summarization, the observed FDR improved for comparisons involving D and E, but the performance decreased dramatically for comparisons B–A and C–A (Tables S10–S12. Supporting Information). The ROC curves also suggest that switching the order of the normalization and summarization steps deteriorates the performance of the limmaMean and limmaMedian workflows (Figure S12, Supporting Information).

8.6. Discussion

Our analysis showed that peptide-based models perspicuously outperform summarization-based methods. Both the linear model without sample effect and the mixed model outperform the other methods in terms of accuracy, precision, sensitivity, and specificity. The ROC curves clearly indicate that these methods produce a more reliable ordering of DE proteins than the competing methods. The linear model with sample effect has a suboptimal performance but still outperforms the other methods in comparisons that do not involve A. Due to selective and periodic sampling in both MS stages, not all peptides are being observed or identified in all

samples. Moreover, intensities from different peptides of the same protein vary considerably due to differences in cleavage and ionization efficiency among others.(26, 27) Summarization thus typically involves different peptides and a different number of peptides in each sample. This leads to protein expression values with distinct characteristics, which induces bias and incorrect precision of the fold change estimates.

Peptide-based models are superior in correcting for individual peptide effects, which are typically quite strong(18, 19) and accounting for the different number of peptides in each sample. Thus, bias is reduced and improved precision estimates are provided, leading to higher sensitivity and specificity. The mixed model can also account for the correlation that is present in peptides from the same protein within a sample. The peptide-based models with a fixed sample effect suffer from the unstable estimation of fold changes and variance components due to the overfitting of sparse proteins identified by a few peptides. Moreover, the inclusion of a fixed sample effect eliminates the between-sample variability from the analysis. Inference between the samples will be based on an underestimated variance, leading to a higher number of false positives in a top list. In the linear model without sample effect, fewer parameters have to be estimated, and the variances within and between samples are combined in the error term. Hence, the method is less prone to overfitting and incorporates both within- and between-sample variability in the test statistics, leading to a better control of the number of false positives. However, the method does not account for the correlation between peptides from a particular protein within a sample. The mixed modeling approach with a random sample effect does incorporate within- and between-sample variances as well as the within-sample correlation between peptides of the same protein. The mixed model and the linear model without sample effect are more or less on par in terms of all assessed performance criteria. Hence, the increased computational complexity of the mixed model cannot be justified for this particular application. However, in real experiments, more correlation can be expected due to the additional biological variation among samples.

We also showed that the use of FDR thresholds might be flawed under certain experimental conditions. This was observed for comparisons involving conditions D and E, i.e., the samples with the highest spiked-in UPS concentrations. Under these conditions, the UPS proteins correspond to a considerable fraction of the total protein mass in the sample. The ROC curves show that peptide-based methods still produce reliable top lists with a superb ordering, but the use of a 5% FDR threshold was too liberal. Hence, long protein lists are produced with many false positives. The majority of these false positives, however, had FC estimates in the opposite direction as those of spiked UPS proteins. This was due to a systematic downward bias in the FC estimates of nondifferentially expressed yeast proteins. Competitive ionization makes the identification and quantification of yeast peptides cumbersome in samples with highly concentrated UPS spikes. Thus, the majority of false positives originate from technological artifacts rather than from flaws in the data analysis pipeline. We therefore recommend that researchers who are planning to use internal controls in their MS experiments avoid overspiking, as this can have detrimental effects on the quantification of the proteins of interest. Moreover, artifacts similar to those from spiked UPS proteins are bound to occur in certain experimental setups (e.g., undepleted blood plasma proteomic samples are known to be dominated by a few highly abundant proteins, and undepleted green tissue samples from plants will suffer from the omnipresence of RuBisCo). Our analysis showed that experimenters should interpret proteins further down the DE list with care. We therefore advise data analysts to use diagnostic plots based on all fold change estimates for assessing the quality of the FC estimates and for detecting potential artifacts. MA plots and boxplots were shown to be well suited for evaluating candidate DE proteins, to flag critical experimental conditions as well as flaws in the data analysis pipeline.

The myriad missing values in quantitative proteomic experiments present severe challenges to the data analysis. The standard MaxQuant pipeline therefore utilizes the match-between runs option to boost the number of peptide intensities that different samples have in common. Moreover, Perseus also incorporates imputation-based routines to deal with missing protein expression values. We showed that imputation is beneficial for detecting differentially expressed proteins with low abundance but performs suboptimally for moderately to highly abundant proteins. Perseus' standard imputation algorithm assumes that missing values originate from lower intensity values. Hence, the imputation can lead to a downward bias for more abundant proteins. Moreover, experimenters should also be aware that imputation comes at the cost of a decreased precision for the FC estimates.

In general, the current peptide-based methods are prone to overfitting and rely on protein-by-protein variance estimates. Hence, the development of robust methods that can borrow information across peptides and proteins would enable proteomics researchers to further deploy label-free quantitative proteomics.

In summary, we have shown that issues inherent to the methodology create challenges in quantitative proteomics, even in highly controlled and standardized samples such as the CPTAC ones. We then go on to show that downstream statistical data analysis approaches differ in their ability to cope with these different issues, and that the importance of these issues depends on the characteristics of the sample under study (e.g., the dominance of a few highly abundant proteins or large protein concentration ratio differences between two samples). Crucial, perhaps, is the fact that although peptide-based approaches fare better than summarization methods, no single method currently exists that can easily tackle all possible issues in quantitative proteomics data. Hence, more sophisticated data processing approaches that recognize these various issues are needed and can compensate for such issues more successfully across the board.

8.7. Conclusion

In this paper, we compared the performance of peptide-based linear models, mean and median summarization followed by limma analysis, and the standard MaxQuant/Perseus workflow for assessing differential abundance in label-free quantitative proteomics experiments. The evaluation of the performance was assessed using the CPTAC benchmark data set. Peptide-based models outperformed the competing data analysis pipelines in terms of sensitivity, specificity, accuracy, and precision. Modeling quantitative proteomics data at the peptide level allows for the correction of strong peptide-specific effects, which avoids the bias associated with summarization-based methods that aggregate different types of peptide intensities into a single value. Moreover, peptide-based models also improve the precision estimates by accounting for the different numbers of peptides that are identified in a sample. We have also shown that the FDR cutoffs used to determine the length of lists with significant differentially expressed (DE) proteins could become problematic in experimental setups with samples that are dominated by a few very abundant proteins. Technological artifacts might induce bias in the non-DE proteins, which can inflate the number of false positives that are returned at a particular FDR level. However, the ordering of the top DE proteins in the lists was shown to remain valid. We therefore advise proteomics researchers to be careful when spiking internal controls, to deplete the highly abundant proteins, and to use diagnostic plots for assessing the candidate DE proteins as well as the overall quality of the obtained fold change estimates. Finally, standard proteomics software provides experimenters with the ability to impute missing values. Perseus' imputation strategy was shown to be beneficial for detecting DE proteins with low abundance but at the cost of reduced precision as well as a suboptimal performance for

moderately to highly abundant DE proteins. Hence, we advise proteomics data analysts to use imputation strategies with care.

8.8. Supporting information

Figures showing the receiver operating characteristic (ROC) curves for the seven analysis methods in comparisons, F1 scores for the studied models, comparison of the bias terms for yeast and UPS proteins, comparisons of the root mean squared error for yeast and UPS proteins, boxplots showing $\log_2$ peptide intensities, MA plots for linear models, and ROC curves for normalization on the peptide and protein level with mean and median aggregation. Tables showing a general overview per spike-in conditions for UPS and yeast proteins, characteristics for various models and workflows, an explanation of the outlined characteristics, partial areas under the curves for a false positive rate, and relative partial areas under the curves for a false positive rate. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/pr501223t.

pr501223t_si_001.pdf (1.38 MB)

8.9. Acknowledgement

Part of this research was supported by IAP research network “StUDyS” grant no. P7/06 of the Belgian government (Belgian Science Policy) and the Multidisciplinary Research Partnership “Bioinformatics: From Nucleotides to Networks” of Ghent University. A.A. is supported by the IWT SBO grant “INSPECTOR” (120025). L.J.E.G. is supported by the IWT SBO grant “Differential Proteomics at Peptide, Protein, and Module Level” (141573). L.M. acknowledges the PRIME-XS project, grant agreement no. 262067, funded by the European Union Seventh Framework Program.

8.10. References

This article references 27 other publications.

1. Vaudel, M.; Sickmann, A.; Martens, L. Peptide and protein quantification: A map of the minefield Proteomics 2010, 10 (4) 650– 670

2. Bluemlein, K.; Ralser, M. Monitoring protein expression in whole-cell extracts by targeted label- and standard-free LC-MS/MS Nat. Protoc. 2011, 6 (6) 859– 869

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

4. Liu, N. Q.; Dekker, L. J. M.; Stingl, C.; Güzel, C.; De Marchi, T.; Martens, J. W. M.; Foekens, J. A.; Luider, T. M.; Umar, A. Quantitative Proteomic Analysis of Microdissected Breast Cancer Tissues: Comparison of Label-Free and SILAC-based Quantification with Shotgun, Directed, and Targeted MS Approaches J. Proteome Res. 2013, 12 (10) 4627– 4641

5. Mosley, A. L.; Sardiu, M. E.; Pattenden, S. G.; Workman, J. L.; Florens, L.; Washburn, M. P. Highly Reproducible Label Free Quantitative Proteomic Analysis of RNA Polymerase Complexes Mol. Cell. Proteomics 2011, DOI: 10.1074/mcp.M110.000687

6. Wang, G.; Wu, W. W.; Zeng, W.; Chou, C.-L.; Shen, R.-F. Label-Free Protein Quantification Using LC-Coupled Ion Trap or FT Mass Spectrometry: Reproducibility, Linearity, and Application with Complex Proteomes J. Proteome Res. 2006, 5 (5) 1214– 1223

7. Silva, J. C.; Gorenstein, M. V.; Li, G.-Z.; Vissers, J. P. C.; Geromanos, S. J. Absolute Quantification of Proteins by LCMSE: A Virtue of Parallel ms Acquisition Mol. Cell. Proteomics 2006, 5 (1) 144– 156

8. Vaudel, M.; Sickmann, A.; Martens, L. Current methods for global proteome identification Expert Rev. Proteomics 2012, 9 (5) 519– 532

9. Peng, M.; Taouatas, N.; Cappadona, S.; van Breukelen, B.; Mohammed, S.; Scholten, A.; Heck, A. J. R. Protease bias in absolute protein quantitation Nat. Methods 2012, 9 (6) 524– 525

10. Schliekelman, P.; Liu, S. Quantifying the Effect of Competition for Detection between Coeluting Peptides on Detection Probabilities in Mass-Spectrometry-Based Proteomics J. Proteome Res. 2013, 13 (2) 348– 361

11. Kumar, C.; Mann, M. Bioinformatics analysis of mass spectrometry-based proteomics data sets FEBS Lett. 2009, 583 (11) 1703– 1712

12. Liu, H.; Sadygov, R. G.; Yates, J. R. A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics Anal. Chem. 2004, 76 (14) 4193– 4201

13. Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B. Quantitative mass spectrometry in proteomics: a critical review Anal. Bioanal. Chem. 2007, 389 (4) 1017– 31

14. Mueller, L. N.; Brusniak, M.-Y.; Mani, D. R.; Aebersold, R. An Assessment of Software Solutions for the Analysis of Mass Spectrometry Based Quantitative Proteomics Data J. Proteome Res. 2008, 7 (1) 51– 61

15. Old, W. M.; Meyer-Arendt, K.; Aveline-Wolf, L.; Pierce, K. G.; Mendoza, A.; Sevinsky, J. R.; Resing, K. A.; Ahn, N. G. Comparison of Label-free Methods for Quantifying Human Proteins by Shotgun Proteomics Mol. Cell. Proteomics 2005, 4 (10) 1487– 1502

16. Théron, L.; Gueugneau, M.; Coudy, C.; Viala, D.; Bijlsma, A.; Butler-Browne, G.; Maier, A.; Béchet, D.; Chambon, C. Label-free Quantitative Protein Profiling of vastus lateralis Muscle During Human Aging Mol. Cell. Proteomics 2014, 13 (1) 283– 294

17. Hubner, N. C.; Bird, A. W.; Cox, J.; Splettstoesser, B.; Bandilla, P.; Poser, I.; Hyman, A.; Mann, M. Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions J. Cell Biol. 2010, 189 (4) 739– 754

18. Clough, T.; Key, M.; Ott, I.; Ragg, S.; Schadow, G.; Vitek, O. Protein Quantification in Label-Free LC-MS Experiments J. Proteome Res. 2009, 8 (11) 5275– 5284

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

20. Paulovich, A. G.; Billheimer, D.; Ham, A.-J. L.; Vega-Montoto, L.; Rudnick, P. A.; Tabb, D. L.; Wang, P.; Blackman, R. K.; Bunk, D. M.; Cardasis, H. L.; Clauser, K. R.; Kinsinger, C. R.; Schilling, B.; Tegeler, T. J.; Variyath, A. M.; Wang, M.; Whiteaker, J. R.; Zimmerman, L. J.; Fenyo, D.; Carr, S. A.; Fisher, S. J.; Gibson, B. W.; Mesri, M.; Neubert, T. A.; Regnier, F. E.; Rodriguez, H.; Spiegelman, C.; Stein, S. E.; Tempst, P.; Liebler, D. C. Interlaboratory Study Characterizing a Yeast Performance Standard for Benchmarking LC-MS Platform Performance Mol. Cell. Proteomics 2010, 9 (2) 242– 254

21. Cox, J.; Hein, M. Y.; Luber, C. A.; Paron, I.; Nagaraj, N.; Mann, M. Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ Mol. Cell. Proteomics 2014, 13 (9) 2513– 2526

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

23. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing J. R. Stat. Soc.: Series B 1995, 57 (1) 289– 300

24. Smyth, G. K., Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl. Genet. Mol. Biol. 2004, 3, Article 3.

25. Satterthwaite, F. E. An approximate distribution of estimates of variance components Biometrics 1946, 2 (6) 110– 4

26. Rodriguez, J.; Gupta, N.; Smith, R. D.; Pevzner, P. A. Does Trypsin Cut Before Proline? J. Proteome Res. 2008, 7 (1) 300– 305

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

9. ROBUST QUANTIFICATION FOR LABEL-FREE MASS SPECTROMETRY-BASED PROTEOMICS

Chapter 9 describes MSqRob, our R software package for improved differential protein abundance analysis in label-free MS-based proteomics. MSqRob is freely available on GitHub (https://github.com/statOmics/MSqRob) and is implemented in a "Shiny" user-friendly graphical interface.

In section 9.1, I introduce MSqRob as a new algorithm for the analysis of quantitative proteomics data that improves protein quantification by combining three innovative statistical approaches: ridge regression, empirical Bayes variance estimation, and M-estimation with Huber weights. MSqRob is both more precise and more accurate than state-of-the-art tools. I developed and implemented MSqRob as an R package and wrote the manuscript together with my supervisors.

Section 9.2 is published as an invited tutorial paper in which I outline key statistical concepts to help researchers to design proteomics experiments and showcases of quantitative proteomics data analysis with MSqRob. For this manuscript, I designed and performed analyses, set up the GitHub repository and wrote the paper together with my supervisors.

9.1. Peptide-level Robust Ridge Regression Improves Estimation, Sensitivity, and Specificity in Data-dependent Quantitative Label-free Shotgun Proteomics

Goeminne L.J.E., Gevaert K. and Clement L. (2016). **Peptide-level Robust Ridge Regression Improves Estimation, Sensitivity, and Specificity in Data-dependent Quantitative Label-free Shotgun Proteomics.** *Molecular & Cellular Proteomics*. 15(2), 657-668

9.1.1. Associated data

Supplementary Materials

Supplemental Data

supp_15_2_657__index.html (2.4K)

GUID: 9B15C066-3091-4E71-980A-55F4B6EE0B27

10.1074_M115.055897_mcp.M115.055897-1.pdf (7.0M)

GUID: 401B1F9E-2C1B-4FA6-9374-9048987008F2

10.1074_M115.055897_mcp.M115.055897-2.xlsx (208K)

GUID: B2490BE1-84E1-49D7-8951-5CB1D9597D83

10.1074_M115.055897_mcp.M115.055897-3.zip (72M)

GUID: 6BF2E185-A0C2-4F19-B71D-5999CF7F6AF2

9.1.2. Abstract

Peptide intensities from mass spectra are increasingly used for relative quantitation of proteins in complex samples. However, numerous issues inherent to the mass spectrometry workflow turn quantitative proteomic data analysis into a crucial challenge. We and others have shown that modeling at the peptide level outperforms classical summarization-based approaches, which typically also discard a lot of proteins at the data preprocessing step. Peptide-based linear regression models, however, still suffer from unbalanced datasets due to missing peptide intensities, outlying peptide intensities and overfitting. Here, we further improve upon peptide-based models by three modular extensions: ridge regression, improved variance estimation by borrowing information across proteins with empirical Bayes and M-estimation with Huber weights. We illustrate our method on the CPTAC spike-in study and on a study comparing wild-type and ArgP knock-out *Francisella tularensis* proteomes. We show that the fold change estimates of our robust approach are more precise and more accurate than those from state-of-the-art summarization-based methods and peptide-based regression models, which leads to an improved sensitivity and specificity. We also demonstrate that ionization competition effects come already into play at very low spike-in concentrations and confirm that analyses with peptide-based regression methods on peptide intensity values aggregated by charge state and modification status (*e.g.* MaxQuant's peptides.txt file) are slightly superior to analyses on raw peptide intensity values (*e.g.* MaxQuant's evidence.txt file).

9.1.3. Introduction

High-throughput LC-MS-based proteomic workflows are widely used to quantify differential protein abundance between samples. Relative protein quantification can be achieved by stable isotope labeling workflows such as metabolic (1, 2) and postmetabolic labeling (3–6). These types of experiments generally avoid run-to-run differences in the measured peptide (and thus protein) content by pooling and analyzing differentially labeled samples in a single run. Label-free quantitative (LFQ)[1] workflows become increasingly popular as the often expensive and time-consuming labeling protocols are omitted. Moreover, LFQ proteomics allows for more flexibility in comparing samples and tends to cover a larger area of the proteome at a higher dynamic range (7, 8). Nevertheless, the nature of the LFQ protocol makes shotgun proteomic data analysis a challenging task. Missing values are omnipresent in proteomic data generated by data-dependent acquisition workflows, for instance because of low-abundant peptides that are not always fragmented in complex peptide mixtures and a limited number of modifications and mutations that can be accounted for in the feature search. Moreover, the overall abundance of a peptide is determined by the surroundings of its corresponding cleavage sites as these influence protease cleavage efficiency (9). Similarly, some peptides are more easily ionized than others (10). These issues not only lead to missing peptides, but also increase variability in individual peptide intensities. The discrete nature of MS1 sampling following continuous elution of peptides from the LC column leads to increased variability in peptide quantifications. Finally, competition for ionization and co-elution of other peptides with similar m/z values may cause biased quantifications (11). However, note that in this respect, using data-independent acquisition (DIA), all peptide ions (or all peptide ions within a certain m/z range, depending on the method used) are fragmented simultaneously, resulting in multiplexed MS/MS spectra (12, 13). Hence, issues of missing fragment spectra are less a problem with DIA, however, some of its challenges lie in deconvoluting MS/MS spectra and mapping their features to their corresponding peptides (14).

Standard data analysis pipelines for DDA-LFQ proteomics can be divided into two groups: spectral counting techniques, which are based on counting the number of peptide features as

a proxy for protein abundance (15), and intensity-based methods that quantify peptide features by measuring their corresponding spectral intensities or areas under the peaks in either MS or MS/MS spectra. Spectral counting is intuitive and easy to perform, but, the determination of differences in peptide and thus protein levels is not as precise as intensity-based methods, especially when analyzing rather small differences (16). More fundamentally, spectral counting ignores a large part of the information that is available in high-precision mass spectra. Further, dynamic exclusion during LC-MS/MS analysis, meant to increase the overall number of peptides that are analyzed, can worsen the linear dynamic range of these methods (17). Also, any changes in the MS/MS sampling conditions will prevent comparisons between runs. Intensity-based methods are more sensitive than spectral counting (18). Among intensity-based methods, quantification on the MS-level is somewhat more accurate than summarizing the MS/MS-level feature intensities (19). Therefore, we further focus on improving data analysis methods for MS-level quantification.

Typical intensity-based workflows summarize peptide intensities to protein intensities before assessing differences in protein abundances (20). Peptide-based linear regression models estimate protein fold changes directly from peptide intensities and outperform summarization-based methods by reducing bias and generating more correct precision estimates (21, 22). However, peptide-based linear regression models suffer from overfitting due to extreme observations and the unbalanced nature of proteomics data; *i.e.* different peptides and a different number of peptides are typically identified in each sample. We illustrate this using the CPTAC spike-in data set where 48 human UPS1 proteins were spiked at five different concentrations in a 60 ng protein/µl yeast lysate. Thus, when comparing different spike-in concentrations, only the human proteins should be flagged as differentially abundant (DA), whereas the yeast proteins should not be flagged as DA (null proteins). Fig. 9.1 illustrates the structure of missing data in label-free shotgun proteomics experiments using a representative DA UPS1 protein from the CPTAC spike-in study: missing peptides in the lowest spike-in condition tend to have rather low $\log_2$ intensity values in higher spike-in conditions compared to peptides that were not missing in both conditions, which supports the fact that the missing value problem in label-free shotgun proteomic data is largely intensity-dependent (23).

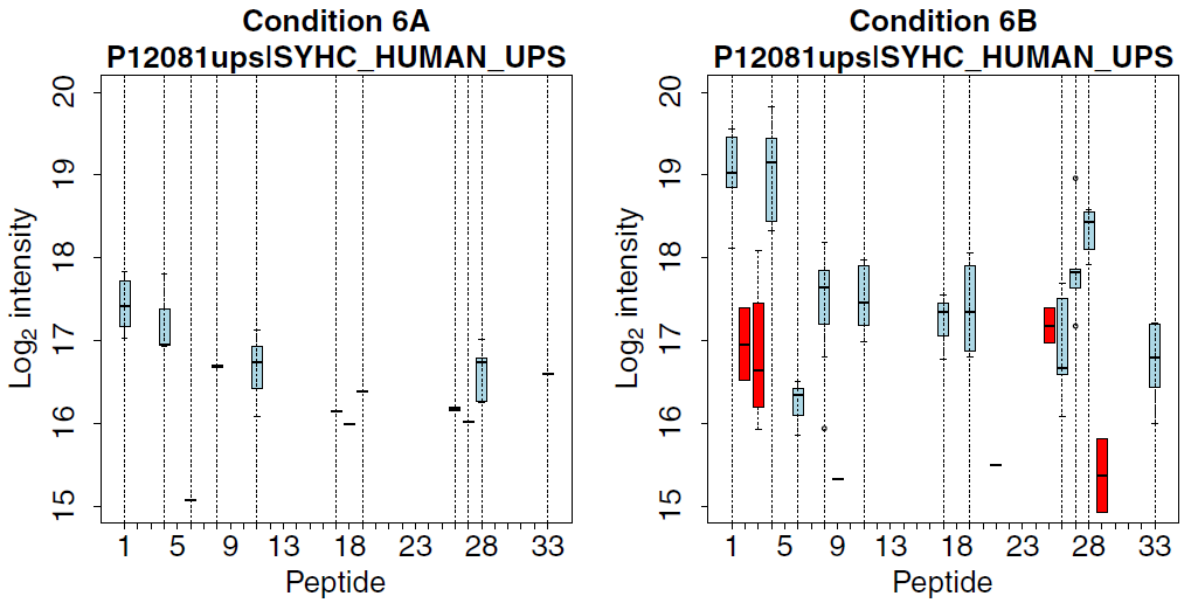


Figure 9.1._Missing peptides are often low abundant. The boxplots show the $\log_2$ intensity distributions for each of the 33 identified peptides corresponding to the human UPS1 protein cytoplasmic Histidyl-tRNA synthetase (P12081) from the CPTAC dataset in conditions 6A (spike-in concentration

0.25 fmol UPS1 protein/µl) and 6B (spike-in concentration 0.74 fmol UPS1 protein/µl). Vertical dotted lines indicate peptides present in both conditions. Note, that most peptides that were not detected in condition 6A exhibit low $\log_2$ intensity values in condition 6B (colored in red).

Fig. 9.2 shows the quantile normalized $\log_2$ intensity values for the peptides corresponding to the yeast null protein CG121 together with average $\log_2$ intensity estimates for each condition based on protein-level MaxLFQ intensities, as well as estimates derived from a peptide-based linear model. Here, three important remarks can be made:

(1) CG121 is a yeast background protein, for which the true concentration is thus equal in all conditions, which appears to be monitored as such by MaxLFQ, except in conditions 6B and 6E (for the latter, no estimate is available). The LM estimate, however, is more reliable but seems to suffer from overfitting.

(2) A lot of shotgun proteomic datasets are very sparse, causing a large sample-to-sample variability. Constructing a linear model based on a limited number of observations will thus lead to unstable variance estimates. Intuitively, a small sample drawn from a given population might “accidentally” show a very small variance while another small sample from the same population might display a very large variance just by random chance. This effect is clear from the sizes of the boxes. The interquartile range is twice as large in condition 6E compared to condition 6C. This issue leads to false positives since some proteins with very few observations are flagged as DA with very high statistical evidence solely due to their low observed variance (24).

(3) Two observed features at $\log_2$ intensities 14.0 and 14.3 in condition 6B have a strong influence on the parameter estimate for this condition. Without these extreme observations, the 6B estimate lies closer to the estimates in the other conditions. As missingness is strongly intensity-dependent, these low intensity values could easily become missing values in subsequent experiments. More generally, a strong influence of only one or two peptides on the average protein level intensity estimate for a condition is an unfavorable property.

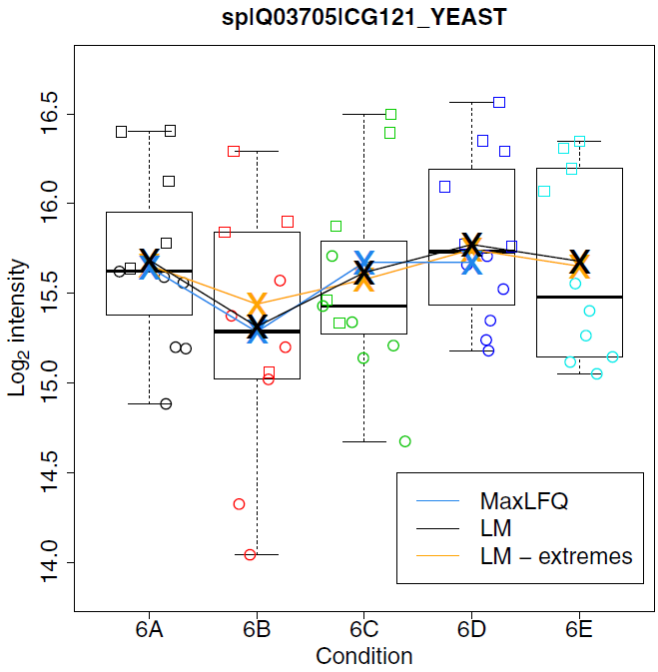


Figure 9.2. Effect of outliers, variability, and sparsity of peptide intensities on abundance estimations. The figure shows $\log_2$ transformed quantile normalized peptide intensities for the yeast null protein CG121 from the CPTAC data set for spike-in conditions 6A, 6B, 6C, 6D, and 6E. Each color denotes a different condition. Connected crosses: average protein $\log_2$ intensity estimates for each condition are provided for a traditional protein level workflow where the mean of the protein-level MaxLFQ values was calculated (MaxLFQ, blue), the estimates of the peptide-based regression model fitted with ordinary least squares (LM, black) and the estimates of the peptide based ordinary least squares fit after omitting the two lowest observations in condition 6B (LM-extremes, orange). In condition 6E there were not enough data points to provide a MaxLFQ protein-level estimate. Boxes denote the interquartile range (IQR) of the $\log_2$ transformed quantile normalized peptide intensities in each condition with the median indicated as a thick horizontal line inside each box. Whiskers extend to the most extreme data point that lies no more than 1.5 times the IQR from the box. Points lying beyond the whiskers are generally considered as outliers. Note, that the presence of two low-intensity peptide observations in concentration 6B has a strong effect on the estimates for both MaxLFQ and LM.

These issues illustrate that state-of-the-art analysis methods experience difficulties in coping with peptide imbalances that are inherent to DDA LFQ proteomics data. We here propose three modular improvements to deal with the problems of overfitting, sample-to-sample variability and outliers:

(1) Ridge regression, which penalizes the size of the model parameters. Shrinkage estimators can strongly improve reproducibility and overall performance as they have a lower overall mean squared error compared to ordinary least squares estimators (25–27).

(2) Empirical Bayes variance estimation, which shrinks the individual protein variances toward a common prior variance, hence stabilizing the variance estimation.

(3) M-estimation with Huber weights, which will make the estimators more robust toward outliers (28).

We illustrate our method on the CPTAC Study 6 spike-in data and a published ArgP knock-out *Francisella tularensis* proteomics experiment and show that our method provides more stable $\log_2$ FC estimates and a better DA ranking than competing methods.

9.1.4. Experimental procedures

CPTAC Spike-in Data Set

The publicly available Study 6 of the Clinical Proteomic Technology Assessment for Cancer (29) is used to evaluate the performance of our method. Raw data can be accessed at https://cptac-data-portal.georgetown.edu/cptac/public?scope=Phase+I. In this study, the Sigma Universal Protein Standard mixture 1 (UPS1, Sigma-Aldrich, St. Louis, MO) containing 48 different human proteins was spiked into a 60 ng protein/µl *Saccharomyces cerevisiae* strain BY4741 (MATa, leu2Δ0, met15Δ0, ura3Δ0, his3Δ1) lysate in five different concentrations (6A: 0.25 fmol UPS1 proteins/µl; 6B: 0.74 fmol UPS1 proteins/µl; 6C: 2.22 fmol UPS1 proteins/µl; 6D: 6.67 fmol UPS1 proteins/µl; and 6E: 20 fmol UPS1 proteins/µl). These samples were sent to five independent laboratories and analyzed on seven different instruments. For convenience, we limited ourselves to the data originating from the LTQ-Orbitrap at site 86, LTQ-Orbitrap O at site 65 and LTQ-Orbitrap W at site 56. Samples were run three times on each instrument. The used dataset thus features five different samples, each analyzed in threefold on three different instruments. Raw data files were searched using MaxQuant version 1.5.2.8 (30) with the following settings. As variable modifications we allowed acetylation (protein N terminus), methionine oxidation (to methionine-sulfoxide) and N-terminal glutamine to pyroglutamate conversion. As a fixed modification, we selected carbamidomethylation on cysteine residues as all samples were treated with iodoacetamide. We used the enzymatic rule of trypsin/P with a maximum of 2 missed cleavages and allowed MaxQuant to perform matching between runs with a match time window of 0.7 min and an alignment time window of 20 min. The main search peptide tolerance was set to 4.5 ppm and the ion trap MS/MS match tolerance was set to 0.5 Da. Peptide-to-spectrum match level was set at 1% FDR with an additional minimal Andromeda score of 40 for modified peptides as these settings are most commonly used by researchers. Protein FDR was set at 1% and estimated by using the reversed search sequences. We performed label-free quantitation with MaxQuant's standard settings. The maximal number of modifications per peptide was set to 5. As a search FASTA file we used the 6718 reviewed proteins present in the *Saccharomyces cerevisiae* (strain ATCC 204508/S288c) proteome downloaded from Uniprot at March 27, 2015 supplemented with the 48 human UPS1 protein sequences. Potential contaminants present in the contaminants.fasta file that comes with MaxQuant were automatically added to the search space by the software. For protein quantification in the proteinGroups.txt file, we used unique and razor peptides and allowed all modifications as all samples originate in essence from the same yeast lysate and the same UPS1 spike-in sample.

Francisella tularensis Data Set

The data of Ramond *et al.* (31) is used to illustrate our method on a real biological experiment. Both raw and processed data are publicly available and can be found in the PRIDE repository at http://www.ebi.ac.uk/pride/archive/projects/PXD001584. The authors explored changes in the proteome of the facultative intracellular pathogenic coccobacillus *Francisella tularensis* after gene deletion of a newly identified arginine transporter, ArgP. Both wild-type and ArgP mutants were grown in biological triplicate. Each biological replicate was analyzed in technical triplicate via label-free LC-MS/MS. Data were processed with MaxQuant version 1.4.1.2 and potential contaminants and reverse sequences were removed. In addition, only proteins present with at least two peptides in at least 9 out of the 18 replicates were retained. Subsequent data analysis via t-tests on imputed LFQ intensities was performed.

Summarization-based Analysis

MaxLFQ+Perseus

This is a standard summarization-based analysis pipeline that is available in the popular MaxQuant-Persues software package (21). Briefly, the MaxQuant ProteinGroups.txt file was loaded into Perseus version 1.5.1.6, potential contaminants that did not correspond to any UPS1 protein as well as reversed sequences and proteins that were only identified by site (thus only by a peptide carrying a modified residue) were removed from the data set. MaxLFQ intensities (32) were $\log_2$ transformed and pairwise comparisons between conditions were done via *t*-tests.

MaxLFQ+limma

The MaxQuant ProteinGroups.txt file is used as input for R version 3.1.2 (Pumpkin Helmet) (33). Potential contaminants and reversed sequences (see above) were removed from the data set. The MaxLFQ intensities were $\log_2$ transformed and analyzed in limma, an R/Bioconductor package for the analysis of microarray and next-generation sequencing data (34). Limma makes use of posterior variance estimators to stabilize the naive variance estimator by borrowing strength across proteins (see also below).

Peptide-based Model Analysis

Data Preprocessing

MaxQuant's peptides.txt file was read into R version 3.1.2, the peptide intensities were $\log_2$ transformed and quantile normalized (35, 36). Many other normalization approaches do exist, however, comparing them is beyond the scope of this paper (24, 36–38). Reversed sequences and potential contaminants were removed from the data. For the CPTAC dataset, we only removed potential contaminants that did not map to any UPS1 protein. MaxQuant assigns proteins to protein groups using an Occam's razor approach. However, to avoid the added complexity of proteins mapped to multiple protein groups, we discarded peptides belonging to protein groups that contained one or more proteins that were also present in a smaller protein group. Next, peptides were grouped per protein group in a data frame. Finally, values belonging to peptide sequences that appeared only once were removed as the model parameter for the peptide effect for these sequences is unidentifiable. For notational convenience, a unique protein or protein group is referred to as a protein in the remainder of this article.

Benchmark Peptide-based Model

We start from the peptide-based linear regression models as proposed by Daly *et al.* (39) Clough *et al.* (22) and Karpievitch *et al.* (40), of which we have independently proven their superior performance compared to summarization-based workflows (21). In general, the following model is proposed:

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

(Eq. 1)

with y_{ijklmn} the n^{th} $\log_2$-transformed normalized feature intensity for the i^{th} protein under the j^{th} treatment (*treat),* the k^{th} peptide sequence (*pep*), the l^{th} biological repeat (*biorep*) and the m^{th} technical repeat (*techrep*) and ε_{ijklmn} a normally distributed error term with mean zero and protein specific variance σ_i^2. The β's denote the effect sizes for *treat, pep, biorep* and *techrep* for the i^{th} protein.

Robust Ridge Model

Our novel approach improves the estimation of the model parameters in (Eq. 1) via three extensions: (1) ridge regression, which leads to shrunken yet more stable $\log_2$ fold change (FC) estimates, (2) Empirical Bayes estimation of the variance, which further stabilizes variance estimators, and (3) M-estimation with Huber weights, which reduces the impact of outlying peptide intensities. For the robust ridge model, degrees of freedom are calculated using the trace of the hat matrix.

1. Ridge Regression

The ordinary least squares (OLS) estimates for protein *i* are defined as the parameter estimates that minimize the following loss function:

$$\sum_{jklmn} e_{ijklmn}^2 = \sum_{jklmn} \left(y_{ijklmn} - \beta_{ij}^{treat} - \beta_{ik}^{pep} - \beta_{il}^{biorep} - \beta_{im}^{techrep}\right)^2$$

(Eq. 2)

With e_{ijklmn} the residual errors and X_{ijklmn} the row of the design matrix corresponding to observation y_{ijklmn}.

Ridge regression shrinks the regression parameters by imposing a penalty on their magnitude. The ridge regression estimator is obtained by minimizing a penalized least squares loss function:

$$\sum_{jklmn} e_{ijklmn}^2 + \lambda_i^{pep} \sum_k \beta_{ik}^{pep\,2} + \lambda_i^{biorep} \sum_l \beta_{il}^{biorep\,2} + \lambda_i^{techrep} \sum_m \beta_{im}^{techrep\,2}$$

(Eq. 3)

With each λ a ridge penalty for the parameter estimator $\hat{\beta}$ corresponding to an effect in Eq. 1. When the λs are larger than zero, the ridge estimators for $\hat{\beta}$ will be shrunken toward 0. This introduces some bias but reduces the variability of the parameter estimator, which makes shrinkage estimators theoretically more stable and more accurate (*i.e.* they tend to have a lower root mean squared error (RMSE) compared to the OLS estimator) (25–27). On the one hand, $\hat{\beta}$'s estimated by only a few observations will experience a strong correction toward 0, protecting against overfitting. On the other hand, $\hat{\beta}$'s that can be estimated based on many observations will exhibit a negligible bias because the ridge penalty will be dominated by the sum of the squared errors, which reflects that these $\hat{\beta}$'s can be estimated more reliably. We choose to tune each λ separately because the variability on the peptide effect seems generally much larger than the variability on the other effect terms. λ penalties can be tuned via cross-validation, but in this work, we exploit the link between mixed models and ridge regression (41) and estimate the penalties by implementing ridge regression within the lme4 package (42) in R. λ_{treat} then equals $\frac{\hat{\sigma}_i^2}{\hat{\sigma}^2_{\beta^{treat},i}}$ with $\hat{\sigma}_i$ the estimated residual variance and $\hat{\sigma}^2_{\beta^{treat},i}$ the estimated variance captured by the treatment effect for peptide intensities from protein *i* (Chapter 5, 41). The standard errors of the parameter estimators and contrasts of interest are based on the bias-adjusted variance estimator (Chapter 6, 41).

2. Empirical Bayes Variance Estimations

In the introduction we argued that data sparsity can lead to unstable variance estimates. In order to stabilize the residual variance estimation, we shrink the estimated protein specific

variances toward a pooled estimate over all proteins using an empirical Bayesian approach (43) implemented in the limma R/Bioconductor-package (44). In that way, the information contained in all proteins is borrowed to stabilize the variance estimates of proteins with few observations. Hence, the small variances will increase, while large variances will decrease. This avoids that proteins that exhibit a small FC and a tiny variance will appear highly significantly. Also, the number of degrees of freedom for all these so-called moderated t-tests will increase compared to normal *t*-tests.

3. M-estimation With Huber Weights

As we have shown in the introduction (see Fig. 9.2), outlying peptides might have a severe impact on the parameter estimates, especially in conditions with few identified features. We therefore propose to adopt M-estimation with Huber weights to diminish the impact of outlying observations. Combining ridge regression with M-estimation leads to the following penalized weighted least squares loss function:

$$\sum_{jklmn} w_{ijklmn} e_{ijklmn}^2 + \lambda_i^{treat} \sum_j \beta_{ij}^{treat\,2} + \lambda_i^{pep} \sum_k \beta_{ik}^{pep\,2} + \lambda_i^{biorep} \sum_l \beta_{il}^{biorep\,2} + \lambda_i^{techrep} \sum_m \beta_{im}^{techrep\,2}$$

(Eq. 4)

Herein, w_{ijklmn} is a Huber weight that weighs down observations with high residuals.

False Discovery Rate

For both peptide-based models and MaxLFQ+limma, *p* values are adjusted for multiple testing with the Benjamini-Hochberg FDR procedure (45). Perseus uses a permutation-based FDR that turns out to be very close to the Benjamini-Hochberg procedure. The FDR is controlled at the 5% level in all analyses.

9.1.5. Results

We decided to compare our method to three competing methods using data from the CPTAC spike-in benchmark study. Additionally, its advantages are demonstrated in a case study that was originally analyzed with a standard summarization-based approach. For peptide-based models the use of MaxQuant's peptides.txt file, which contains summed up peptide level data, slightly increased the discriminative power as compared to analyses using the MaxQuant's evidence.txt file, which contains individual intensities for each identified feature (supplemental Fig. S7, File S1). Therefore, all peptide-based models are based on the peptides.txt file.

1. Evaluation Using Data From a Spike-in Benchmark Study

The performances of the different methods are assessed by comparing different spike-in concentrations in the CPTAC Study 6 data set. This data set consists of identical samples containing a trypsin-digested *Saccharomyces cerevisiae* proteome spiked with different concentrations (0.25, 0.74, 2.22, 6.67, and 20 fmol/µl) of a trypsin-digested UPS1 mix containing 48 human proteins. As the high spike-in concentrations are known to suffer from ionization competition effects (18, 21), we focus mainly on comparing 6B-6A, 6C-6A and 6C-6B, which have the lowest spike-in concentrations.

We compared our robust ridge approach to three competing approaches: (1) MaxLFQ-Perseus, (2) MaxLFQ-limma, and (3) a petide based linear model (LM). (1) MaxLFQ-Perseus

is a standard summarization-based approach where MaxLFQ normalized $\log_2$ protein intensities are compared between conditions by FDR-corrected t-tests. (2) MaxLFQ-limma is a summarization-based approach in which MaxLFQ protein level data are analyzed via limma (34), an R/Bioconductor package that can stabilize variance estimation by borrowing information across proteins via an empirical Bayesian approach. (3) The linear model (LM) is a peptide-based linear regression model (Eq. 1) that is known to outperform summarization-based approaches (21). This model contains a treatment effect, a peptide effect and an instrument effect, and its structure is motivated in supplemental File S1. Note, that our robust ridge method (RR) is an improvement of the LM approach by implementing ridge regression, empirical Bayes variance estimation and M-estimation with Huber weights. We compared the results of the four methods in terms of precision and accuracy of the $\log_2$ fold change estimates, as well as sensitivity and specificity.

Precision and Accuracy

Log_2 fold change (FC) estimates using our robust ridge model for yeast null proteins are clearly more precise compared to the other methods; the interquartile range of the $\log_2$ FC estimates is on average three times smaller compared to the LM fit and 4.5 times smaller for comparison 6B-6A; Fig. 9.3). The accuracy is also increased as most DA estimates for yeast null proteins are estimated very close to zero. This is due to an effect of the ridge penalty, which strongly shrinks the estimates for null proteins identified by a few peptides toward zero. Fig. 9.3 also shows a general trend for each method: as spike-in concentration differences increase, a negative bias of the $\log_2$ FC estimates appears. This likely reflects ionization suppression effects (18, 21). Note that for our method, the $\log_2$ FC distributions in all comparisons (including 6B *versus* 6A) in Fig. 9.3 are skewed toward negative fold changes and that the skewness increases with higher spike-in concentrations, suggesting that ionization suppression effects already occur at very low spike-in concentrations.

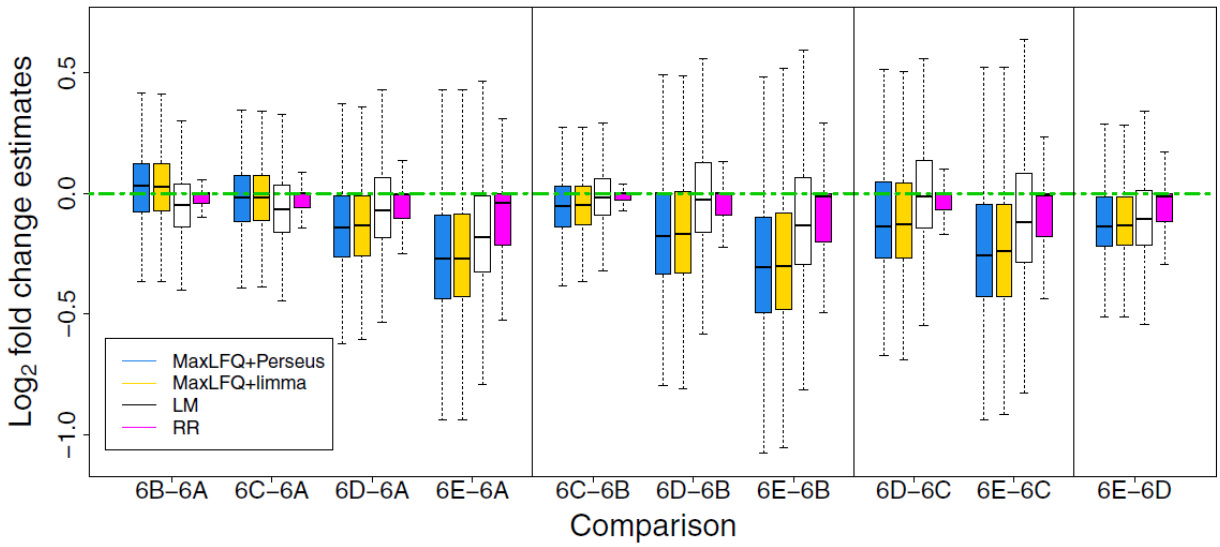


Figure 9.3. Precision and accuracy of fold change (FC) estimates for null proteins in the CPTAC study. The boxplots show the distributions of the FC estimates of the null yeast proteins for each of the ten comparisons for 4 different approaches. Outliers (here defined as data points that lie more than 1.5 times the interquartile range from the box) are not shown. The horizontal dotted green line denotes the true $\log_2$ fold change for the yeast proteins ($\log_2$ FC = 0). Blue (MaxLFQ+Perseus): protein-level analysis consisting of MaxLFQ normalization followed by *t*-tests in Perseus, yellow (MaxLFQ+limma): protein-level analysis consisting of MaxLFQ normalization followed by limma analysis, black (LM): peptide-based linear regression model containing treatment, peptide and instrument effects, purple (RR): peptide-based ridge regression model containing treatment, peptide and instrument effects with

empirical Bayes variance estimator and M-estimation with Huber weights. An identical figure with outliers is provided in supplemental Fig. S12, File S1.

When assessing the DA UPS1 proteins (Fig. 9.4), the $\log_2$ FC estimates for the summarization-based approaches MaxLFQ-Perseus and MaxLFQ-limma are always more biased and more variable compared to the LM and RR methods, except in condition 6B-6A, where MaxLFQ-limma has the lowest bias of all approaches. Median $\log_2$ FC estimates for UPS proteins are very comparable between our RR and the LM method. For eight out of ten comparisons, the RR estimates are even closer to the true $\log_2$ FC. The interquartile range of the $\log_2$ FC estimates of the DA proteins for RR is on average 1.2 times smaller compared to those of the LM model. Thus, shrinkage estimation does not negatively affect estimates for proteins with a strong evidence for DA.

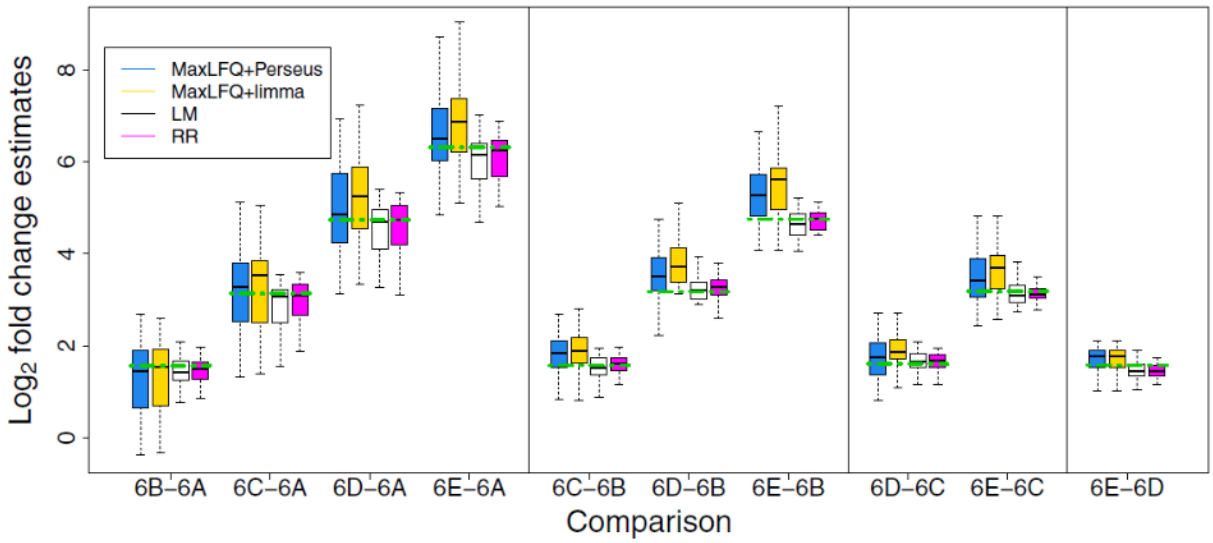


Figure 9.4. Precision and accuracy of fold change (FC) estimates for differential abundant proteins in the CPTAC study. The boxplots show the distributions of the FC estimates of the spiked-in UPS1 proteins for each of the ten comparisons for four different approaches. Outliers (here defined as data points that lie more than 1.5 times the interquartile range from the box) are not shown. The horizontal dotted green lines denote the true $\log_2$ FC for the UPS1 proteins in each comparison. Blue (MaxLFQ+Perseus): protein-level analysis consisting of MaxLFQ normalization followed by *t*-tests in Perseus, yellow (MaxLFQ+limma): protein-level analysis consisting of MaxLFQ normalization followed by limma analysis, black (LM): peptide-based linear regression model containing treatment, peptide and instrument effects, purple (RR): peptide-based ridge regression model containing treatment, peptide and instrument effects with empirical Bayes variance estimator and M-estimation with Huber weights. An identical figure with outliers is provided in supplemental Fig. S13, File S1.

Sensitivity and Specificity

The sensitivity and specificity of the different methods are assessed using receiver operator characteristics (ROC) curves (Fig. 9.5, Tables 9.1 and 9.2). For the summarization-based approaches, it turns out that MaxLFQ+limma outperforms MaxLFQ+Perseus. This is not surprising, as it has been shown that limma outperforms standard t-tests, also in proteomics data sets (46). As we have shown before, both peptide-based regression models outperform the summarization-based approaches (21). Our RR method further improves on the LM model in comparison 6B-6A, in which the detection of DA is most challenging since it involves the two lowest spike-in concentrations. For all other comparisons LM and RR have a similar performance (Table 9.2), which was expected because the ROC curves of the LM method for these comparisons are already very steep.

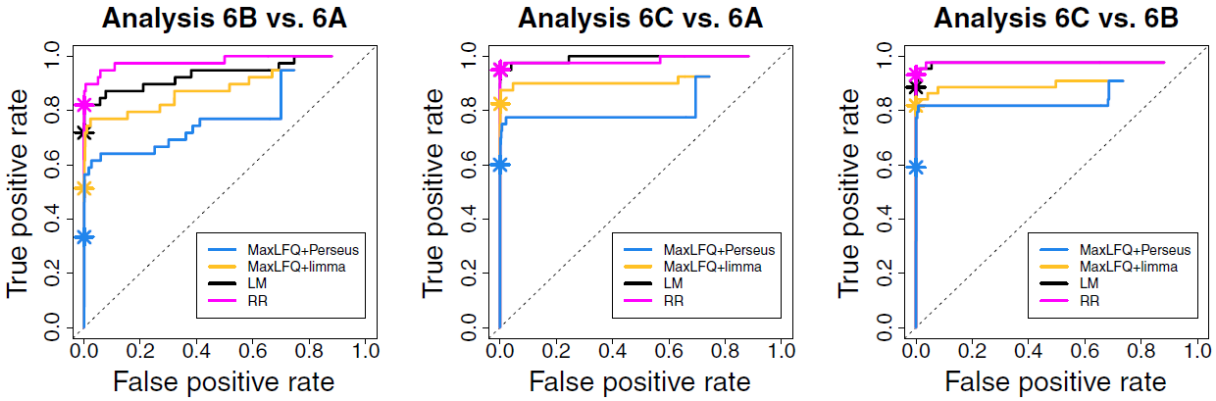


Figure 9.5. Comparison of sensitivity and specificity. Receiver operator characteristic (ROC) curves show the superior performance of our robust ridge approach compared to other standard data analysis techniques for comparisons of conditions 6B-6A, 6C-6A, and 6C-6B in the CPTAC spike in study. Stars denote the cut-offs at an estimated 5% FDR level. Blue (MaxLFQ+Perseus): protein-level analysis consisting of MaxLFQ normalization followed by *t*-tests in Perseus, yellow (MaxLFQ+limma): protein-level analysis consisting of MaxLFQ normalization followed by limma analysis, black (LM): peptide-based linear regression model containing treatment, peptide and instrument effects, purple (RR): peptide-based ridge regression model containing treatment, peptide and instrument effects with empirical Bayes variance estimator and M-estimation with Huber weights.

Table 9.1. Total areas under the curve (AUC) for three standard approaches and our robust ridge method for comparisons 6B-6A, 6C-6A and 6C-6B in the CPTAC spike-in study. MaxLFQ+Perseus: protein-level analysis consisting of MaxLFQ normalization followed by t-tests in Perseus, MaxLFQ+limma: protein-level analysis consisting of MaxLFQ normalization followed by limma analysis, LM: peptide-based linear regression model containing treatment, peptide and instrument effects, RR: peptide-based ridge regression model containing treatment, peptide and instrument effects with empirical Bayes variance estimator and M-estimation with Huber weights.

Comparison	MaxLFQ+Perseus	MaxLFQ+limma	LM	RR
6B-6A	0.536	0.634	0.817	0.862
6C-6A	0.583	0.672	0.877	0.869
6D-6A	0.583	0.680	0.880	0.878
6E-6A	0.564	0.660	0.883	0.880
6C-6B	0.607	0.655	0.860	0.861
6D-6B	0.618	0.657	0.858	0.859
6E-6B	0.601	0.644	0.863	0.863
6D-6C	0.654	0.680	0.856	0.860
6E-6C	0.663	0.678	0.864	0.864
6E-6D	0.683	0.685	0.836	0.837

Table 9.2. Partial areas under the curve (pAUC) for a false positive rate (FPR) < 0.1 for three standard approaches and our robust ridge method by comparing conditions 6B-6A, 6C-6A and 6C-6B in the CPTAC spike-in study. MaxLFQ+Perseus: protein-level analysis consisting of MaxLFQ normalization followed by t-tests in Perseus, MaxLFQ+limma: protein-level analysis consisting of MaxLFQ normalization followed by limma analysis, LM: peptide-based linear regression model containing treatment, peptide and instrument effects, RR: peptide-based ridge regression model containing treatment, peptide and instrument effects with empirical Bayes variance estimator and M-estimation with Huber weights.

Comparison	MaxLFQ+Perseus	MaxLFQ+limma	LM	RR
6B-6A	0.061	0.075	0.083	0.091
6C-6A	0.076	0.088	0.096	0.097
6D-6A	0.077	0.089	0.098	0.097
6E-6A	0.076	0.089	0.097	0.097
6C-6B	0.081	0.086	0.096	0.097
6D-6B	0.082	0.086	0.096	0.096
6E-6B	0.081	0.087	0.096	0.096
6D-6C	0.076	0.083	0.094	0.095
6E-6C	0.088	0.090	0.095	0.095
6E-6D	0.090	0.092	0.092	0.093

FDR Control

None of the adopted methods are able to control the true FDR at the nominal 5% level in the majority of the comparisons (Table 9.3). MaxLFQ+Perseus can only control the FDR at the 5% level in comparisons 6C-6A and 6C-6B. MaxLFQ+limma only controls the FDR accurately in comparison 6C-6B and both LM and RR can control the FDR only in comparisons 6B-6A and 6C-6B. When comparing RR and LM, RR does a better job in controlling the FDR in comparisons 6B-6A, 6C-6A, 6D-6B and 6D-6C, but not for the other comparisons.

Table 9.3. Observed FDR when using a 5% FDR cut-off level for each of the three standard approaches and our robust ridge method for all pairwise comparisons between conditions 6A, 6B, 6C, 6D and 6E. MaxLFQ+Perseus: protein-level analysis consisting of MaxLFQ normalization followed by t-tests in Perseus, MaxLFQ+limma: protein-level analysis consisting of MaxLFQ normalization followed by limma analysis, LM: peptide-based linear regression model containing treatment, peptide and instrument effects, RR: peptide-based ridge regression model containing treatment, peptide and instrument effects with empirical Bayes variance estimator and M-estimation with Huber weights. This table shows that most methods are unable to control the FDR at 5%, especially for comparisons involving higher spike-in concentrations (e.g. 6D and 6E).

Comparison	MaxLFQ+Perseus	MaxLFQ+limma	LM	RR
6B-6A	0.071	0.048	0.034	0.030
6C-6A	0.040	0.083	0.095	0.050
6D-6A	0.456	0.794	0.466	0.050
6E-6A	0.917	0.922	0.870	0.883
6C-6B	0.037	0	0	0.024
6D-6B	0.640	0.879	0.528	0.494
6E-6B	0.918	0.924	0.863	0.870
6D-6C	0.429	0.799	0.481	0.434
6E-6C	0.886	0.906	0.842	0.848
6E-6D	0.321	0.584	0.386	0.561

2. Case Study

We further illustrate the performance of our novel method on true biological data in which a single trigger was expected to have an impact on several tightly regulated, but highly interconnected pathways. Thus, contrary to a spike-in data set where differential abundance typically heads in one direction (either up or down-regulated) and stays limited to the spiked-in proteins, a biological dataset consists of a plethora of both strongly as well as weakly differentially regulated proteins. Moreover, in the CPTAC data set, the same sample was always used to spike in different amounts of UPS1 proteins in the same yeast background instead of isolating a new yeast proteome each time. Hence, variability in biological repeats will be much larger compared to the spike-in data set. Although detection of differential abundance in real biological data is the ultimate goal, there is no known ground truth available for these data sets and our evaluation is based on visual inspection of selected findings.

We made use of the publicly available data published by Ramond *et al.* (31) in which the authors compared the proteome of *Francisella tularensis* mutant for the arginine transporter ArgP to wild-type (WT) bacteria in biological triplicate. For each biological repeat, three technical replicates were also available. We compared the authors' results with the results of our RR method based on the authors-supplied peptides.txt instead of re-searching the raw spectra. In their study, Ramond *et al.* (31) performed the analysis at the protein level using intensities of at least two different peptides and keeping those proteins with at least 9 out of 18 valid values, and as such analyzed 842 proteins in total. With our approach—analysis at the peptide level and filtering out peptides that appeared only once in the data set—we were able to analyze a total of 989 proteins. Both the results of our ranking as well as the ranking from the original *Francisella* article can be found in supplemental File S2. When we compared the top 100 proteins with the lowest *p* values in both methods, only 52 proteins overlapped. Ramond *et al.* (31) found 309 DA proteins at the 5% FDR level, whereas we only found 159 proteins significantly DA at the same FDR level. Thus, our method appears to be more conservative.

We evaluated the differential abundance of the ten proteins that were present in the RR DA list, but not in the original DA list, as well as the ten highest ranked proteins from the original DA list missing in our list. The ten proteins that were only discovered with our method were all lost during the preprocessing procedure of Ramond *et al.* (31). Among those proteins, six are more highly abundant in the mutant: exodeoxyribonuclease V subunit gamma and ABC transporter membrane protein, as well as four hypothetical proteins (the membrane protein FTN_0835, an AAA^+ superfamily member FTN_0274, an alpha/beta hydrolase FTN_0721 and FTN_1244). Four out of the ten proteins that were only discovered by our method have a lower abundance in the mutant. These proteins are Radical SAM superfamily protein, DNA helicase II, C32 tRNA thiolase and the hypothetical protein FTN_0400. Log_2 intensity plots of the individual peptide intensities suggest a differential abundance for most of these proteins (supplemental Figs. S14-S23, File S1).

Eight out of the ten highest-ranked proteins in the original DA list all seem to be highly abundant gauged by the number of peptides identified (supplemental Figs. S24-S33, File S1). Further inspection reveals that most of these proteins do not appear to bear strong evidence for DA between WT and mutant. Except for the hypothetical protein FTN_1397 and the Mur ligase family protein, all variance estimates are smaller than 0.01, which leads to extreme T statistics. Empirical Bayesian variance estimation does increase these small variances, but not enough to make them insignificant at the 5% significance level. There seems to be a relatively clear effect for hypothetical protein FTN_1397 (supplemental Fig. S31, File S1), which in our method just did not pass the 5% FDR level (*p* value of $9.4*10^{-3}$ and FDR adjusted *p* value 0.057). The Mur ligase family protein shows no strong visual DA (supplemental Fig. S32, File S1), but combines a moderate DA estimate (-0.34) with a small variance (0.01). The two other proteins only present in the list of Ramond *et al.* (31) are rather low-abundant. DNA-binding protein HU-beta (supplemental Fig. S30, File S1) shows no visual evidence for DA at all, but hypothetical protein FTN_1199 (supplemental Fig. S33, File S1) might possess weak evidence for DA.

Ramond *et al.* (31) also noted that several protein modules had an increased enrichment in DA proteins. In fact, all ribosomal proteins and all proteins involved in branched-chain amino acid (BCAA) synthesis were either unchanged (as the *p* value did not reach the 5% FDR cut-off) or present in lower amounts in the mutant. Based on their data, one could also suspect an up-regulation of several tricarboxylic acid (TCA) cycle proteins (nine out of 12) in the mutant.

Ribosomal Proteins

It is known that a global down-regulation of ribosome synthesis is a common response to nutrient starvation in Bacteria and Eukarya (47, 48). One might thus expect a similar drop in abundance in the arginine transporter-mutated *F. tularensis* as its delay in phagosomal escape can be fully restored by supplementation of the medium with excess arginine (31). When considering all 30S and 50S ribosomal proteins (both significant and insignificant in terms of DA), both our method and the original article report log_2 FC estimates for 49 ribosomal proteins. As expected, all log_2 FC estimates from our method pointed toward down-regulation in the mutant. Contrary, in the analysis of Ramond *et al.* (31), two proteins, ribosomal protein L7/L12 and 30S ribosomal protein S18 showed, although deemed insignificant, quite large log_2 FCs (0.12 resp. 0.54) in favor of up-regulation in the mutant. This again suggest a more stable log_2 FC estimation by our method (supplemental Fig. S34, File S1).

BCAA Synthesis Proteins

Based on KEGG, we could only identify nine proteins as BCAA synthesis proteins although the authors report on 12 BCAA synthesis proteins. As we were unable to find the remaining 3 proteins in this pathway, we further focus on nine proteins only. Supplemental Fig. S35, File

S1 shows the distribution of the $\log_2$ FC estimates for both the methodology of Ramond *et al.* (31) and our method. Here, one insignificant protein, dihydroxy-acid dehydratase shows a $\log_2$ FC of 0.22. Using our method, all $\log_2$ FCs indicate either down-regulation in the mutant (*i.e.* $\log_2$ FCs smaller than 0) or no change at all ($\log_2$ FCs smaller than $1*10^{-17}$). On average, $\log_2$ FC estimates for BCAA synthesis proteins are more negative in our method. Thus, our method provides stronger evidence for down-regulation of some BCAA synthesis proteins in the mutant compared to the method of the original article.

TCA Cycle Proteins

Based on KEGG, we identified 13 TCA cycle proteins, while Ramond *et al.* (31) reported 12 proteins. When assessing the $\log_2$ FC estimates of these authors for the 13 proteins, 10 TCA cycle proteins appear to be up-regulated in the mutant (supplemental Fig. S36, File S1), 8 of which are declared significant. Contrary, in our method, only 1 protein of the TCA cycle (2-oxoglutarate dehydrogenase complex, E2 component, dihydrolipoyltranssuccinase) is found significant. Strikingly, 8 out of 13 TCA cycle proteins even show $\log_2$ FC estimates that are in absolute value smaller than 1×10^{-9}. We therefore zoomed in onthe individual $\log_2$ FC estimates of the peptides mapping to these proteins (supplemental Figs. S37-S50, File S1).

2-oxoglutarate dehydrogenase complex, E2 component, dihydrolipoyltranssuccinase (supplemental Fig. S37, File S1) is the only TCA cycle protein that is found at significantly different levels by our analysis. It is also denoted as significant in the original paper's methodology. Nonetheless, the evidence does not seem to be very strong. 2-oxoglutarate dehydrogenase E1 component, dihydrolipoamide acetyltransferase and succinate dehydrogenase iron-sulfur subunit are also denoted as significant in the original analysis with *p* values of 5.3×10^{-10}, 8.1×10^{-4} and 1.1×10^{-3} respectively, although the spectral evidence also appears quite weak (supplemental Figs. S38, S40, and S41, File S1). In our opinion, any visual evidence for differential abundance for malate dehydrogenase, aconitate hydratase, succinyl-CoA synthetase, alpha subunit and isocitrate dehydrogenase is negligible. Nonetheless, *p* values corresponding to DA for these proteins are estimated at 7.4×10^{-4}, 3.6×10^{-3}, 0.02 and 0.02 respectively in the paper of Ramond *et al.* (31) (supplemental Figs. S42, S43, and S47, File S1).

9.1.6. Discussion

In this work, we introduced three extensions to existing peptide-based linear models that significantly improve stability and precision of fold change estimates. These extensions include minimizing a penalized least squares loss function (ridge regression), weighing down outliers via M-estimation with Huber weights and variance stabilization via empirical Bayes. Our estimation approach is inevitably computationally more complex than the linear regression model, albeit much faster than fully Bayesian approaches that have to be fitted by computationally intensive Markov chain Monte Carlo algorithms (49). In this contribution, we normalized $\log_2$ transformed peptide intensities using quantile normalization. Many other types of transformation and normalization exist and can be adopted prior to applying our method. Our focus however, is on robust estimation procedures for peptide-based linear models. Therefore, a thorough comparison of preprocessing methods is beyond the scope of this paper.

We compared our novel estimation method to three other methods: a standard protein-level summarization followed by t-tests (MaxLFQ+Perseus), a standard protein-level summarization followed by limma (MaxLFQ+limma) and a peptide-based linear regression model. Evaluation was done based on the CPTAC Study 6 data set, where the ground truth is known as well as on a biological data set, where ArgP mutated *versus* wild-type *Francisella tularensis* proteomes

are compared. For the CPTAC dataset, we found our method to give more precise and more stable $\log_2$ FC estimates for both the yeast null proteins and the spiked-in UPS1 proteins. Thus, our method sufficiently shrinks abundance estimates that are driven by data sparsity in the null proteins, while retaining good DA estimates when sufficient evidence is present (such as for most UPS1 proteins in the CPTAC spike-in study). These findings are supported by theory as ridge regression shrinkage indeed reduces the variability of the estimator and generates overall more precise FC estimates (lower overall RMSE compared to ordinary least squares estimators) (25–27). Furthermore, empirical Bayes variance estimation squeezes the individual residual variances of all models toward a pooled variance which stabilizes the variance estimation. Finally, M-estimation with Huber weights weakens the impact of individual outliers.

The systematic underestimation of the $\log_2$ FC estimates provided by our method, even in comparison 6B-6A (Fig. 3), suggests that ionization competition effects can already come into play at spike-in concentration levels of 0.74 fmol/µl (condition 6B). These effects might partly explain why none of the considered methods performs well in controlling the FDR at the nominal 5% level. When analyzing the *Francisella* dataset, RR is more conservative than the method used by Ramond *et al.* (31), which might suggest a better protection against false positives. Indeed, ionization competition effects are also relevant for true biological data sets as an increase in a number of highly abundant proteins in a certain condition might generate a downwards bias in peptide intensities corresponding to non-DA proteins in this condition. Researchers should thus carefully reflect whether a low $\log_2$ fold change is truly biologically relevant, as ionization suppression effects already seem to appear at low differences in concentration. Indeed, even when disregarding these effects, the abundance of a protein typically has to differ by a reasonable amount to be of interest to a researcher. Therefore, we suggest testing against a minimal $\log_2$ FC value that is biologically interesting in a particular experiment; *e.g.* 0.5 or 1 (50). A similar approach can be easily adopted within our framework, but we have chosen to test against an FC of 0 to make our results comparable with the analysis of Ramond *et al.* (31).

In practice, only a handful of true positives will typically be selected for further experimental validation. Therefore, the ranking that is produced by a method is more important than its capability to accurately control the FDR at the 5% level. Here, RR produces superior ranking lists compared to all other methods except in comparisons 6D-6A, 6E-6A and 6E-6B, where large ionization suppression effects are expected due to huge spike-in concentration differences. Hence, it will be very difficult to discern the ionization bias from real DA for these comparisons. Each component of our model contributes to an improvement in performance. ROC curves in supplemental Fig. S4, File S1 show that empirical Bayesian variance estimation slightly but consistently improves the performance of the LM model, while M-estimation seems to cause the largest gain (supplemental Fig. S5, File S1). Ridge regression clearly improves the LM model, EB variance estimation slightly improves the ridge regression model while M-estimation again seems to deliver the largest gain in performance. Corresponding AUC and pAUC values can be found in (supplemental Tables S7-S10, File S1).

In the CPTAC case study, we also confirmed that a peptide-based model starting from summed up intensities over different charge states and modifications but corresponding to identical peptide sequences in the same sample (peptides.txt file) tends to have a higher discriminative ability than a model starting from the individual feature intensities (evidence.txt file, see supplemental Fig. S7, File S1). Others have also shown that analysis on the lowest level of summarization does not automatically lead to the best performance (18).

We also applied our method on a biological dataset and compared our results to the performance of the MaxLFQ summarization-based method used in the original publication (31). Proteins identified as DA by our method that were missed in the original publication were all filtered out because too few peptides were identified. Most of these proteins contain relatively strong evidence for DA based on visual inspection of the individual quantile normalized peptide intensities, illustrating that our method can reliably discover DA in proteins that suffer considerably from missing peptides across samples. Proteins which were denoted as DA in the original article but not by our method were typically identified by a lot of peptides but with a weak visual evidence for DA. Indeed, when a lot of peptide intensities are present, modeling at the peptide level does not contribute much to stabilize the overall $\log_2$ FC estimate. Mostly though, these small fold changes are not biologically relevant. However, when one seeks DA in sparse data, as is the case in most experiments, our method clearly outperforms classical approaches. We indeed identify possibly interesting effects in low abundant proteins without overfitting, omitting the need for extensive *a priori* data filtering (*e.g.* dropping proteins with less than two peptides present in at least nine out of 18 samples as done by Ramond *et al.* (31)). Consequently, our ability to detect DA in low abundant proteins combined with robustness against irrelevant changes in high abundant proteins is a favorable property.

We also showed that our fold change estimates are more stable for both the ribosomal and the BCAA synthesis modules, which are denoted as DA by Ramond *et al.* (31). For TCA cycle proteins, nine out of 12 were described as significantly more abundant in the mutant by these authors, while in our method, eight out of 13 TCA cycle proteins have $\log_2$ FC estimates that are in absolute value smaller than 1×10^{-9}. Upon visual inspection, most of these proteins indeed contain very limited evidence for DA. Our method thus appears to provide more reliable data for follow-up analysis, thereby aiding researchers in drawing more correct conclusions. The fact that almost 82% of the proteins in our DA list of *Francisella* proteins indeed contain less peptides in the condition with the lowest abundance of this protein, advocates the inclusion of imputation or censoring approaches, or even the combination of our method with estimates derived from spectral counting, which could be used as a rough but very simple validation technique. For example, hypothetical protein FTN_0400 could be a false positive because more peptides are identified in the mutant, although it appears to be less abundant (supplemental Fig. S18, File S1). This could be an indication of intensity-dependent censoring in the WT. Just like the peptide-based linear regression model, our RR model can handle missing values. The models we presented here assume missingness completely at random, an assumption that is flawed when analyzing shotgun proteomics data. Note, however, that peptide-based models partially correct for this by incorporating peptide-specific effects. Moreover, our strategies to improve robustness of the estimators can be easily plugged into censored regression methods or estimation approaches that adopt advanced imputation techniques for handling missing data (40).

Another interesting outlook is to model all proteins together by incorporating pathway or module level effects in the model in order to make stronger inferences on individual proteins belonging to a certain pathway. Finally, we want to stress the importance of data exploration. Plots of $\log_2$ peptide intensities for proteins that are flagged as differential abundant are very useful for assessing the biological relevance and the degree of belief one can have in the DA proteins that are returned by a method.

We are currently preparing an R/Bioconductor package for our method. Meanwhile, all code and data needed to repeat the data analysis in this manuscript is available in Supplemental File 3.

9.1.7. Footnotes

Author contributions: L.J.G. and L.C. designed research; L.J.G. and L.C. performed research; L.J.G. analyzed data; L.J.G., K.G., and L.C. wrote the paper.

* This research was supported in part by IAP research network “StUDyS” grant no. P7/06 of the Belgian government (Belgian Science Policy) and the Multidisciplinary Research Partnership “Bioinformatics: from nucleotides to networks” of Ghent University. L.G. is supported by a Ph.D. grant from the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen) entitled “Differential proteomics at peptide, protein and module level” (141573).

This article contains supplemental Files S1 to S3.

9.1.8. References

1. Oda Y., Huang K., Cross F. R., Cowburn D., and Chait B. T. (1999) Accurate quantitation of protein expression and site-specific phosphorylation. Proc. Natl. Acad. Sci. U.S.A. 96, 6591–6596

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

3. Gygi S. P., Rist B., Gerber S. A., Turecek F., Gelb M. H., and Aebersold R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat. Biotech. 17, 994–999

4. Hsu J.-L., Huang S.-Y., Chow N.-H., and Chen S.-H. (2003) Stable-isotope dimethyl labeling for quantitative proteomics. Anal. Chem. 75, 6843–6852

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

6. Thompson A., Schäfer J., Kuhn K., Kienle S., Schwarz J., Schmidt G., Neumann T., and Hamon C. (2003) Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal. Chem. 75, 1895–1904

7. Bantscheff M., Schirle M., Sweetman G., Rick J., and Kuster B. (2007) Quantitative mass spectrometry in proteomics: a critical review. Anal. Bioanal. Chem. 389, 1017–1031

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

9. Rodriguez J., Gupta N., Smith R. D., and Pevzner P. A. (2008) Does trypsin cut before proline? J. Proteome Res. 7, 300–305

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

11. Schliekelman P., and Liu S. (2013) Quantifying the effect of competition for detection between coeluting peptides on detection probabilities in mass-spectrometry-based proteomics. J. Proteome Res. 13, 348–361

12. Venable J. D., Dong M.-Q., Wohlschlegel J., Dillin A., and Yates J. R. (2004) Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat. Meth. 1, 39–45

13. Gillet L. C., Navarro P., Tate S., Röst H., Selevsek N., Reiter L., Bonner R., and Aebersold R. (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis. Mol. Cell. Proteomics 11, 1–17

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

15. Liu H., Sadygov R. G., and Yates J. R. (2004) A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics. Anal. Chem. 76, 4193–4201

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

17. Bantscheff M., Lemeer S., Savitski M., and Kuster B. (2012) Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. Anal. Bioanal. Chem. 404, 939–965

18. Milac T. I., Randolph T. W., and Wang P. (2012) Analyzing LC-MS/MS data by spectral count and ion abundance: two case studies. Statistics Interface 5, 75–87

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

20. Zhang Y., Fonslow B. R., Shan B., Baek M.-C., and Yates J. R. (2013) Protein analysis by shotgun/bottom-up proteomics. Chem. Rev. 113, 2343–2394

21. Goeminne L. J. E., Argentini A., Martens L., and Clement L. (2015) Summarization vs peptide-based models in label-free quantitative proteomics: Performance, pitfalls, and data analysis guidelines. J. Proteome Res. 14, 2457–2465

22. Clough T., Key M., Ott I., Ragg S., Schadow G., and Vitek O. (2009) Protein quantification in label-free LC-MS experiments. J. Proteome Res. 8, 5275–5284

23. Karpievitch Y. V., Dabney A. R., and Smith R. D. (2012) Normalization and missing value imputation for label-free LC-MS analysis. BMC Bioinformatics 13, S5.

24. Ting L., Cowley M. J., Hoon S. L., Guilhaus M., Raftery M. J., and Cavicchioli R. (2009) Normalization and statistical analysis of quantitative proteomics data generated by metabolic labeling. Mol. Cell. Proteomics 8, 2227–2242

25. Ahmed S. E., and Raheem S. M. E. (2012) Shrinkage and absolute penalty estimation in linear regression models. Computational Stat. 4, 541–553

26. Stein C. (1956) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and

Probability, Volume 1: Contributions to the Theory of Statistics, pp. 197–206, University of California Press, Berkeley, Calif.

27. Copas J. B. (1983) Regression, prediction and shrinkage. J. Roy. Statist. Soc.. 45, 311–354

28. Huber P. J. (1964) Robust estimation of a location parameter. The Annals of Mathematical Statistics. 35, 73–101

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

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

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

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

33. R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

34. Ritchie M. E., Phipson B., Wu D., Hu Y., Law C. W., Shi W., and Smyth G. K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47.

35. Amaratunga D., and Cabrera J. (2001) Analysis of data from viral DNA microchips. J. Am. Statist. Assoc. 96, 1161–1170

36. Bolstad B. M., Irizarry R. A., Åstrand M., and Speed T. P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19, 185–193

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

38. Rudnick P. A., Wang X., Yan X., Sedransk N., and Stein S. E. (2014) Improved normalization of systematic biases affecting ion current measurements in label-free proteomics data. Mol. Cell. Proteomics 13, 1341–1351

39. Daly D. S., Anderson K. K., Panisko E. A., Purvine S. O., Fang R., Monroe M. E., and Baker S. E. (2008) Mixed-effects statistical model for comparative LC-MS proteomics studies. J. Proteome Res. 7, 1209–1217

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

41. Ruppert D., Wand M. P., and Carroll R. J. (2003) Semiparametric Regression, Cambridge University Press, New York

42. Bates D M. M., Bolker BM, Walker S. (2014) lme4: Linear mixed-effects models using Eigen and S4. J. Statistical Software 67, 1–48

43. Lönnstedt I., and Speed T. (2002) Replicated microarray data. Statistica Sinica 12, 31–46

44. Smyth G. K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. 3, Article3

45. Benjamini Y., and Hochberg Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. Royal Statist. Soc. 57, 289–300

46. Schwämmle V., León I. R., and Jensen O. N. (2013) Assessment and improvement of statistical tools for comparative proteomics analysis of sparse data sets with few experimental replicates. J. Proteome Res. 12, 3874–3883

47. Rudra D., and Warner J. R. (2004) What better measure than ribosome synthesis? Genes Dev. 18, 2431–2436

48. Dressaire C., Redon E., Gitton C., Loubiere P., Monnet V., and Cocaign-Bousquet M. (2011) Investigation of the adaptation of Lactococcus lactis to isoleucine starvation integrating dynamic transcriptome and proteome information. Microbial Cell Factories 10, S18.

49. Henao R., Thompson J. W., Moseley M. A., Ginsburg G. S., Carin L., and Lucas J. E. (2012) Hierarchical factor modeling of proteomics data. Computational Advances in Bio and Medical Sciences (ICCABS), 2012 IEEE 2nd International Conference on, pp. 1–6

50. McCarthy D. J., and Smyth G. K. (2009) Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics 25, 765–771

9.1.9. Appendix

Ridge regression

The untransformed, preprocessed intensities for each peptide p in each run r are assumed to follow a log-normal distribution. After log-transformation, these intensities become normally distributed. In all generality, for each protein, we propose the following peptide-based regression model that has also been proposed by Daly *et al.* (2008) [1]:

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

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

We now want to introduce an extra penalization on the fixed effects beta by exploiting the link between ridge regression and mixed models (see section 4.2.4). Except for a fixed intercept

β^0, we penalize the parameters corresponding to each covariate group g by assuming: $\beta_{m_g}^{g} \sim \mathrm{N}\left(0, \sigma^2/\lambda_g\right)$ for $m_g = 1, \ldots, M_g$. Herein, g refers to the gth covariate and accounts for the fact that certain covariates are modeled with more than one parameter. For example, a treatment effect with three levels will be modeled with three dummy parameters $\beta_1^{\text{treatment}}$, $\beta_2^{\text{treatment}}$, and $\beta_3^{\text{treatment}}$ whereby $\beta_1^{\text{treatment}} + \beta_2^{\text{treatment}} + \beta_3^{\text{treatment}} = 0$. We also assume $\beta_p^{\text{peptide}} \sim \mathrm{N}\left(0, \sigma^2/\lambda_{\text{peptide}}\right)$. Therefore, the BLUP estimator for $\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\beta}^{\text{peptide}} \\ \boldsymbol{u}^{\text{run}} \end{bmatrix}$ can be written as follows:

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\beta}}^{\text{peptide}} \\ \widehat{\boldsymbol{u}}^{\text{run}} \end{bmatrix} = \left(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C} + \boldsymbol{B}\right)^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{y}$$

With $\boldsymbol{y} = \begin{bmatrix} y_{11} \\ \cdots \\ y_{1R} \\ \cdots \\ y_{pr} \\ \cdots \\ y_{P1} \\ \cdots \\ y_{PR} \end{bmatrix}$ and $\boldsymbol{C} = \begin{bmatrix} & \boldsymbol{x}_1^{\text{peptide}} & \\ \boldsymbol{x}_{11} & \cdots & \boldsymbol{x}_1^{\text{run}} \\ \cdots & & \cdots \\ \boldsymbol{x}_{1R} & \boldsymbol{x}_1^{\text{peptide}} & \boldsymbol{x}_R^{\text{run}} \\ \cdots & \cdots & \cdots \\ \boldsymbol{x}_{pr} & \boldsymbol{x}_p^{\text{peptide}} & \boldsymbol{x}_r^{\text{run}} \\ \cdots & \cdots & \cdots \\ \boldsymbol{x}_{P1} & \boldsymbol{x}_P^{\text{peptide}} & \boldsymbol{x}_1^{\text{run}} \\ \cdots & \cdots & \cdots \\ \boldsymbol{x}_{PR} & & \boldsymbol{x}_R^{\text{run}} \\ & \boldsymbol{x}_P^{\text{peptide}} & \end{bmatrix}$. Herein $\boldsymbol{x}_p^{\text{peptide}}$ is a row vector of dummies, for which the pth element is equal to 1 and all other elements equal to 0. $\boldsymbol{x}_r^{\text{run}}$ is a row vector of dummies with the rth element equal to 1 and all other elements equal to 0. $\boldsymbol{B}$ is an $(1 + M + P + R) \times (1 + M + P + R)$ diagonal matrix with diagonal elements $[0 \;\; \boldsymbol{g}\, \boldsymbol{p}\, \boldsymbol{r}]$, with $\boldsymbol{g}$ a vector of length M containing the λ_g that corresponds to each parameter estimate $\hat{\beta}_{m_g}^{g}$ for $m_g = 1, \ldots, M_g$ and $g = 1, \ldots, G$, $\boldsymbol{p}$ a vector of length P containing the λ_{peptide} that corresponds to each parameter estimate $\hat{\beta}_p^{\text{peptide}}$ for $p = 1, \ldots, P$ and $\boldsymbol{r}$ a vector of length R containing the $\frac{\hat{\sigma}_u^2}{\hat{\sigma}^2}$ that corresponds to each parameter estimate $\hat{u}_r^{\text{run}}$ for $r = 1, \ldots, R$.

Robust regression with M estimation

To robustify our procedure against outliers, we use a weighted maximum likelihood method with Huber weights, as proposed by Zhou (2009) [2].

$$\sum_{j=1}^{J} w_j \, \mathrm{l}\left(y_j, \boldsymbol{\beta}, \boldsymbol{u}\right),$$

with $j = 1, \ldots, J$ an indicator for observation. This weighted log-likelihood is solved iteratively. The mixed model is fitted while the weights are kept constant. Then, the weights are recomputed using Huber's weight function on the residuals scaled with the residual standard deviation. This procedure is repeated until convergence. After convergence, the weighted BLUP estimator is given by:

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\beta}}^{\text{peptide}} \\ \widehat{\boldsymbol{u}}^{\text{run}} \end{bmatrix} = \left(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{C} + \boldsymbol{B}\right)^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{y},$$

with $\boldsymbol{W} = \left[w_1 \ldots w_j \ldots w_J\right]\boldsymbol{I}_{J\times J}$ and w_1 to w_J the weights corresponding to these observations and $\boldsymbol{I}_{J\times J}$ a $J \times J$ unity matrix. Zhou (2009) [2] showed that the weighted BLUP estimator is better

than the unweighted one in terms of bias and efficiency when the data contains some outliers but provides the same asymptotic efficiency when the model is correctly specified. Robust M estimation with Huber weights has also been used to robustify the negative binomial model in the popular RNA sequencing R package EdgeR [3].

Empirical Bayes variance estimation

Finally, we robustify our inference with limma's empirical Bayes variance estimation (see section 4.2.2). In brief, limma assumes the following prior distribution on the error variance σ_i^2 for each protein i ($i = 1, \dots, I$):

$$\frac{1}{\sigma_i^2} \sim \frac{1}{d_0 \sigma_0^2} \chi^2_{d_0},$$

with σ_0^2 a prior variance and $\chi^2_{d_0}$ a χ^2 distribution with d_0 degrees of freedom. A maximum a posteriori residual standard deviation $\tilde{s}_i$ for each protein is given by:

$$\tilde{s}_i = \sqrt{\frac{d_i \hat{\sigma}_i^2 + d_0 \hat{\sigma}_0^2}{d_i + d_0}}$$

We then plug in this posterior residual standard deviation in the estimator for the standard deviation of the model parameter of interest, $\hat{\beta}^g_{m_g}$ (suppressing the indicator i for notational convenience):

$$\tilde{\sigma}_{\hat{\beta}^g_{m_g}} = \tilde{s} \sqrt{(\boldsymbol{C}^{\mathrm{T}} \boldsymbol{W} \boldsymbol{C} + \boldsymbol{B})^{-1}_{m_g, m_g}}$$

Herein, m_g, m_g denotes the m_gth diagonal element of the matrix. This enables statistical inference with a moderated t-test with $d_i + d_0$ degrees of freedom:

$$\tilde{t}_{im_g} = \frac{\hat{\beta}^g_{im_g}}{\tilde{\sigma}_{\hat{\beta}^g_{im_g}}}$$

Herein, d_i is calculated as $J - tr(\boldsymbol{H})$, with J the total number of observations and $\boldsymbol{H}$ the hat matrix, which is calculated as follows:

$$\boldsymbol{H} = \boldsymbol{C}\left(\boldsymbol{C}^{\mathrm{T}} \boldsymbol{W} \boldsymbol{C} + \boldsymbol{B}\right)^{-1} \boldsymbol{C}^{\mathrm{T}} \boldsymbol{W}$$

Implementation

MSqRob builds on the lme4 R package for parameter estimation and statistical inference [4]. Shrinkage on fixed effect parameters is obtained by encoding them as random effects. To allow for robust M-estimation, a loop is placed around the model fitting procedure: after model fitting, Huber weights are calculated on the residuals scaled with the residual standard deviation. These weights are provided as arguments to the lmer function of the lme4 package, which allows to estimate the parameters via weighted log-likelihood. This procedure is repeated until convergence.

References for the Appendix

1. Daly, D.S. *et al.*, *Mixed-Effects Statistical Model for Comparative LC−MS Proteomics Studies.* Journal of Proteome Research, 2008. **7**(3): p. 1209-1217.
2. Zhou, T., *Weighting Method for a Linear Mixed Model.* Communications in Statistics - Theory and Methods, 2009. **39**(2): p. 214-227.

3. Zhou, X., H. Lindsay, and M.D. Robinson, *Robustly detecting differential expression in RNA sequencing data using observation weights.* Nucleic Acids Research, 2014. **42**(11): p. e91-e91.
4. Bates, D. *et al.*, *Fitting Linear Mixed-Effects Models Using lme4.* Journal of Statistical Software; Vol 1, Issue 1 (2015), 2015.

9.2. Experimental design and data-analysis in label-free quantitative LC/MS proteomics: A tutorial with MSqRob

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

9.2.1. Highlights

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

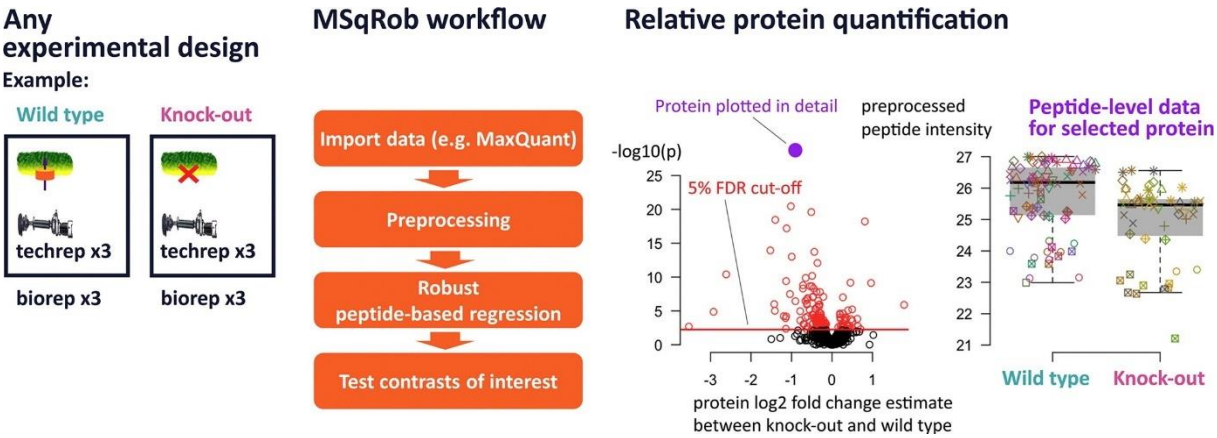
9.2.2. Abstract

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

9.2.3. Significance

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

9.2.4. Graphical abstract



9.2.5. Keywords

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

9.2.6. Historical background

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

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

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

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

In data-dependent acquisition (DDA), a software identifies multiply charged peptide precursor ions with the highest intensities from deconvoluted MS (or MS^1) spectra. In a next step, such peptide ions are individually selected and further fragmented, typically by collision-induced dissociation, whereby MS/MS (or MS^2) spectra are recorded. The frequency by which peptide ions are fragmented depends on both the LC resolution and the scanning speed of the mass spectrometer, with increasingly faster instruments now mapping larger fractions of the expressed proteome than ever before [16], [17]. Fig. 9.6 gives an overview of a contemporary label-free mass spectrometry-based proteomics workflow.

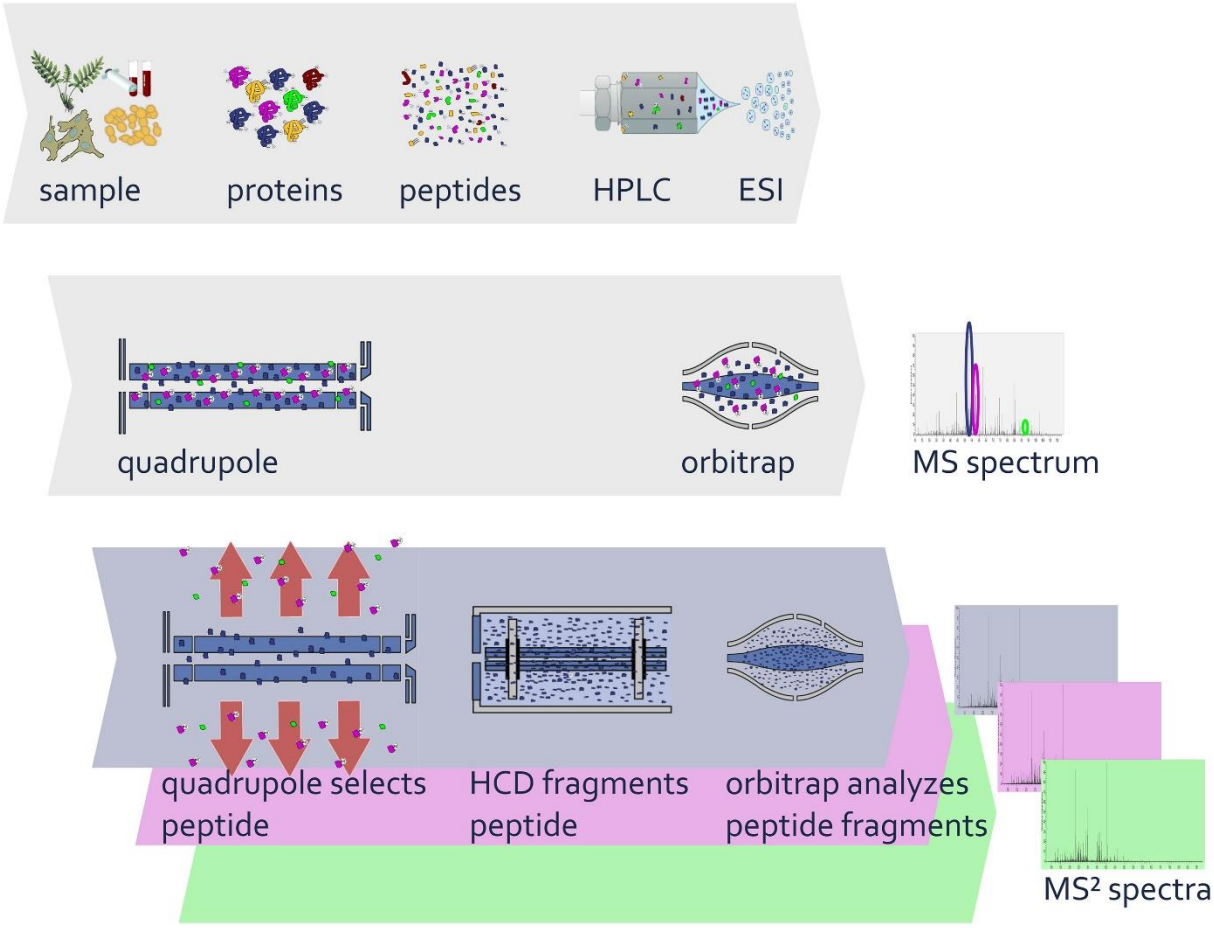


Figure 9.1. A typical DDA shotgun proteomics workflow using a quadrupole Orbitrap instrument. Extracted proteins are enzymatically digested to peptides, using a specific protease such as trypsin [18]. Peptides are then separated over a reverse-phase column and eluting peptides are transformed to gas-phase ions by electrospray ionization (ESI) [19]. At discrete time points, the eluting set of ionized

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

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

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

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

exclusion settings of the mass spectrometer (i.e. the same MS peak is fragmented only once in order to boost the number of identifications) might obscure the relationship between the number of counts and protein abundance [51], [52], [53]. Further, when machine settings are changed, runs become incomparable.

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

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

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

9.2.7. Basic concepts

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

Basic concepts on experimental design

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

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

Another important concept is that of **blocking** [77], which randomizes the different treatments to experimental units that are arranged within groups/blocks (e.g. batches, time periods) that are similar to each other. Due to practical constraints, it is often impossible to perform all experiments on the same day, or even on the same HPLC column or mass spectrometer, leading to unwanted sources of technical variation. In other experiments, researchers might test the treatment in multiple cultures or in big experiments that involve multiple labs. A good experimental design aims to mitigate unwanted sources of variability by including all or as many treatments as possible within each block. That way, variability between blocks can be factored out from the analysis when assessing treatment effects (Fig. 9.7).

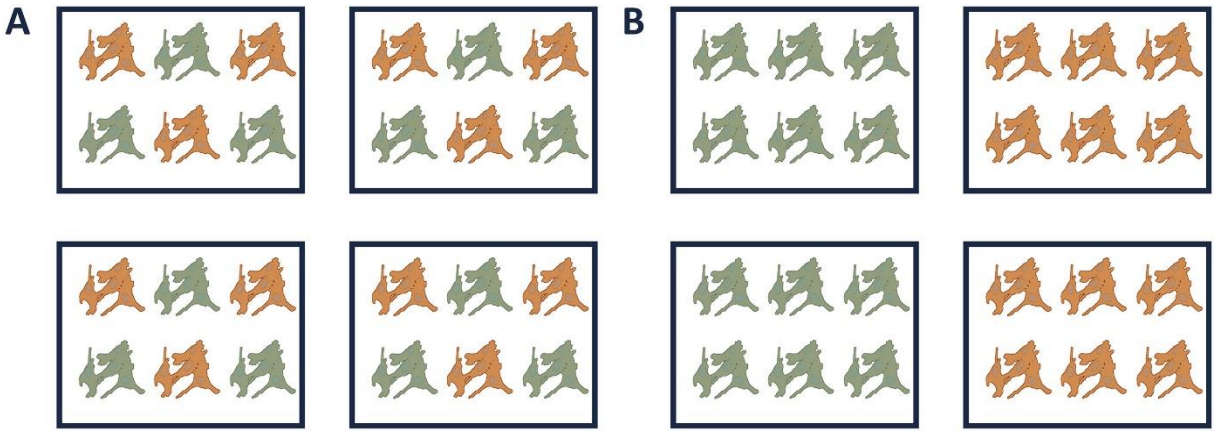


Figure 9.2. Example of a good (A) and a bad (B) design. In design A, both the green and orange treatments are divided equally within each block. That way, the treatment effect can be estimated within a block. In design B, each block contains only one treatment, so the treatment effect is entirely confounded with the blocking effect and it is thus impossible to draw meaningful conclusions on the treatment (unless one would be willing to assume that the blocking effect is negligible, which is a very strong assumption that cannot be verified based on the design).

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

A general label-free quantitative proteomics data analysis workflow

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

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

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

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

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

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

When working with the MSqRob package, one needs to discriminate between **fixed** and **random effects** [82], as MSqRob handles them differently. Fixed effects are those effects for

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

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

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

9.2.8. How is MSqRob used in research?

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

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

knowledge required. Moreover, MSqRob is capable of handling virtually any experimental design.

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

9.2.9. Case studies

Prerequisites

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

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

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

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

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

The *Francisella* example

Experimental set-up

The study on the facultative pathogen *Francisella* tularensis was conceived by Ramond et al. [88]. *F. tularensis* enters the cells of its host by phagocytosis. The authors showed that *F. tularensis* must import arginine from the host cell via a novel arginine transporter, ArgP, in order to efficiently escape from the phagosome and reach the cytosolic compartment, where

it can actively multiply. In their study, they compared the proteome of wild type *F. tularensis* (WT) to ArgP-gene deleted *F. tularensis* (knock-out, KO). For this experiment, bacterial cultures were grown in biological triplicate and each sample was run three times on a nanoRSLC-Q Exactive PLUS instrument. Hence, pseudo-replication occurs on different levels of the experiment, i.e. multiple peptides for the same protein in each MS-run (technical repeat) and 3 technical repeats for each biological repeat. The data were searched with MaxQuant version 1.4.1.2. Below, we give an overview on how to process the data with the MSqRob Shiny App.

The input tab (Fig. 9.8)

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

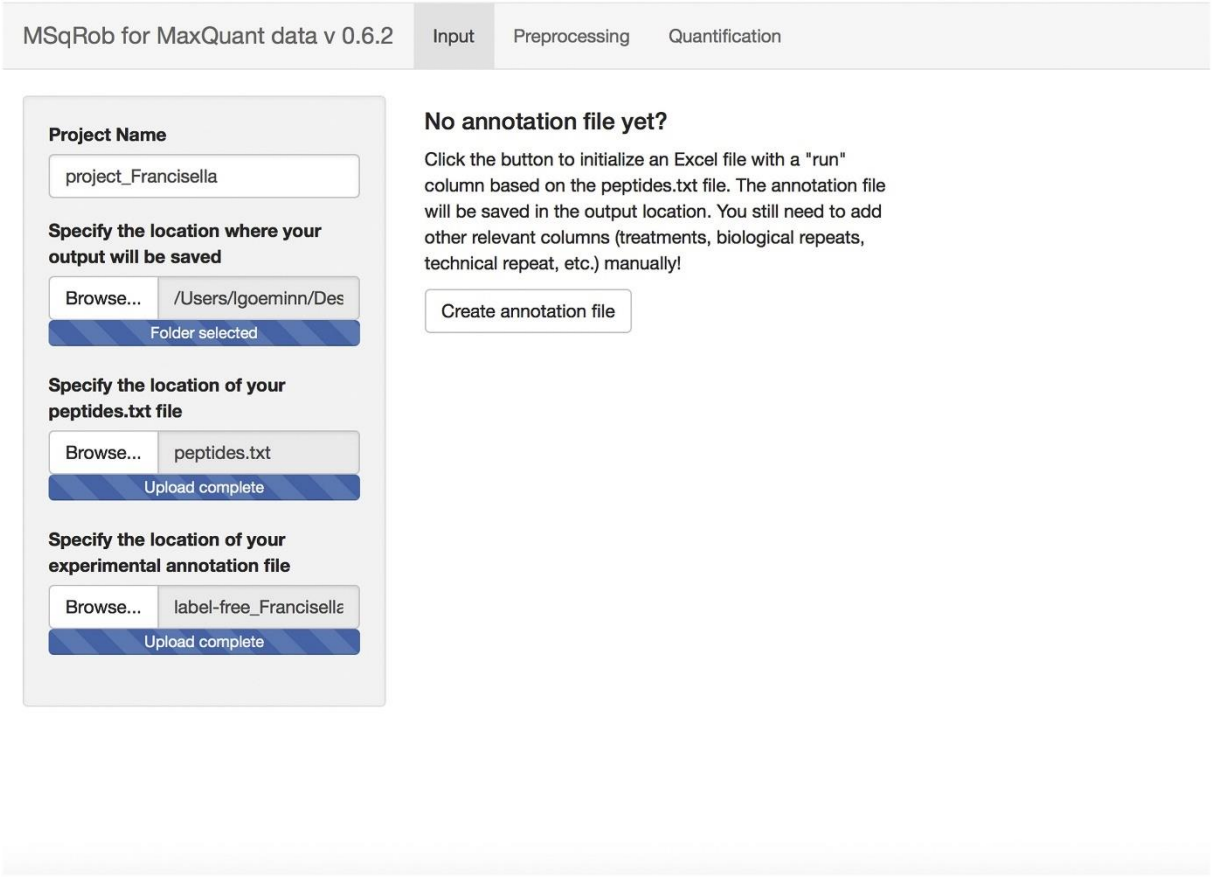


Figure 9.3. Overview of MSqRob's input tab.

Similarly, upload the **experimental annotation file**. This file should be a tab-delimited file or an Office Open XML spreadsheet file (“.xlsx” file). If needed, this file can be made based on Fig. 9.9. If the file location was already specified and the peptides.txt file was uploaded, one can generate the “run” column of this file automatically by clicking the “Create annotation file” button. The other columns need to be filled in manually based on the experimental design. Alternatively, one can download the file from

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

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

Figure 9.4. Experimental annotation file for the Francisella dataset.

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

The preprocessing tab (Fig. 9.10)

Left panel

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

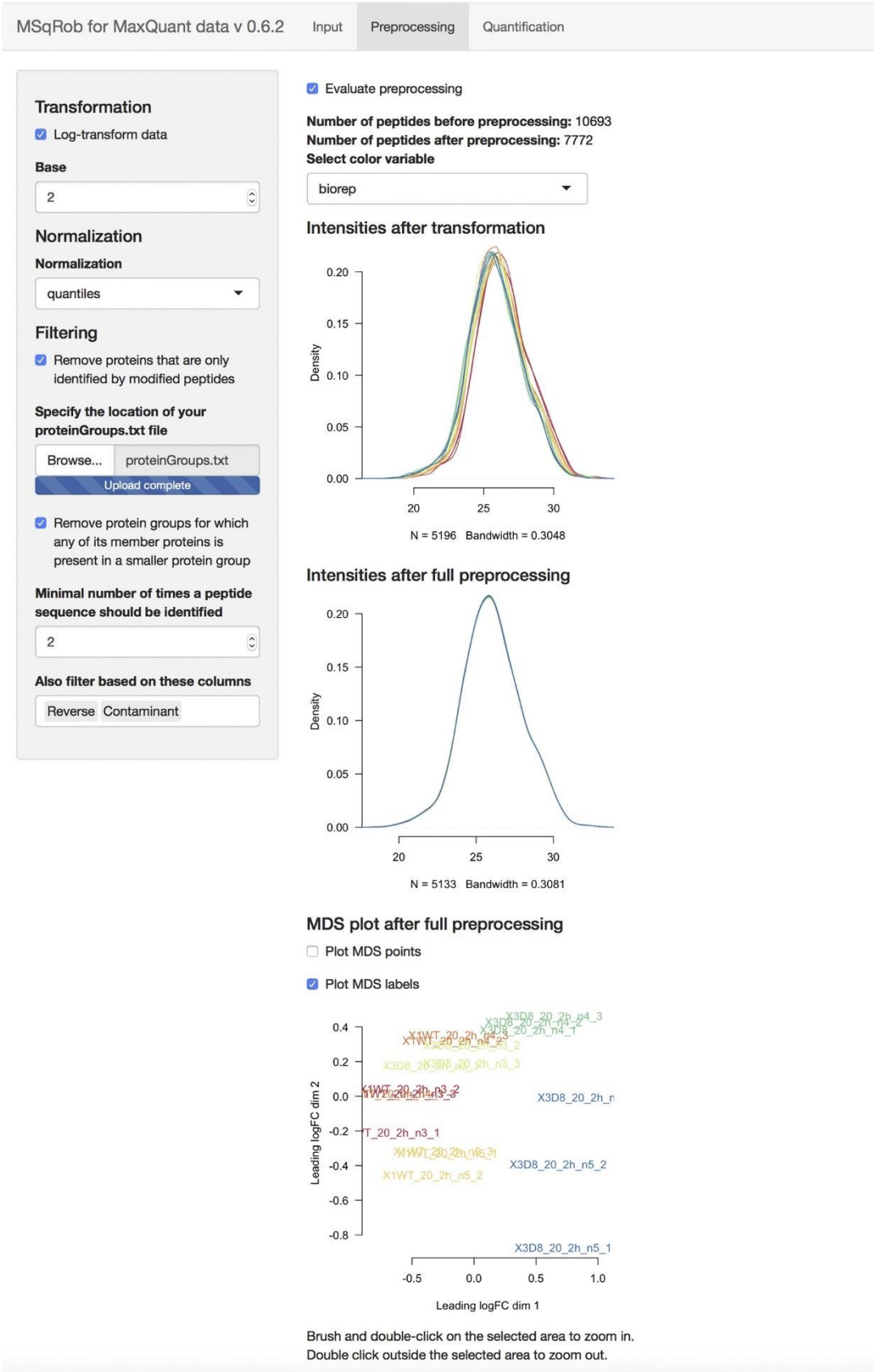


Figure 9.5. Overview of MSqRob's preprocessing tab.

MS-based proteomic intensity distributions are nearly always strongly skewed to the right. Therefore, a log-**transformation** is highly recommended. We suggest to log-transform the

data with **base** 2. This has the added advantage that the model estimates will be interpreted as $\log_2$ FC. For the remainder of this work, we assume that intensities have been $\log_2$-transformed.

We provide different **normalization** approaches. As a default, we suggest quantile normalization [79], [90]. Quantile normalization imposes the same empirical intensity distribution on all runs. More information on other normalization methods that are implemented can be found in the documentation of the 'normalise' function in the R package MSnbase [91]. The effect of quantile normalization on the distribution of the $\log_2$-transformed peptide intensities is shown in Fig. 9.11.

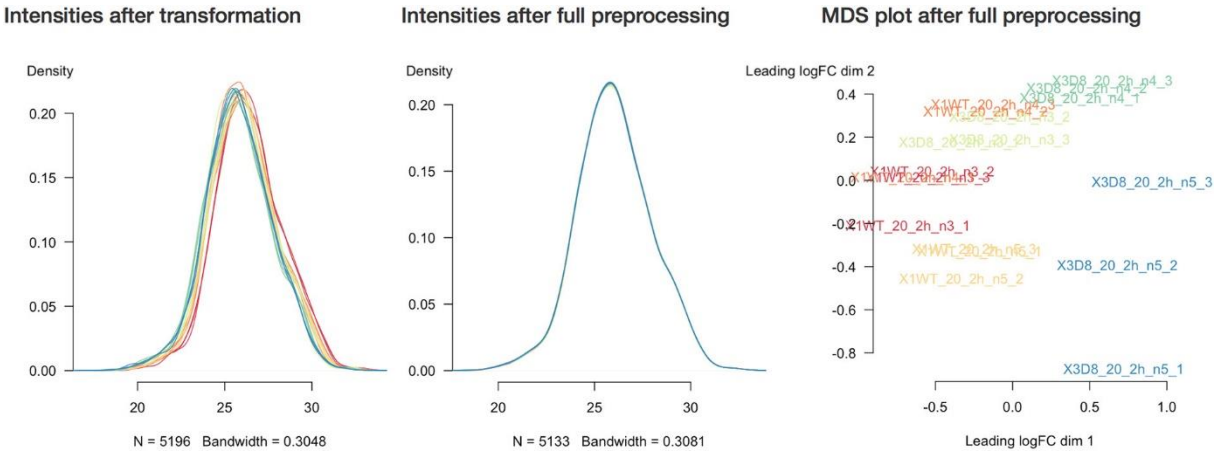


Figure 9.6. Overview of the $\log_2$-transformed peptide intensities for the Francisella dataset before and after preprocessing. Left: $\log_2$-transformed peptide intensities before preprocessing, center: $\log_2$-transformed peptide intensities after preprocessing. Note that the densities are forced onto the same distribution. Right: MDS plot that clusters similar MS runs together.

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

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

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

identified by only one peptide can contribute to the estimation provided that the peptide is identified in multiple samples, say t with $t \geq T$.

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

Right panel

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

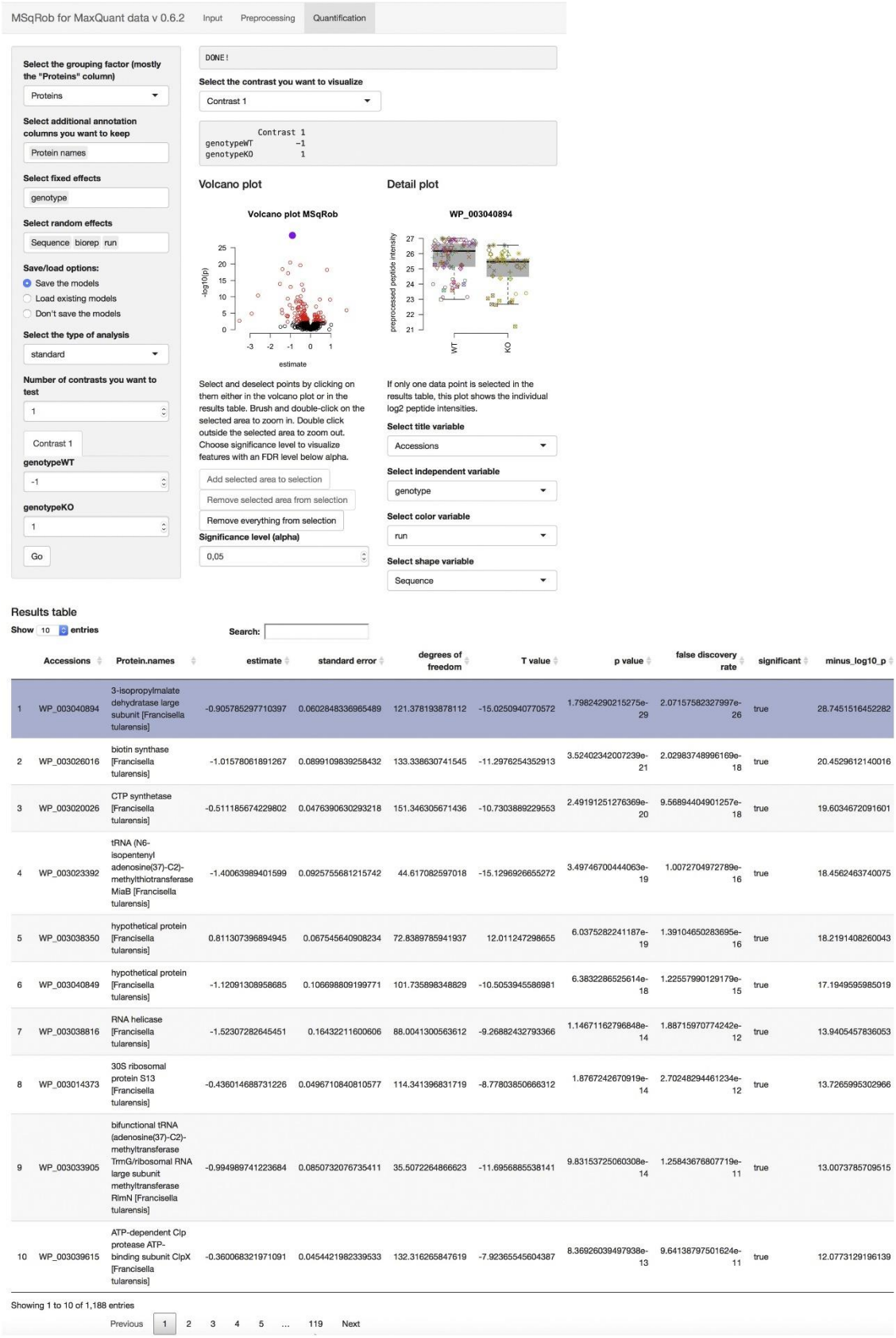
The effect of preprocessing can be assessed by ticking the **"Evaluate preprocessing"** box. A plot will be generated that shows the densities of the (log-transformed) peptide intensities after full preprocessing (i.e. after normalization and filtering) (Fig. 9.11). A multidimensional scaling (MDS) plot will also be produced, which shows a dot for each MS run such that the distance between two dots is equal to the root-mean-square deviation for the top 500 peptides that distinguish the two corresponding runs. Options are provided to show only dots, only labels or both. It is also possible to zoom in on a particular part of the plot by dragging the mouse to select a particular area on the plot and then double-click to zoom in.

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

The quantification tab (Fig. 9.12)

Left panel

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



Results table

Show 10 entries Search:

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

Showing 1 to 10 of 1,188 entries

Previous 1 2 3 4 5 … 119 Next

Figure 9.7. Overview of MSqRob's quantification tab.

Next, the **fixed effects** need to be specified. Fixed effects are effects that remain constant when repeating the experiment. For factor variables (e.g. genotype), the number of levels are

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

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

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

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

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

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

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

Right panel

When the analysis is finished, MSqRob prints “DONE!” at the top of the right panel. In this case study we only evaluated one contrast (KO vs. WT). If multiple contrasts are specified, one can select the contrast one would like to explore further. The **“Volcano plot”** shows $-\log_{10}$(p-values) as a function of the “estimate” (i.e., here the $\log_2$ FC between KO and WT for the *Francisella* example). One can select an area on this plot using the computer mouse and double clicking zooms in on this area. Upon selecting such an area, one can add all points in the area to a selection or remove all points in this area from a selection using their respective buttons. By clicking on a dot, one selects/deselects it. When only one protein is selected, a **“Detail plot”** is made for this protein, which shows the preprocessed peptide intensities as a function of a predictor variable from the model. Boxplots show the median preprocessed peptide intensity as a thick black line, the box itself comprises the interquartile range (IQR) and

whiskers extend to the most extreme data point that lies within 1.5 times the IQR on each side [93]. Each peptide intensity in the Detail plot can be given a color and a shape value according to any model parameter. Fig. 9.13 shows a detail plot for the most significant protein in the study, WP_003040894 or 3-isopropylmalate dehydratase large subunit, an enzyme required for the biosynthesis of leucine. Note that all identified enzymes required for the synthesis of branched chain amino acids were found either unchanged or downregulated in the ArgP mutant. Here, we specified "genotype" as independent variable, "run" as color variable and "Sequence" as shape variable.

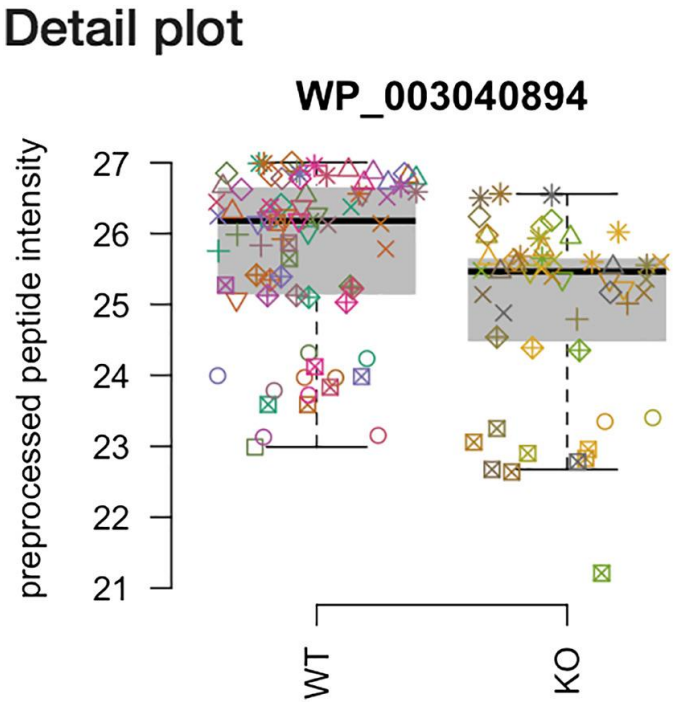


Figure 9.8. Example of a detail plot for the most significant protein in the study performed: 3-isopropylmalate dehydratase large subunit.

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

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

Note that in the file location one provided in the input tab, a folder is created, which is named "project_Francisella*_[date and time of analysis]*". In this folder, one finds the

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

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

The CPTAC example

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

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

The input tab

Again, an appropriate **name** is chosen for the **project**. Here, use “project_CPTAC”, select the **file location** where the output has to be saved. Next, the location of the experimental annotation file and the peptides.txt file is specified, and MaxQuant's peptides.txt file is imported. An example of the experimental annotation for the CPTAC dataset is given in Fig. 9.14. This file can be downloaded from https://github.com/statOmics/MSqRobData/blob/master/inst/extdata/CPTAC/label-free_CPTAC_annotation.xlsx.

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

Figure 9.9. Top 30 rows of the annotation file for the CPTAC dataset.

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

The quantification tab

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

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

Upon comparing condition 6B to condition 6A on the Volcano plot, we noticed that most hits have positive FC estimates. These red circles on the right of “0” are indeed the UPS1 spike-in proteins and their levels in condition 6B are higher than in condition 6A. There appear to be two false positive hits (red circle left of “0” and the selected purple circle in Fig. 9.15). The selected yeast protein sp | P53115 | INO80_YEAST exhibits a strong negative $\log_2$ FC estimate of − 2.09 on the Volcano plot (Fig. 9.15). Upon inspecting this protein in the Detail plot, we found that this protein was only identified by two different peptides with very different intensity patterns (NAPSEGVMASLLNVEK: square and VSTTPLLK: circle). The intensities of the former peptide remain basically unchanged over the different spike-in concentrations, while those of the latter show a clear upwards trend with increasing spike-in concentration. Based on its intensity pattern, this latter peptide is very likely an incorrectly annotated UPS1 peptide. Indeed, our model assumes that all peptides behave in a similar way when comparing over samples. Here, the effect of the VSTTPLLK peptide is on average lower in condition 6B compared to the rest of the dataset, pulling the estimated average $\log_2$-intensity in this condition down. This effect is not at play for condition 6A, as this peptide was not identified, and therefore, the difference between 6B and 6A will be strongly negative (− 2.09). This example clearly demonstrates the added value of using Detail plots, as these enable detecting aberrations in the data that would otherwise go unnoticed, preventing researchers from drawing wrong conclusions.

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

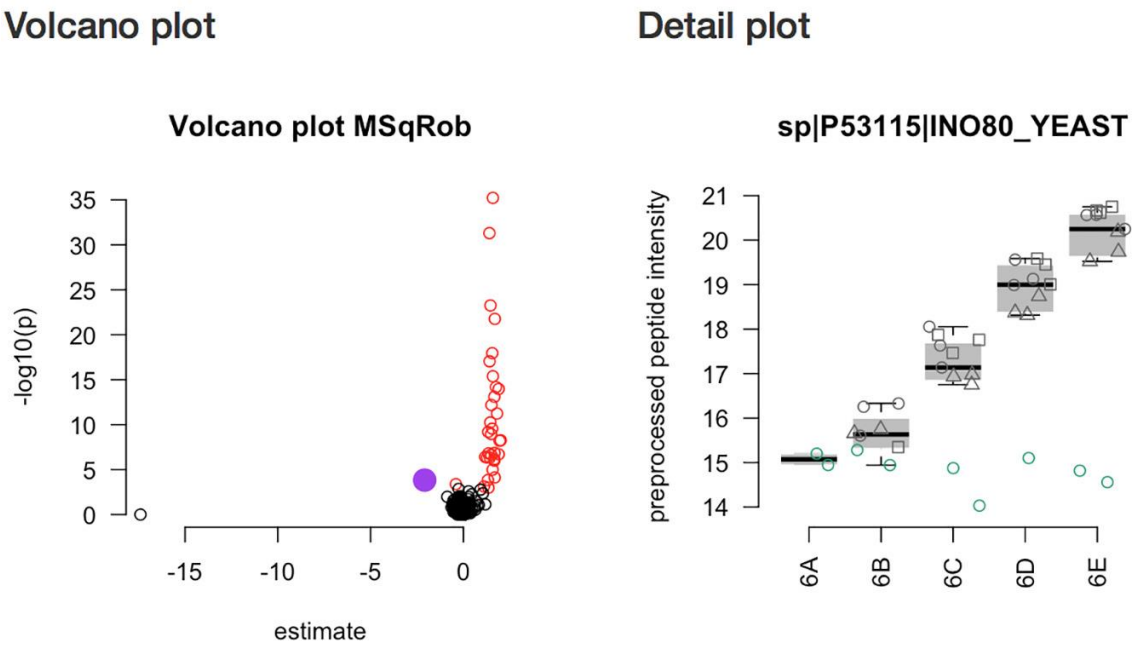


Figure 9.10. Use of the MSqRob output plots to find peculiar proteins. Protein sp | P53115 | INO80_YEAST has a log2 FC of − 2.09 and is identified by two different peptides, one of which is likely mis-annotated. In the Detail plot, points are colored by lab, while a different shape denotes a different peptide sequence.

9.2.10. Current limitations and useful working limits

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

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

9.2.11. Future developments

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

A constant theme in improvements of mass spectrometry instruments has been their increase in analysis speed and proteome coverage. We expect this trend to continue, which could, in the long run, reduce or even eliminate intensity-dependent missingness. Faster machines will also allow biologists to analyze an increasing number of biological repeats, which will boost the power of their experiments and allow them to detect small, but sometimes very relevant perturbations with greater confidence. Thanks to such increasing coverage, each generation

of machines allows us to dive deeper into the proteome than ever before. As machine duty cycles continue to increase, DIA and DDA are expected to come closer together as DIA windows will become smaller and smaller [102], while the analysis depth in DDA will continue to increase, so that one day, they might merge into a single technique that is capable of identifying all peptides in a sample. When that happens, the need to handle missing data in DDA will become obsolete.

9.2.12. Acknowledgements

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

9.2.13. Appendix A

Tutorial slide show (5MB)

9.2.14. References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

9.2.15. Appendix

We propose the same peptide-based regression model as in section 9.1:

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

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

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

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

$$\boldsymbol{X}^{\text{fixed}} = \boldsymbol{Q}\boldsymbol{R}$$

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

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

Implementation

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

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

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

$X_{ridgeGroup}$ `<- model.matrix(…)`

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

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

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

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

And finally fit the model:

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

optimizerOutput <- optimizeLmer(devianceFunction)

mRidge <- mkMerMod(
                    rho = environment(devianceFunction),

                    opt = optimizerOutput,

                    reTrms = parsedFormula$reTrms,

                    fr = parsedFormula$fr)
```

When doing inference, the part of the design matrix that was replaced with $Q_{ridgeGroup}$, is post-multiplied with $R_{ridgeGroup}$. A similar procedure to create a single ridge penalty for multiple covariates in lm4 models has also exploited by the gamm4 R package [1].

Reference for the Appendix

1. Wood, S.N., *Generalized Additive Models: An Introduction with R*. 2006: Chapman and Hall/CRC

10. MSQROB TAKES THE MISSING HURDLE: UNITING INTENSITY- AND COUNT-BASED PROTEOMICS

In chapter 10, we describe unpublished work. The motivation for this work is the high degree of missing data in label-free quantitative shotgun proteomics. Current solutions to the missing data problem, including those implemented by Perseus or MSstats, rely on imputation, which is often highly biased by assumptions that oversimplify reality.

Our manuscript proposes a novel approach to address this important issue of missing data, by defining a hurdle model for the peptide intensities. The hurdle model is a mixture of a binary component that distinguishes between log-transformed peptide intensities that are missing and observed, and a normal component to model the magnitude of the $\log_2$-transformed peptide intensities passing the detection hurdle. The components of the hurdle model naturally combine the strength of count-based and intensity-based workflows for differential analysis in quantitative label-free MS-based proteomics within one model framework. In our manuscript, we first illustrate that imputation assumptions used in leading tools (i.e., Perseus and MSstats) often have a detrimental effect on the quantification of proteins with missing data. We then show that our novel hurdle approach rescues this situation without needing to rely on these unrealistic imputation assumptions. Importantly, we demonstrate that our method outcompetes existing approaches, and continues to provide reliable quantification for proteins where peptides are absent in one (or more) conditions. For this manuscript, I designed and performed analyses, set up the GitHub repository and wrote the paper together with my co-authors.

Ludger J.E. Goeminne, Adriaan Sticker, Lennart Martens, Kris Gevaert, and Lieven Clement

10.1. Abstract

Missing values present a major issue in quantitative data-dependent mass spectrometry-based proteomics. We therefore present an innovative solution to this key issue by introducing a hurdle model, which is a mixture between a binomial peptide count and a peptide intensity-based model component. It enables dramatically enhanced quantification of proteins with many missing values without having to resort to harmful assumptions for missingness. We demonstrate the superior performance of our method by comparing it with state-of-the-art methods in the field.

10.2. Introduction

Label-free data-dependent quantitative mass spectrometry (MS) is the preferred method for deep and high-throughput identification and quantification of thousands of proteins in a single analysis [1]. However, this approach suffers from many missing values, which strongly reduce the amount of quantifiable proteins [2, 3]. There are three common causes for this missingness: (1) true absence of signal, or signal below detection limit in the MS1 spectrum; (2) lack of fragmentation and hence missed identification of the MS1 peak; and (3) failed identification of the acquired fragmentation spectrum. As a result, missingness is more likely to occur for low-abundant proteins and/or poorly ionizing peptides. However, missingness may also extend to mid- and even high-range intensities, e.g. when co-eluting peptides suppress an MS1 signal or when poor quality of an MS2 spectrum interferes with correct identification. Missingness due to lack of fragmentation can be mitigated by "matching between runs", where unidentified MS1 peaks in one run are aligned to identified

peaks in another run in narrow retention time and mass-over-charge (m/z) windows [4]. Nevertheless, missing values remain widespread; a survey of 73 recent public proteomics data sets from the PRIDE database demonstrates an average of 44% missing values (Fig. 1 a, Supplementary Table 1).

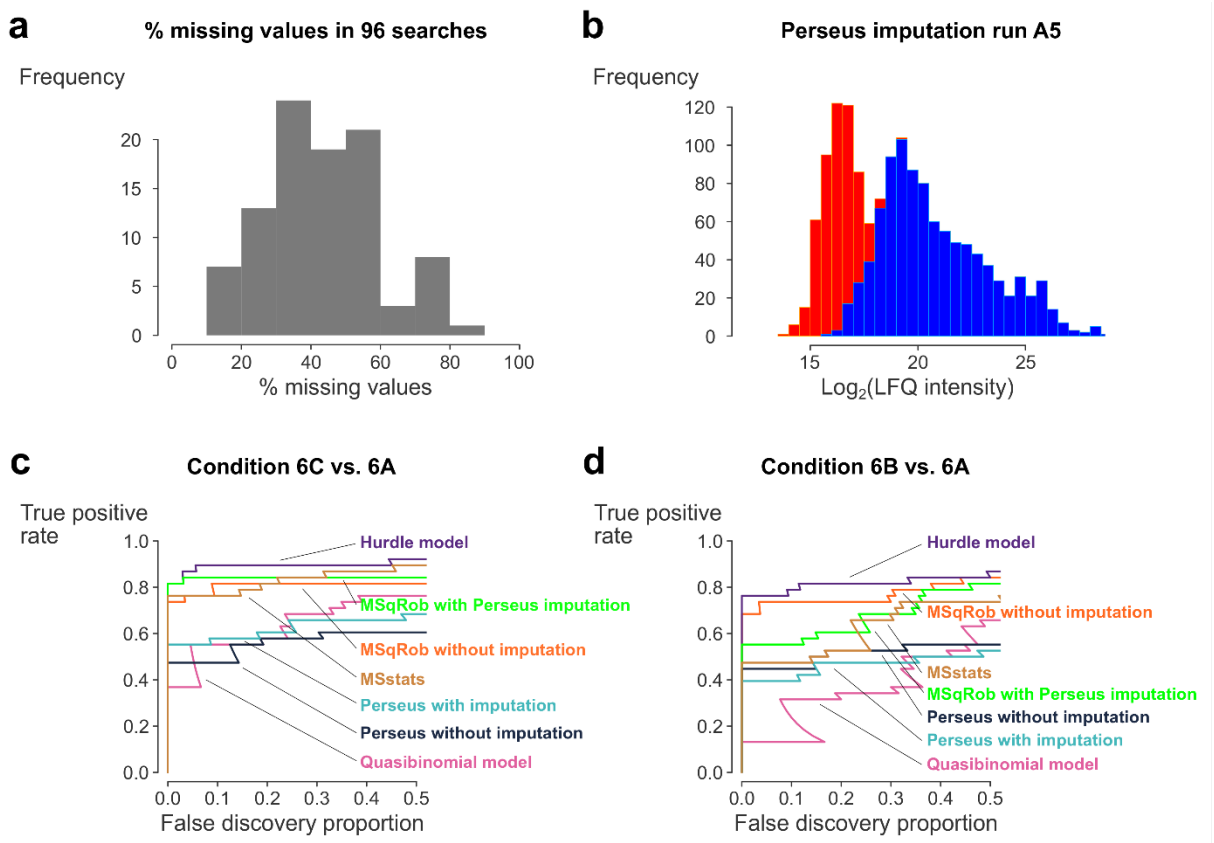


Figure 1. The impact of missing values and the superior performance of the hurdle model. (a) Missing values are highly prevalent in recent proteomics data. The histogram shows the distribution of the percentage of missing values in 96 peptides.txt files from 73 PRIDE projects with a PRIDE publication date in 2017 that applied shotgun proteomics to full or partial proteomes. Missingness ranges from 16 to 82%, with an average of 44% missing values. (b) Perseus imputation fails when too many missing values are present: the frequency distribution of the $\log_2$-transformed LFQ intensities in the CPTAC dataset becomes bimodal after Perseus imputation due to the many imputed values, as exemplified by run A5. Observed values (978 LFQ intensities) are colored in blue, imputed values (484 LFQ intensities) are colored in red. Similar results for the other runs can be seen in Supplementary Fig. 1. (c, d) In the CPTAC dataset (38 UPS1 and 1,343 yeast proteins), the hurdle model outperforms other methods. The fraction of true positive UPS1 proteins flagged as DA in conditions C vs. A and B vs. A is plotted as a function of the fraction of false positive yeast proteins in the total number of DA proteins.

A common solution for missingness is to impute the missing values. But, this comes at the expense of additional assumptions, e.g. k-nearest neighbors (kNN) assumes intensity-independent missingness, and the popular software packages Perseus [5] ("Perseus imputation", PI) and MSstats [6] assume missingness by low intensity. In reality, however, missing values originate from a mix of intensity-dependent and -independent mechanisms, which can moreover be strongly data set-specific [3, 7]. As a result, state-of-the-art imputation profoundly changes the distribution of protein-level intensities as it typically produces a second mode due to many imputed values, (shown for MaxQuant/Perseus in Fig. 1 b and Supplementary Fig. 1 – 2; and for kNN in Supplementary Fig. 3 – 4). This can have a deep impact on the downstream differential analysis. We demonstrate these effects on the widely studied CPTAC dataset [8], where more than 47% of peptide data are missing.

Because the spike-in concentrations in this dataset cover a wide range, the perils of imputation can be clearly shown. Indeed, while our MSqRob [9] quantification algorithm shows very good performance when combined with PI for comparison C (high spike-in concentration: 2.2 fmol/µL) vs. A (lowest spike in concentration: 0.25 fmol/µL) (Fig. 1 c), MSqRob with PI performs very poorly for comparisons B (intermediate amount of spike-in: 0.74 fmol/µL) vs. A (Fig. 1 d) and C vs. B, where the differences in spike-in concentration are smaller (Supplementary Fig. 5, 6), thus rapidly accumulating false positives. MSqRob therefore omits imputation by default to avoid a severe backlash in performance [10]. However, without imputation, intensity-based methods cannot cope with complete missingness in one condition as it is impossible to calculate a fold change (FC) relative to a missing value. Thus, potentially interesting cases such as strong protein synthesis/stabilization or protein degradation remain undetected.

Conversely, peptide counting approaches naturally handle missing peptides in a run by a zero count. Yet relative quantification by peptide counting generally performs very poorly (pink line in Fig. 1 c, d). This because counting disregards the inherent abundance-intensity relationship (within a given dynamic range) for each peptide [11, 12]. However, differentially abundant (DA) proteins for which no FC can be estimated due to missingness do differ in peptide counts (Supplementary Fig. 7) and proteins with a higher concentration have on average both higher peptide ion intensities and higher peptide counts, as it is more likely that even their poorly ionizing peptides are detected.

Combining intensity-based methods with counting-based methods to exploit this complementary information therefore seems promising. Webb-Robertson *et al.* combine intensity- and count-based statistics to filter out peptides prior to differential analysis [13]. ProPCA uses principal components to summarize intensities and spectral counts into one value [14] and IDPQuantify uses Fisher's method to combine the p-values from a two-sample t-test with those of a quasipoisson regression on spectral counts [15]. However, ProPCA's combined metric lacks an intuitive interpretation, and does not include any downstream statistical analysis. Like Perseus, IDPQuantify only handles simple pairwise comparisons without any possibility to correct for confounding effects. Moreover, none of these methods is able to pinpoint whether the statistical significance is driven by the count component, the intensity component, or both, thus making it impossible to interpret the results in terms of FCs.

10.3. Results and discussion

Here, we introduce a hurdle model that unites the advantages of MSqRob with the complementary information present in peptide counts, that avoids unrealistic imputation assumptions, and that provides interpretable results. This hurdle model takes peptide-level information as input and consists of a mixture model of two components: (1) a binary component that distinguishes between $\log_2$-transformed peptide intensities that are either missing or observed; and (2) an MSqRob-based component to model the magnitude of $\log_2$ peptide intensities passing the detection hurdle. Inference on the parameters of both model components has an intuitive interpretation: the binary component can be used to assess *differential detection (DD)* and returns *log odds ratios* (log ORs), while the MSqRob component allows to test for *differential abundance (DA)* of a protein and returns *$\log_2$ fold changes* ($\log_2$ FCs). When peptides are completely missing in one condition, the hurdle model reduces to a binomial model. However, when peptides are present in both conditions, the hurdle approach allows to combine information in the OR and FC test statistics, and can be used to infer on DD, DA, or both in a post-hoc analysis. Note, that the peptide counts

can be over- or underdispersed, which is why the variance component is estimated via quasi-likelihood and the model is termed a quasibinomial model.

We use the CPTAC study 6 dataset [8] to compare our hurdle approach to the quasibinomial model alone, to MSqRob alone, to Perseus (with and without imputation), and to MSstats. Fig. 1 c, d demonstrates the superior performance of the hurdle model: it consistently outperforms the other algorithms. Moreover, the hurdle model also always outperforms MSqRob with imputation under a low abundance assumption such as PI (Supplementary Fig. 5, 6). In the lowest spike-in condition, A, missing values in the UPS1 proteins are mainly caused by low abundance, allowing the count-component of the hurdle model to add to the strength of the intensity-component. In comparison C vs. B, the UPS1 spike-in concentrations are higher, reducing the difference in peptide counts between both conditions, and mainly driving significance of the hurdle model by the difference in average intensities. In this case, the hurdle model performs on par with MSqRob without imputation, as is expected, while PI approaches still perform poorly (Supplementary Fig. 5, 6).

Next, we assess the human heart dataset of Doll *et al.* (2017) [16]. For the 7,822 gene identifiers in common after preprocessing, we assessed the overlap between the 1,500 most significantly regulated identifiers in the Doll *et al.*, hurdle, and MSqRob comparisons between the atrial to the ventricular proteome (Fig. 2 a, Supplementary Table 2). The 652 identifiers shared between all methods correspond to proteins with a strong DA and many identified peptides (Supplementary Fig. 8). The 209 identifiers shared between Perseus and hurdle mostly have strong DD and only few peptides in one of the heart chambers (Supplementary Fig. 9). The 204 identifiers unique to hurdle differ strongly in peptide counts, but the $\log_2$ FCs are close to one. (Supplementary Fig. 10). The 543 identifiers unique to Perseus often show very few peptide identifications or very small FCs, making these results more questionable (Supplementary Fig. 11). Indeed, the imputation strategy again has a profound impact on the results: there is a 22% non-overlap between Perseus with and without the PI strategy. Moreover, because of the stochastic nature of the PI imputation, on average 4.5% of the first 1,500 proteins declared DA did not overlap between two repeated PI analyses. The 96 identifiers shared between MSqRob and Perseus, finally, mainly have many peptides, but a relatively small FC (Supplementary Fig. 12). These are not included in the top 1,500 differential proteins by the hurdle model as their place in the ranking is taken by identifiers with a strong DD. A similar analysis for the 500 and 1,000 most differential proteins for each method is given in Supplementary Fig. 13.

For our novel approach we also observe a strong association between the OR and FC estimates (Fig. 2 b). Out of the 1,496 hurdle identifiers that are significant at the 1% FDR level, 466 were significantly DD, 499 significantly DA, 297 both significantly DD and DA, and 234 could not be attributed to either DA or DD in our *post hoc* analysis. All 297 identifiers declared both DD and DA have a FC and OR estimate in the same direction (i.e. both up or both down).

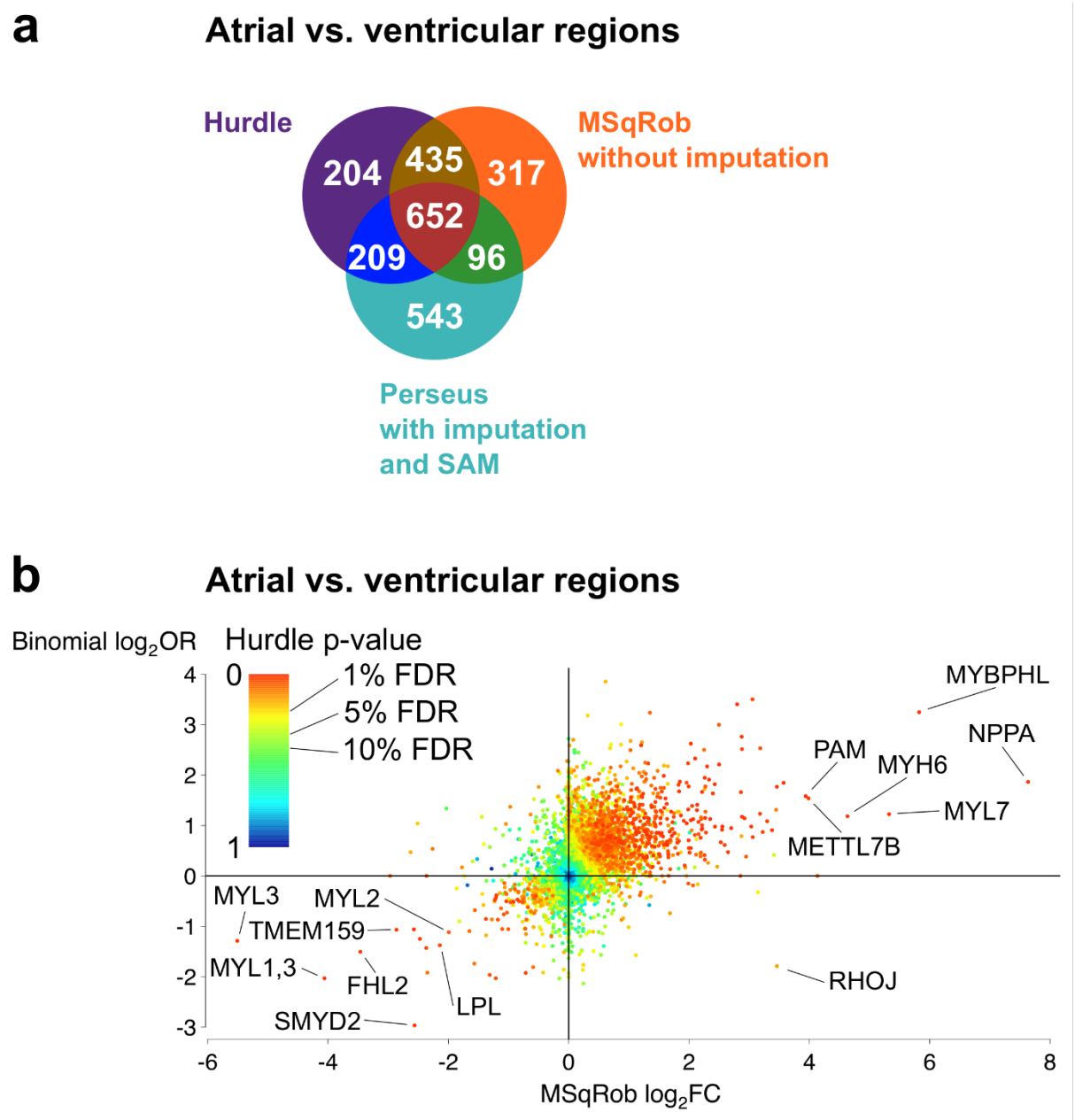


Figure 2. The hurdle model detects both DD and DA proteins. (a) The Venn diagram shows the overlaps in the 1,500 most significant proteins when contrasting atrial (average of left atrium, LA; right atrium, RA; atrial septum, SepA) to ventricular regions (average of left ventricle, LV; right ventricle, RV; ventricular septum, SepV) in the HEART dataset between hurdle, MSqRob without prior imputation, and Perseus with imputation and SAM as implemented by Doll et al. (b) Proteins that are significant with the hurdle model often display a strong positive association between FC and OR. Here, $\log_2$ FCs are plotted as a function of $\log_2$ ORs in the atrial vs. ventricular comparison for all 6,489 proteins with an FC estimate. Proteins are colored by p-value. Common FDR cut-offs are indicated on the color bar.

As in the original publication, comparing the left to the right ventricle did not render any significant proteins. However, when comparing the left to the right atrium, we found twelve significant proteins compared to only three in the original publication (Supplementary Table 3). The higher abundance of the muscle contraction protein myotilin [17] in the left atrium was confirmed by both methods, but the hurdle model additionally indicates several heavy and light myosin chains such as MYH7 and MYL3 as DA. Moreover, the hurdle model also flags the voltage-dependent calcium channel CACNA2D2 and BMP10, an essential factor for embryonal cardiac development [18]. Finally, SERINC3 and PNMA1, which were

declared DA by Doll *et al.* actually come with very limited evidence for DA and their significance is likely driven by the inherent randomness of PI (Supplementary Table 4).

In summary, we demonstrated the negative impact of imputation in protein quantification, and the potential flaws it induces in the downstream analysis when imputation assumptions are violated. We therefore propose a hurdle approach that can handle missingness without having to resort to imputation and its associated assumptions by leveraging both an intensity-based approach (MSqRob) and a count-based quasibinomial approach. Our hurdle approach outcompetes state-of-the-art methods with and without imputation in the presence of both strong missingness as well as when missingness is limited. Moreover, our hurdle method continues to provide valid inference when peptides are absent in one condition. It is therefore a highly promising approach to detect the sudden appearance of post-translationally modified peptides in protein regulation alongside more traditional differential protein expression.

10.4. Methods

In this section, we first demonstrate how we analyzed the frequency of missing values in recent PRIDE projects. Then, we discuss the nature and the origin of the CPTAC and the HEART datasets, followed by an overview of how these datasets were preprocessed and imputed with different strategies. We then introduce the Perseus, MSstats, MSqRob, the quasibinomial model and the hurdle model as methods for statistical inference. Finally, we refer to our GitHub page which enables to reproduce all the analyses in this publication.

10.4.1. Missing values in recent PRIDE projects

For our analysis of the frequency of missing values of recent datasets in PRIDE, we downloaded 96 peptides.txt files corresponding to 73 PRIDE projects that adhere to the following conditions:

- label-free shotgun proteomics,

- full proteomes or enriched subsets of the proteome (i.e. no projects where there was an enrichment step for a chemical modification and no projects investigating protein-protein interactions),

- published in PRIDE in 2017 and

- searched with MaxQuant and peptides.txt file available from PRIDE.

An overview of the projects and peptides.txt files with their number of missing values, number of observed values and percentage of missing values is given in Supplementary Table 1.

10.4.2. Data availability

We made use of two example datasets.

1. In the benchmark CPTAC study 6, which can be downloaded from https://cptac-data-portal.georgetown.edu/cptac/dataPublic/list?currentPath=%2FPhase_I_Data%2FStudy6&nonav=true, the human UPS1 standard is spiked in 5 different concentrations in a yeast proteome background [8]. This allows to make comparisons for which the ground truth is known: when comparing two different spike-in conditions, only the UPS1 proteins are truly differentially abundant (DA), while the yeast proteins are not.

The three lowest spike-in conditions (A, B and C) from LTQ-Orbitrap at site 86, LTQ-Orbitrap O at site 65 and LTQ-Orbitrap W at site 56 were searched with MaxQuant 1.6.1.0 against a database containing 6,718 reviewed *Saccharomyces cerevisiae* (strain ATCC 204508/S288c) proteins downloaded from UniProt on September 14, 2017, supplemented with the 48 human UPS1 protein sequences provided by Sigma Aldrich. Carbamidomethylcysteine was set as a fixed modification and methionine oxidation, protein N-terminal acetylation and N-terminal glutamine to pyroglutamate conversion were set as variable modifications. Detailed search settings are described in Supplementary Material.

We only performed pairwise comparisons between the 3 lowest spike-in concentrations because of the huge ionization competition effects in the higher spike-in concentrations, as described earlier [9].

2. The HEART dataset originates from a large-scale proteomics study of the human heart, where 16 different regions and 3 different cell types from three healthy human adult hearts were studied [16].

For our analysis, we made use of the peptides.txt and proteinGroups.txt files made available by Doll *et al.* in ProteomeXchange via the PRIDE Archive repository under the identifier PXD006675. Here, we limited ourselves to the data from 6 regions: the atrial and ventricular septa (SepA and SepV), the left and right atrium (LA and RA) and the left and right ventricle (LV and RV). We compare the atrial regions (LA, RA and SepA) to the ventricular regions (LV, RV and SepV), as well as LA vs. RA and LV vs. RV.

10.4.3. Preprocessing for MSqRob and the quasibinomial model

Peptide intensities obtained from MaxQuant's peptides.txt file were $\log_2$-transformed and quantile normalized. Next, potential contaminants, reverse sequences and proteins that were only identified by peptides carrying a modification were removed from the data. Finally, proteins identified by only a single peptide were also removed.

10.4.4. Imputation methods

Missing values were either not imputed or imputed with kNN or Perseus imputation.

No imputation

For MSqRob without imputation, we additionally removed peptides that are only identified in a single run. Indeed, MSqRob directly models peptide intensities and the peptide-specific effect for peptides with one identification cannot be estimated because no replicates are available. This additional filtering step is not required for MSqRob with imputation, because missing values in the peptide intensity matrix are imputed first.

kNN imputation

k-nearest neighbors (kNN) imputation calculates a Euclidean distance metric on all peptide intensities to find the k most similar peptides and imputes the missing value with the average of the corresponding values from the k neighbors [19]. We used the default value of $k = 10$ neighbors.

Perseus imputation

We call "Perseus imputation (PI)" the standard imputation approach from the popular proteomics computational platform Perseus [5]. The imputation is achieved by using the

"Replace missing values from normal distribution" function in Perseus 1.0.6.7. We applied PI on peptide intensities for MSqRob with PI and on LFQ intensities for Perseus with PI.

PI first constructs a rescaled normal distribution with: (1) a downshifted mean equal to the average of all the observed data minus d times the standard deviation of the observed data and (2) a standard deviation equal to w times the standard deviation of the observed data. Missing values are then imputed with random draws from this rescaled distribution. We adopt Perseus' default values for w (0.3) and d (1.8).

10.4.5. Statistical inference

In this subsection, we discuss differential protein analysis with Perseus and MSstats, followed by MSqRob, the quasi-binomial model and our hurdle model. All methods mentioned below model the data protein by protein. To improve readability, we will suppress the protein indicator in the remainder of this section.

Perseus

For the CPTAC dataset, LFQ intensities were imported into Perseus 1.6.0.7. Potential contaminants, reversed sequences and proteins only identified by peptides carrying modification sites were removed from the data. Next, empty columns were removed and via "Categorical annotation rows", runs were sorted according to their spike-in conditions (A, B or C). Data were either imputed with PI ("Perseus with imputation") or not ("Perseus without imputation"). Finally, we performed two-sample t-tests for each of the three comparisons.

For the atrial vs. ventricular comparison in the HEART dataset, Doll *et al.* provided the results of their Perseus with imputation approach in their Supplementary Data 5 file, available from https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-017-01747-2/MediaObjects/41467_2017_1747_MOESM7_ESM.xlsx.

MSqRob

MSqRob has been described in detail in Goeminne *et al.* (2016) [9]. Briefly, for each protein, the $\log_2$-transformed peptide intensities y_{pr} for peptide $p = 1, \ldots, P$ in run $r = 1, \ldots, R$ are modeled as follows:

$$y_{pr} = \beta^0 + \boldsymbol{x}_{\boldsymbol{pr}}^{\mathrm{T}}\boldsymbol{\beta} + \beta_p^{\mathrm{peptide}} + u_r^{\mathrm{run}} + \varepsilon_{pr}, \qquad (\mathrm{Eq.}\ 10.1)$$

where β^0 is the intercept, $\beta_p^{\mathrm{peptide}}$ is the fixed effect of the individual peptide sequence p, u_r^{run} is a random run effect ($u_r^{\mathrm{run}} \sim \mathrm{N}(0, \sigma_u^2)$) that accounts for the correlation of peptide intensities y_{jr} and y_{kr} from the same protein within the same run r, $\boldsymbol{x}_{\boldsymbol{pr}}^{\mathrm{T}} = \left(x_{1,pr}, \ldots, x_{m,pr}\right)^{\mathrm{T}}$ is a vector with the covariate pattern of the m remaining predictors, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_m]^{\mathrm{T}}$ is a vector of parameters modeling the effect of each predictor on the peptide intensity conditionally on the remaining covariates, and ε_{pr} is the error term that is assumed to be normally distributed ($\varepsilon_{pr} \sim \mathrm{N}(0, \sigma^2)$).

For the CPTAC and HEART datasets, the MSqRob model can be written as follows:

$$y_{pr} = \beta^0 + \sum_{t=1}^{T} x_{t,pr}^{\mathrm{treat}} \beta_t^{\mathrm{treat}} + \sum_{b=1}^{B} x_{b,pr}^{\mathrm{block}} \beta_b^{\mathrm{block}} + \beta_p^{\mathrm{peptide}} + u_r^{\mathrm{run}} + \varepsilon_{pr}, \qquad (\mathrm{Eq.}\ 10.2)$$

where β^0 is the intercept, β_t^{treat} is the effect of interest ($t = 1$ to 3 corresponding to spike-in conditions A, B, C for CPTAC, $t = 1$ to 6 for the six cardiac regions in the HEART dataset),

β_b^{block} is a blocking factor (lab $b = 1$ to 3 for CPTAC, patient $b = 1$ to 3 for HEART), $x_{t,pr}^{\text{treat}}$ and $x_{b,pr}^{\text{block}}$ are dummy variables which are equal to 1 if run r corresponds to treatment t or block b, respectively, and 0 otherwise. β_p^{peptide} is the effect of the individual peptide sequence p, u_r^{run} is a random effect that accounts for the fact that peptides within each run r are correlated. Due to the parameterization of the model, the following restrictions apply: $\sum_{t=1}^{T} \hat{\beta}_t^{\text{treat}} = 0$, $\sum_{b=1}^{B} \hat{\beta}_b^{\text{block}} = 0$, $\sum_{p=1}^{P} \hat{\beta}_p^{\text{peptide}} = 0$ and $\sum_{r=1}^{R} \hat{u}_r^{\text{run}} = 0$.

The effect sizes $\beta_{(.)}^{(.)}$ (except for the intercept) are estimated using penalized regression. Distinct ridge penalties are used for the treatment, block and peptide parameters, respectively and the ridge penalties are tuned by exploiting the link between ridge regression and mixed models (see e.g. Ruppert *et al.* (2003), chapter 4 [20]). Outliers are accounted for using M-estimation with Huber weights. Protein-wise degrees of freedom are now calculated in a less liberal way as: $R - (1 + (T - 1) + (B - 1))$. Variances are stabilized by borrowing information over proteins using limma's empirical Bayes approach which results in a moderated t-test. Scripts to run MSqRob are provided on our GitHub page (see below under "Code availability").

MSstats

MSstats has been described by Choi *et al.* (2014) [6]. For our analysis, we used the default settings of MSstats version 3.12.2. During preprocessing, feature intensities are $\log_2$-transformed and normalized by equalizing the run medians. Then, missing values are imputed with the default MSstats settings. Features are summarized to the protein level with Tukey's median polish method. Treatment effects are specified as "Condition" and blocking factors as "BioReplicate" in the MSstats workflow.

Quasi-binomial model

When assuming that the number of observed peptides n_r^{peptide} in each run r (after preprocessing) for a protein are binomially distributed

$$n_r^{\text{peptide}} | \boldsymbol{x}_r \sim \text{Binomial}(P, \pi_r), \qquad \text{(Eq. 10.3)}$$

with P the total number of unique peptides observed over all runs $r = 1, \ldots, R$ for this protein, and π_r the probability to identify a peptide for this protein; the peptide counts can be modeled using logistic regression. However, they are often under- or over-dispersed with respect to the binomial distribution. We therefore adopt quasi-binomial regression (McCullagh and Nelder (1989), section 4.5 [21]) where we model the first two moments (mean $\text{E}[n_r^{peptide}]$ and variance $\text{Var}[n_r^{peptide}]$) as follows:

$$\text{E}\left[n_r^{peptide}\right] = P\pi_r, \qquad \text{(Eq. 10.4)}$$

$$\text{logit}(\pi_r) = \log\left(\frac{\pi_r}{1 - \pi_r}\right) = \boldsymbol{x}_r^{\text{T}}\boldsymbol{\beta}, \qquad \text{(Eq. 10.5)}$$

$$\text{Var}[n_r^{peptide}] = \varphi P \pi_r (1 - \pi_r), \qquad \text{(Eq. 10.6)}$$

with φ a dispersion parameter accommodating for a more flexible variance function than that of the binomial regression model. Note that the quasibinomial approach models the log odds, i.e. the logarithm of the probability that a peptide is detected divided by the probability that a peptide is not detected, i.e. the odds on detection. Hence, the model will return log

ORs when contrasting different treatments to each other, which can be used to infer on differential peptide detection for a specific protein.

Hurdle model

The normalized intensities for each peptide p in each run r are typically assumed to follow a log-normal distribution. Upon $\log_2$-transformation, missing (zero) intensities are set at $-\infty$ and cannot be modeled with intensity-based methods such as Perseus, MSstats and MSqRob. Missing values are therefore either omitted or imputed.

Here, we consider a hurdle model that consists of two parts: a binary component z_{pr} that distinguishes between peptide intensities in run r that are missing ($z_{pr} = 0$) or observed ($z_{pr} = 1$) with detection probability π_r; and a normal component y_{pr} with mean μ_{pr} and variance σ^2 to model $\log_2$ peptide intensities passing the detection hurdle. Note, that the detection probability π_r and the mean μ_{pr} can be further parameterized using peptide specific effects, a random run effect u_r^{run}, and additional covariates $\boldsymbol{x}_{pr}$. More formally, the hurdle model for $\log_2$-transformed intensities y_{pr} for peptide $p = 1, \dots, P$ in run $r = 1, \dots, R$ can be specified as follows:

$$z_{pr}|\boldsymbol{x}_{pr} \sim \text{Bernoulli}(\pi_r) \qquad \text{(Eq. 10.7)}$$

$$y_{pr}|z_{pr} = 1, \boldsymbol{x}_{pr}, u_r^{\text{run}} \sim \text{N}\left(\mu_{pr}, \sigma^2\right) \qquad \text{(Eq. 10.8)}$$

The log-likelihood for $\boldsymbol{\pi} = [\pi_1, \dots, \pi_R]^{\text{T}}, \boldsymbol{\mu} = [\mu_{11}, \dots, \mu_{PR}]^{\text{T}}$ and σ^2 given $\boldsymbol{y} = [y_{11}, \dots, y_{PR}]^{\text{T}}$, $\boldsymbol{z} = [z_{11}, \dots, z_{PR}]^{\text{T}}$, $\boldsymbol{u}^{run} = [u_1, \dots, u_R]^{\text{T}}$, and $\boldsymbol{X}$ the matrix with rows $\boldsymbol{x}_{pr}^{\text{T}}$ can then be written as:

$$l(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma^2|\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{X}, \boldsymbol{u}) = \sum_{pr}(1 - z_{pr}) \log(1 - \pi_r) + \sum_{pr} z_{pr} \log(\pi_r) + \sum_{pr} z_{pr} \log\left[\text{N}\left(\mu_{pr}, \sigma^2\right)\right]$$

$$\text{(Eq. 10.9)}$$

Note, that the log-likelihood implies an estimation orthogonality between π_r and μ_{pr} and that the first two terms in the equation are equivalent to the log-likelihood of a Bernoulli process. Further, we omit a peptide-specific effect for the parameterization of π_r because this leads to complete separation for too many proteins. The detection probability is thus considered constant for all peptides of a particular protein in run r. When summing over the peptides $p = 1, \dots, P$, peptide counts $n_r^{\text{peptide}} = \sum_p z_{pr}$ are obtained and it is computationally more efficient to estimate π_r using a binomial model for the peptide counts. To account for under- and/or over-dispersion in the counts, we will again estimate the detection probability π_r using quasi-binomial regression with Model $(\text{Eq. } 10.4 - 10.6)$. The last term in the log-likelihood corresponds to a normal log-likelihood and we propose to estimate μ_{pr} and σ^2 using the MSqRob Model $(\text{Eq. } 10.2)$.

Both model components model their means using the same covariate pattern $\boldsymbol{x}_{pr}$ and they will allow us to assess the same contrast of interest to infer on the log OR on detection and a $\log_2$ FC between conditions, respectively. Assuming independence between both statistics under the null hypothesis (see section 10.7. Appendix), the hurdle model allows to assess the omnibus null hypothesis of no differential detection and no differential expression by combining the p-values of MSqRob without imputation and of the quasi-binomial model component. To this end, we first transform the p-values to z-values and combine them in a chi-square statistic:

$$\chi^2_{\text{hurdle}} = z^2_{\text{MSqRob}} + z^2_{\text{quasi-binomial}} \quad \text{(Eq. 10.10)}$$

If an MSqRob $\log_2$ FC can be estimated the chi-square statistic follows a chi-square distribution with 2 degrees of freedom under the omnibus null hypothesis, otherwise the corresponding p-value is equivalent to that of the quasi-binomial model.

Two-stage inference

We used stageR version 1.3.29 to implement a two-stage inference procedure [22]. It is a two-stage testing paradigm that leverages power of aggregating multiple tests per protein (here a test for differential detection, DD and differential abundance, DA) in the screening stage. Upon rejection of the omnibus null hypothesis, stageR performs post-hoc tests to assess on DD and DA. We adopt the Benjamini-Hochberg False Discovery Rate (FDR) procedure on the aggregated tests in the first stage. In the post-hoc analysis, we use the modified Holm procedure implemented in stageR to control the Family Wise Error Rate for the DD and DA tests within a protein at the FDR-adjusted significance level of the first stage.

Multiple testing correction

All p-values for MSstats, MSqRob, the quasibinomial and the hurdle model were corrected for multiple testing using the Benjamini-Hochberg FDR correction. Perseus uses a permutation-based FDR based on 250 iterations.

10.4.6. Code availability

All scripts to reproduce the results in this contribution and in supplementary material are available at https://github.com/statOmics/MSqRobHurdlePaper.

10.5. Acknowledgements

L.G. is supported by a Ph.D. grant from the Flanders Innovation & Entrepreneurship agency, Flanders (Agentschap Innoveren & Ondernemen – Vlaanderen) entitled 'Differential proteomics at peptide, protein and module level' (141573). L.M. acknowledges funding from the Research Foundation Flanders (FWO) under Grant number G042518N. We also thank the students of the Statistical Genomics course, 2016/2017, Ghent University, who assisted us in assessing an initial implementation of the quasi-binomial regression component for their project work.

10.6. Author contributions

L.G. contributed R code, analyzed the data and wrote the paper, A.S. contributed R code and wrote the paper, K.G. and L.M. wrote the paper. L.C. contributed R code, conceived the idea and wrote the paper.

10.7. References

1. Hindupur, S.K. *et al.*, *The protein histidine phosphatase LHPP is a tumour suppressor.* Nature, 2018. **555**: p. 678.
2. Webb-Robertson, B.-J.M. *et al.*, *Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics.* Journal of Proteome Research, 2015. **14**(5): p. 1993-2001.
3. Lazar, C. *et al.*, *Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies.* Journal of Proteome Research, 2016. **15**(4): p. 1116-1125.

4. Tyanova, S., T. Temu, and J. Cox, *The MaxQuant computational platform for mass spectrometry-based shotgun proteomics.* Nature Protocols, 2016. **11**(12): p. 2301-2319.
5. Tyanova, S. *et al.*, *The Perseus computational platform for comprehensive analysis of (prote)omics data.* Nature Methods, 2016. **13**: p. 731.
6. Choi, M. *et al.*, *MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments.* Bioinformatics, 2014. **30**(17): p. 2524-2526.
7. Karpievitch, Y.V., A.R. Dabney, and R.D. Smith, *Normalization and missing value imputation for label-free LC-MS analysis.* BMC Bioinformatics, 2012. **13 Suppl 16**: p. S5.
8. Paulovich, A.G. *et al.*, *Interlaboratory Study Characterizing a Yeast Performance Standard for Benchmarking LC-MS Platform Performance.* Molecular & Cellular Proteomics, 2010. **9**(2): p. 242-254.
9. Goeminne, L.J.E., K. Gevaert, and L. Clement, *Peptide-level Robust Ridge Regression Improves Estimation, Sensitivity, and Specificity in Data-dependent Quantitative Label-free Shotgun Proteomics.* Molecular & Cellular Proteomics, 2016. **15**(2): p. 657-668.
10. Goeminne, L.J.E. *et al.*, *Summarization vs Peptide-Based Models in Label-Free Quantitative Proteomics: Performance, Pitfalls, and Data Analysis Guidelines.* Journal of Proteome Research, 2015. **14**(6): p. 2457-2465.
11. Blein-Nicolas, M. and M. Zivy, *Thousand and one ways to quantify and compare protein abundances in label-free bottom-up proteomics.* Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 2016. **1864**(8): p. 883-895.
12. Liu, K. *et al.*, *Relationship between Sample Loading Amount and Peptide Identification and Its Effects on Quantitative Proteomics.* Analytical Chemistry, 2009. **81**(4): p. 1307-1314.
13. Webb-Robertson, B.-J.M. *et al.*, *Combined Statistical Analyses of Peptide Intensities and Peptide Occurrences Improves Identification of Significant Peptides from MS-Based Proteomics Data.* Journal of Proteome Research, 2010. **9**(11): p. 5748-5756.
14. Dicker, L., X. Lin, and A.R. Ivanov, *Increased Power for the Analysis of Label-free LC-MS/MS Proteomics Data by Combining Spectral Counts and Peptide Peak Attributes.* Molecular & Cellular Proteomics, 2010. **9**(12): p. 2704-2718.
15. Chen, Y.-Y. *et al.*, *IDPQuantify: Combining Precursor Intensity with Spectral Counts for Protein and Peptide Quantification.* Journal of Proteome Research, 2013. **12**(9): p. 4111-4121.
16. Doll, S. *et al.*, *Region and cell-type resolved quantitative proteomic map of the human heart.* Nature Communications, 2017. **8**(1): p. 1469.
17. Wang, J. *et al.*, *Myotilin dynamics in cardiac and skeletal muscle cells.* Cytoskeleton, 2011. **68**(12): p. 661-670.
18. Huang, J. *et al.*, *Myocardin regulates BMP10 expression and is required for heart development.* The Journal of Clinical Investigation, 2012. **122**(10): p. 3678-3691.
19. Beretta, L. and A. Santaniello, *Nearest neighbor imputation algorithms: a critical evaluation.* BMC Medical Informatics and Decision Making, 2016. **16**(3): p. 74.
20. Ruppert, D., M.P. Wand, and R.J. Carroll, *Semiparametric Regression*. 2003: Cambridge University Press.
21. McCullagh, P. and J. Nelder, *Generalized Linear Models SECOND EDITION*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. 1989: Chapman and Hall/CRC 532.
22. Van den Berge, K. *et al.*, *stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage.* Genome Biology, 2017. **18**(1): p. 151.

10.8. Appendix

Under the null hypothesis of no differential abundance, the combined test statistic χ^2_{hurdle} ($\text{Eq. } 10.10$) will follow a chi-squared distribution with two degrees of freedom if both z-statistics are independent. To assess this assumption, we performed a mock analysis on the repeats of the spike-in conditions within each lab. We also included the combination of spike-in condition and lab as a blocking factor in the analysis (see Table 10.1). This set-up ensures that none of the proteins are differentially abundant.

Table 10.1. Overview of the mock treatment levels and the condition-lab combinations for each MS run in the mock analysis.

Condition:lab	**Mock treatment**		
	One	**Two**	**Three**
6A_Orbitrap_86	6A1	6A2	6A3
6A_Orbitrap_65	6A4	6A5	6A6
6A_Orbitrap_56	6A7	6A8	6A9
6B_Orbitrap_86	6B1	6B2	6B3
…	…	…	…

The results of the three pairwise comparisons, both for the MSqRob and for the quasi-binomial model compontent of the hurdle model are shown in Fig. 10.3. and confirm that the estimates for the $\log_2$ FCs and the log ORs are not correlated in the mock analysis.

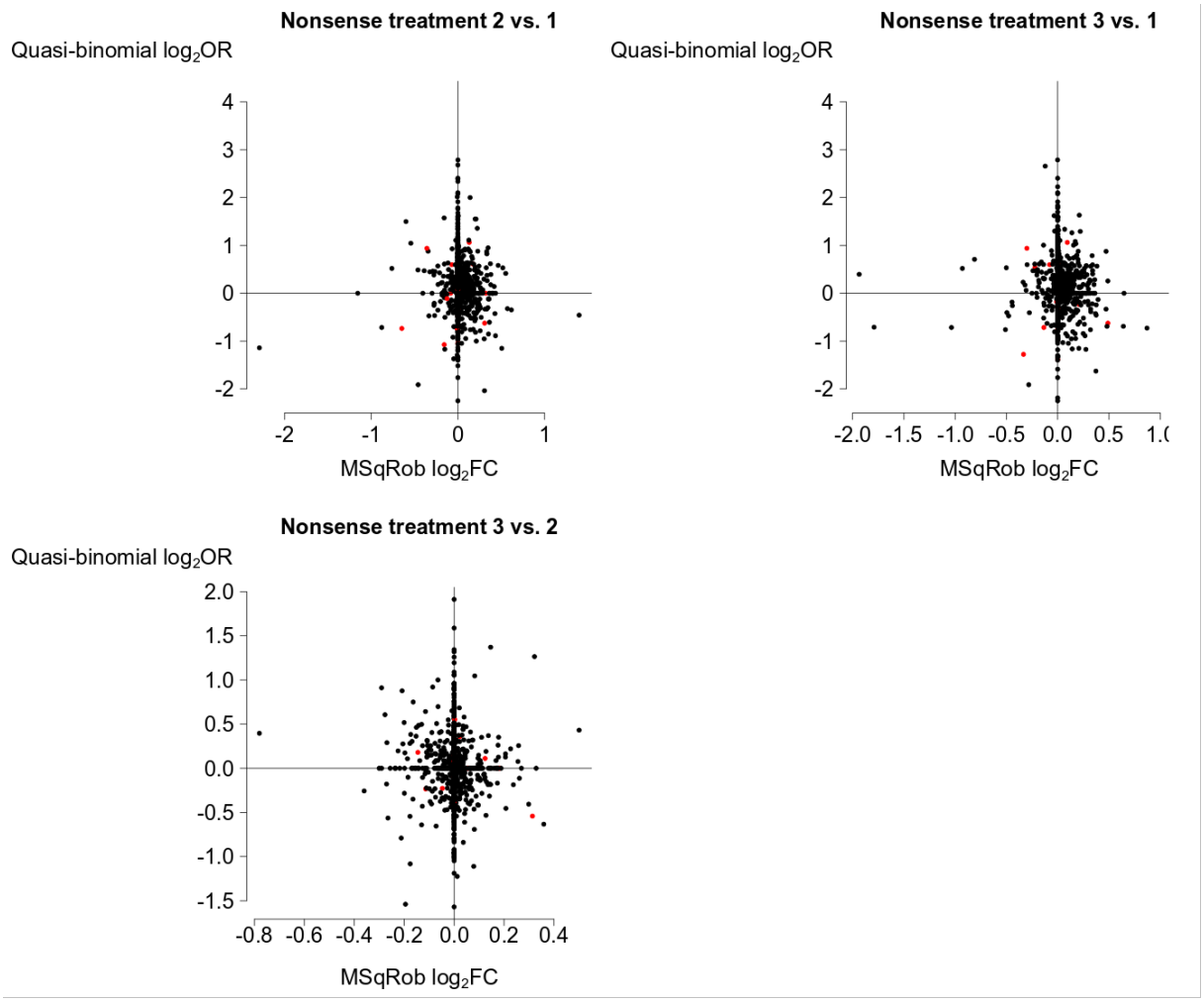


Figure 10.3. There seems to be no correlation between the estimates from the MSqRob and quasi-binomial model component. The scatter plots show the $\log_2$ FC estimates on the horizontal axis and log OR estimates on the vertical axis of the hurdle model for all proteins in the CPTAC dataset for which a $\log_2$ FC could be estimated in all three comparisons of the mock analysis (UPS1 proteins are indicated in red, while yeast proteins are indicated in black).

PART III: DISCUSSION AND RESEARCH PERSPECTIVES

11. DISCUSSION

In recent years, enormous progress has been made in MS instrumentation and quantitative proteome analysis protocols, not only on how experiments are planned and executed, but also on how the data are analyzed. Given the quite large costs of a proteomics experiment, it is important to extract the maximum out of the obtained data. Nonetheless, because of historical reasons, the number of statisticians studying proteomics data is limited compared to the number of statisticians in the transcriptomics field. Indeed, proteomics initially only aimed for protein identification and not quantification, and the low duty cycles rendered proteomic analyses laborious, time-consuming and costly. Therefore, biological repeats were often not included in these early studies, which made it impossible for statisticians to infer the results of these studies towards the population (see 4.2.1). The enormous technological advances in instrumentation over the latest 10 years led to a massive increase in throughput and reproducibility. This enabled the adoption of experimental designs that include proper biological replication, which is now required by all major proteomics journals. This evolution towards more statistical rigor occurred much earlier in the transcriptomics field, and most statisticians who used to work on microarray data migrated towards RNA-Seq data as they were used to these types of designs.

Given the importance of quantitative proteomics in the study of diseases, it is unacceptable that biologically relevant proteins go unnoticed because of suboptimal protein quantification strategies [1]. Therefore, I focused in my PhD project on understanding and improving quantification in label-free shotgun proteomics. Nowadays, large and complex experimental designs are being analyzed more and more routinely and many different methods to quantify proteomes have emerged of which the performances differ widely. For a non-expert, the choice of different methods presented in the literature seems overwhelming. In this discussion, I do not aim at giving a complete overview of every single method that is available now, as new approaches are emerging every other day. Alternatively, I present my view on why certain approaches work better than others in order to provide experimentalists with basic tools to critically reflect on new methods and select that workflow that is the most adapted to their needs. Hopefully, this work will not only help researchers in getting the most out of their data, but also inspire data scientists to improve existing methods.

11.1. Comparing performances

An optimal differential analysis method can discern truly differentially abundant proteins (true positives) from non-differentially abundant proteins (false positives) as good as possible. Therefore, the most important performance metric is the protein ranking: truly differentially abundant proteins should appear near the top of the ranking, while proteins with very little evidence for differential abundance should be found near the bottom of the ranking. Once a ranking is obtained, a threshold should be set above which all proteins are deemed significantly differentially abundant at a certain FDR level. In practice, only a few proteins that are deemed differentially abundant with a high certainty will be selected for further experimental validation. Setting the boundary straight between differentially abundant and non-differentially abundant proteins is therefore of secondary importance compared to generating a useable ranking in the first place.

High variabilities in peptide intensities relative to protein fold changes are characteristics of label-free shotgun proteomics experiments. Importantly, much of this variability is non-random

(e.g. due to differences in ionization efficiencies), and methods that fail to appropriately model the different sources of variability in the data are doomed to return many false positives. Indeed, the high variability in observed peptide intensities might sometimes lead to large apparent fold changes with low apparent uncertainties, especially when very few peptides are identified for the corresponding protein. Such false positives will however fail to pass the validation stage in follow-up experiments, which results in a waste of valuable resources and contributes to the replication crisis in life sciences.

Not all methods are easy to include in a comparative study as some published data analysis methods do not provide scripts, let alone user-friendly interfaces. Other tools function like a black box: they do not provide any insight in how protein level fold changes are calculated, nor do they give any indication on the certainty of the quantification. Finally, many scripts and tools have been abandoned because the scientist that introduced them has moved on to another project. Luckily, the usage of such methods is expected to diminish strongly, as the proteomics field is undergoing a rapid transition towards more transparent and reproducible ways of data analysis.

Benchmark datasets, in which the ground truth is known, are necessary to compare methods. There are three main ways of generating benchmark datasets: computationally simulating a dataset, computationally introducing differences in a biological dataset, and generating a spike-in dataset in the lab.

Simulated datasets are entirely computer-generated and therefore easy to create. However, simulation studies often fail to capture all sources of variability in a biological dataset. For example, ionization competition effects are rarely considered in simulation studies, and there are many other known and unknown influences on peptide intensity that interact amongst each other in complex ways, that are very difficult to predict. Moreover, it is often tempting to simulate the data under assumptions that are very close to those of the model under study or even use this model to generate the data, therefore artificially boosting its performance.

Artificial differences between different treatments can also be generated by computational manipulation of an existing dataset. The intensities of all peptides mapping to a group of proteins can for example be computationally decreased or increased in half of the samples from the control arm of an existing case-control experiment. Although such datasets have a representative data structure for non-differentially abundant proteins, differential effects of e.g. enzymatic digestion or ionization are generally ignored for the differentially abundant proteins. Furthermore, if all peptide intensities for each differentially abundant protein are changed by a fixed amount, the variability in the peptide intensities might be underestimated compared to a biological dataset, which makes it easier to detect the differentially abundant proteins in the artificial dataset.

In spike-in studies such as the CPTAC study 6, different amounts of proteins are spiked into a background proteome and analyzed on the HPLC-MS system. Although the generation of such datasets requires setting up an actual MS experiment, well-executed spike-in studies capture the data structure of biological datasets. When setting up a spike-in study, it is important to incorporate biological repeats. Unfortunately, many spike-in studies, such as CPTAC, spike the samples only once and repeatedly analyze the same samples. The variability in such datasets is therefore only technical in nature and inevitably lower due to the absence of biological repeats.

In chapter 8, we used the CPTAC dataset to rigorously compare summarization-based methods to peptide-based methods. We provided an objective comparison and showed that methods starting from normalized peptide intensity data are superior to methods using a summary step, and this in terms of sensitivity, specificity, accuracy and precision. A crucial element of that study is that we provided clarity on the different mechanisms causing the differences in performance. This insight is extremely important, both for practitioners who need to understand why certain methods outperform others, and for data scientists who want to develop new quantification methods.

Naive summarization methods such as mean and median summarization produce biased fold change estimates because their protein abundance summaries are based on different peptides in each run without properly controlling for peptide-specific effects. MaxLFQ summaries are theoretically unbiased, but its protein summaries are inefficiently estimated as MaxLFQ only uses the PSM intensities that are shared between each pair of runs. This unstable estimation of the MaxLFQ summaries is particularly apparent in the samples with low spike-in concentrations, where very few peptides are found due to intensity-dependent missingness and hence very few ratios are used to calculate the MaxLFQ estimates. Furthermore, summarization-based methods cannot propagate the differences in uncertainty on the summary estimates to the downstream analysis because they ignore the fact that the standard errors on the summaries are dependent on the number of peptides used in each summary. Peptide-based models, on the contrary, can naturally incorporate a peptide effect, hence avoiding bias and accounting for the fact that estimates based on many peptides are more precise than those based on only a few peptides.

When constructing a peptide-based model, it is important to correctly model different sources of variability as failure to correctly take the data structure into account might lead to imprecise or biased results. Indeed, the treatment of interest is not the only source of variability and variability due to different peptides and sample-to-sample variability should also accounted for. Do note however, that including too many parameters in a model will lead to over-parameterization, the phenomenon whereby a model is too complex for the dataset at hand (i.e. it has too many parameters compared to the amount of data). Models for proteins with very sparse peptide evidence are particularly at risk for over-parameterization. An over-parameterized model will be tuned too well to the data: part of the random residual variability will be absorbed by the model's parameter estimates as if that variability reflects meaningful variation that can be explained by the model. As a result, the model's parameter estimates will be extremely imprecise.

In chapter 8, we also showed that including a fixed sample effect leads to over-parameterization for most proteins in the CPTAC dataset. This is not surprising as most proteins are identified with only a few peptides per sample. Hence, fold change estimates and variance components for proteins with very few peptides will be unstable. More importantly, estimating a sample effect removes the between-sample variability from the analysis. Ignoring this between-sample variability causes a vast underestimation of the standard errors on the fold changes and will thus lead to an increase the number of false positives. This is the reason why the peptide-based model without sample-effect was shown to be more accurate and more precise. Indeed, not only is this model less prone to over-fitting, the standard error on the fold change estimates is also more correctly estimated as both the within- and between-sample variances are lumped together in the model's error term. However, a model without a sample effect does not account for the fact that peptides within the same sample are correlated with each other. Using a mixed model in which the sample effect is encoded as random provides a

solution as random effects allow to model the correlation structure in the data in a natural way by correctly taking the between-sample variance into account while still acknowledging the within-sample correlation. This explains the slightly better performance of the mixed model compared to the peptide-based model without a sample effect.

We observed that the performance gain of the mixed model compared to the peptide-based model without a sample effect is very limited in the CPTAC dataset. This is likely because the CPTAC dataset is highly artificial as UPS1 proteins were spiked into a yeast background in five different concentrations and these samples were repeatedly analyzed in multiple labs. Hence, the dataset only contains technical variability. In a biological experiment, one expects the differences between the individual samples to be bigger because of the additional biological variability between samples. Consequently, the within-sample correlations will be stronger and the contribution of the sample effects to the total variability in the data will be larger.

11.2. The impact of MSqRob

The fact that peptide-based models outperform naive summarization-based models and that mixed models have a theoretical advantage for dealing with hierarchical data inspired us to develop MSqRob. This was motivated by the observation that state-of-the-art peptide-based mixed models declare many proteins as differentially abundant for which there is in reality very little evidence as based on visual inspection of the data. These false positives were often proteins for which very little peptides were identified. Sometimes, only one or two peptides seemed to drive the differential abundance estimate, but more often, the fold change estimates of such proteins were relatively small. Closer inspection of these models revealed that their residual variances and those of the sample-effects were very close to zero. In other words, these models were clearly over-fitted. MSqRob tackles these issues with three modular extensions: ridge regression, empirical Bayes variance estimation and M-estimation with Huber weights.

Ridge regression imposes a penalty on the magnitude of a parameter by which the parameter estimate is shrunk towards zero. This will introduce a small bias compared to an ordinary least-squares estimation, but causes a large gain in stability, which protects against over-fitting. In the literature, it has been shown that shrinkage estimators outperform ordinary least squares estimators in terms of root mean squared error [2-4].

To estimate our parameters, we make use of the link between ridge regression and mixed modeling. The best linear unbiased predictions of random effects in a mixed model can indeed be considered as shrinkage estimates where the estimated ridge penalty $\hat{\lambda}_\beta$ on a parameter β is equal to $\hat{\lambda}_\beta = \frac{\hat{\sigma}^2}{\hat{\sigma}_\beta^2}$ with $\hat{\sigma}^2$ the residual variance and $\hat{\sigma}_\beta^2$ the estimated variance on the parameter β. This shrinkage estimator is equivalent to assuming a normal prior with mean 0 on the parameter β in the Bayesian framework. Because of this, parameter estimates for proteins with very few identified peptides will be strongly shrunken towards the prior (zero). Contrary, in protein models with many identified peptides, the parameter estimates will be mainly driven by the data.

To stabilize the residual variance, we make use of empirical Bayes variance estimation, which relies on the fact that thousands of protein models are fitted in parallel. The data structures of protein models with many identified peptides might be useful templates for the model structures of proteins for which there are too little peptides to precisely estimate the model structure. It therefore makes sense to squeeze the unstable variance estimates for all proteins towards a

common pooled variance estimate. Similar to ridge, the squeezing of the variances will be done in such a way that the residual variances for protein models with very sparse peptide evidence will be strongly shrunken towards the common variance, while the shrinkage effect will be smaller for proteins with many identified peptides. Moreover, the resulting moderated t-statistic will follow a t-distribution with increased degrees of freedom. Implemented in the popular R package limma, empirical Bayes variance estimation is routinely performed on microarray data [5], but has also demonstrated its use in the proteomics field [6]. limma is executed after the model fitting since only the residual variances and the degrees of freedom for each model are required as input. The limma procedure is extremely fast and provides a noticeable boost in performance: proteins that would otherwise be very significant due to seemingly small variances originating from sparse peptide evidence are much less significant with the moderated t-test.

Down-weighing outliers with robust M-estimation with Huber weights, finally, reduces the impact of outlying peptides, which is especially important for proteins with sparse peptide evidence. Such outliers correspond for example to peptides with differential modification statuses, misidentified peptides or even peptides with ionization efficiencies that are drastically different from the bulk of the peptides. A disadvantage of this procedure is that it markedly increases MSqRob's run time as each model needs to be refitted until convergence of the weights. In our most recent implementation of MSqRob, we did some optimizations to increase the speed of the M-estimation procedure. First of all, we ensured that the models were iteratively updated instead of letting R build a completely new model object with each iteration. Second, we increased the tolerance level of the loop and reduced the maximum number of iterations since we noticed that most models already reached convergence after one or two iterations.

An important remark is that MSqRob's major boost in performance truly originates from the combined effects of these three improvements: ridge regression, empirical Bayes variance estimation and M-estimation with Huber weights. Indeed, irrespective of the order in which these improvements are added to the mixed model, the first two additions only cause slight gains in performance. However, the third addition generates a large performance boost, irrespective of whether it is ridge regression, empirical Bayes variance estimation or M-estimation. An important advantage of MSqRob is that it reduces the need for filtering out sparse proteins. Indeed, fold changes and variance estimates for proteins with sparse peptide evidence are automatically shrunken unless there is enough evidence for differential abundance in the data.

A superior performance is not at all a guarantee for a new tool to gain wide adoption by a scientific community. Hence, we made MSqRob available as a Shiny App. Our Shiny App is an important asset as it provides an easy-to-use point-and-click interface to users who are not proficient in R scripting. In my opinion, such an interface is essential for the wider adoption of any data analysis method. This is one of the reasons for the huge success of the MaxQuant-Perseus workflow in the proteomics field which allows users to perform the whole data analysis workflow from raw files to quantification in a single, intuitive point-and-click environment. However, we need to note that the Perseus workflow is clearly outcompeted by peptide-based models because the MaxLFQ summaries are inefficient and Perseus cannot handle the blocking effect for lab.

MSqRob's graphical user interface needs a proper preprocessing pipeline. Therefore, we offer the most commonly-used preprocessing algorithms. Alternatively, a user can import preprocessed data into MSqRob's Shiny App. Upon the request of our users, I made the

MSqRob GUI compatible with MaxQuant, the open mzTab data standard [7], Progenesis QI for proteomics (Nonlinear Dynamics, Newcastle, UK), and our own moFF tool [8]. Furthermore, I included the automatic generation of density plots before and after preprocessing and a multidimensional scaling (MDS) plot after preprocessing. Such diagnostic plots are extremely important for exploratory analyses, in which the data quality and the quality of the preprocessing workflow are evaluated. Density plots might for example reveal that the peptide intensities in some runs deviate strongly from those in other runs, which might point to runs of inferior quality (Fig. 11.1). MDS plots are useful as they clearly show the major sources of variability and might even reveal sources of variability that were unaccounted for. This might probe the data analysist to re-discuss the experimental protocol and experimental design with the experimenters to assess if e.g. a blocking factor might have been forgotten in the analysis.

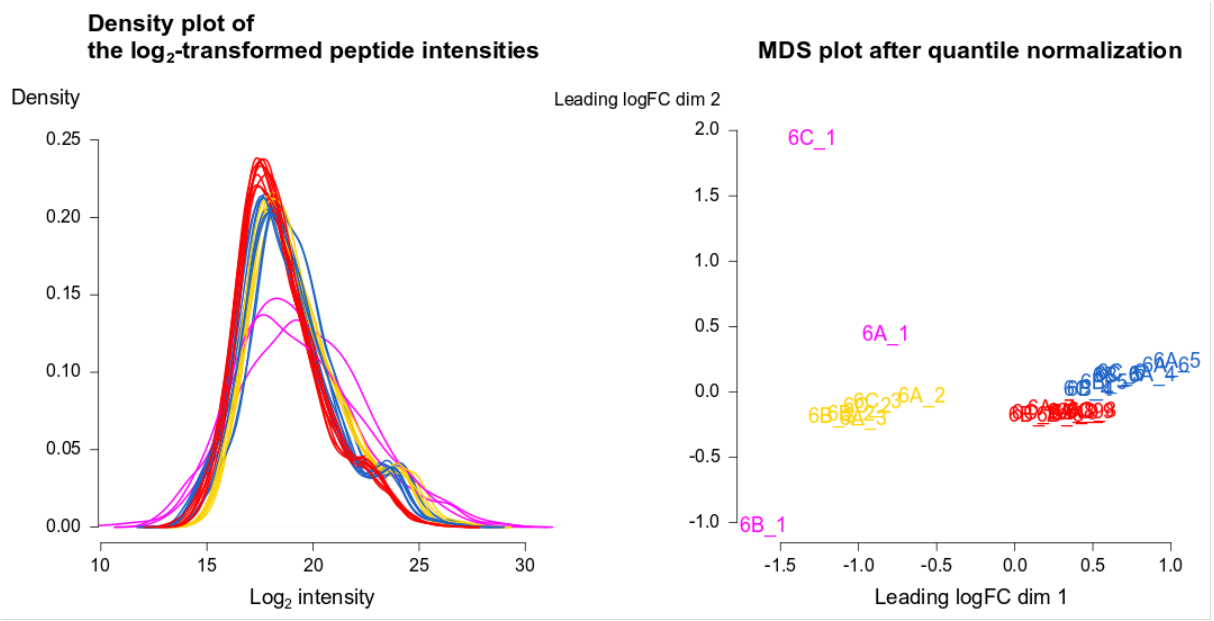


Figure 11.1. Left: density plots of the $\log_2$-transformed peptide intensities in the CPTAC dataset [9] after artificial introduction of noise and missing values in the first three runs from lab 1 (purple). The other runs are colored according to the lab that generated the data (red: orbitrap at site 56, yellow: orbitrap at site 65 and blue: orbitrap at site 86). Such a pattern is typical for MS runs in which something went wrong during the analysis (e.g. with the HPLC flow or during peptide ionization). When some runs deviate strongly from others, it is always advised to inspect the raw data (e.g. are there much more missing values) and to try to find out what exactly went wrong during the analysis. It is generally advisable to remove runs of poor quality from the analysis. Right: MDS plot after quantile normalization. As demonstrated in the introduction, runs cluster together according to the lab that generated the data. This demonstrates the usefulness of the blocked design. The poor-quality runs clearly deviate from the other runs in lab 1 but are still closer to those runs than to the runs in labs 2 and 3.

Volcano plots show the $-\log_{10}$-transformed p-values as a function of the $\log_2$ fold change estimates for all fitted proteins. This gives a general idea of the number of up- and down-regulated proteins in the comparison of interest as well as of the factors contributing to statistical significance for each protein. Indeed, proteins with relatively high p-values and large fold changes can be expected to be highly variable and/or identified by very little peptides. Conversely, the significance of proteins with extremely low p-values and relatively small fold changes are either driven by a relatively large number of peptides of which the intensities are relatively stable, or by lower numbers of peptides for which the variance was underestimated due to the larger uncertainty on variance estimates for small sample sizes.

Clicking on a dot or a protein in the protein list reveals a detail plot that shows the underlying preprocessed $\log_2$-transformed intensities for all peptides in all samples for that particular protein. This is very useful information as it allows to visually inspect the underlying data that determine each protein's significance. Therefore, users can focus on validating proteins for which there is the most visual evidence for differential abundance (i.e. a clear difference in abundance and enough identified peptides). If proteins with very weak or even negligible visual evidence for differential abundance would show up as differentially abundant anyway, this might be an indication for experimental flaws or flaws in the data analysis prior to differential analysis (e.g. erroneous protein inference for one or more peptides). If several of such proteins share some characteristics in their data structure[35], this can provide a first clue about why MSqRob fails in those particular instances. This information is very useful for statisticians who aspire to improve MSqRob's quantification algorithm. Alternatively, the detail plot can help users to uncover potentially interesting biological phenomena such as one or more peptides that behave differently from the other peptides in a protein. Such peptides might indicate differential modification statuses or changes in protein splicing resulting from the treatment.

When I started my PhD in September 2013, the majority of the proteomics data analysis pipelines consisted of preprocessing pipelines that were often borrowed from the microarray context followed by summarization-based methods. Apart from a few exceptions, these pipelines were often constructed *ad hoc* without any statistical rationale. Some Bayesian methods existed, but these are slow and therefore not routinely used in practice [10]. MSstats had just been published as a peptide-based model and gained a considerable user base over the years [11]. However, such peptide-based models suffer from over-fitting (chapter 8).

MSqRob provides a statistically sound data analysis pipeline that copes better with proteins with sparse peptide evidence. This should result in an improved ability to detect differential abundance, which will in turn lead to improved biological insights. To facilitate the adoption of MSqRob by experimenters, we provide a GUI next to our R package. Since the publication of our tutorial paper in 2018, MSqRob is gaining some traction. Indeed, MSqRob has already been used in a paper by other groups on carbon fixation in *Arthrospira platensis* cells [12] and in a multi-omics study that discovered potential therapeutic drug targets for diffuse large B-cell lymphoma [13]. We also demonstrated MSqRob ourselves in a recent study in which we demonstrated Immune-responsive gene 1 to be a target of A20 [14]. Here, many A20 targets could not be quantified with Perseus as they had very sparse peptide evidence and were therefore filtered out. MSqRob was however sensitive enough to quantify such low-abundant proteins. Finally, the impact of MSqRob reaches beyond academia. Indeed, the development and dissemination of more efficient and user-friendly data analysis methods and software for differential proteomics is also immediately relevant for other research and development in sectors where high-throughput proteomics profiling is used, such as the clinical, biomedical and pharmaceutical sectors.

11.3. The impact of the hurdle model

We expect our hurdle model to have a major impact on the field. Indeed, missing values have many causes and can be seen as a dataset-dependent mix of missingness completely at random and not at random. In reality, the precise mechanisms behind missing values remain unknown and imputation is often used to avoid having to remove too much information from

[35] These characteristics can be very diverse, e.g. proteins with weak evidence for differential abundance might have one or more outlying peptide intensities that drive the fold change, or their fold changes might be driven by the intensities in only a few lower-quality runs, etc.

datasets. In chapter 8, we demonstrated the unpredictable consequences of imputation. Imputation under the missing completely at random assumption cause an upwards quantification bias for low-abundant proteins, while imputation based on missingness by low abundance might lead to a downwards bias for highly-abundant proteins.

MSqRob does not impute missing values by default to avoid drawing conclusions based on unrealistic imputation assumptions. The hurdle model combines the power of MSqRob with complementary information that is available from peptide counts. Indeed, count-based approaches can naturally handle missing values as zero counts. Moreover, the hurdle model's fold change and odds ratio estimates have meaningful interpretations in terms of differential abundance and differential detection probability, respectively. This strongly expands the range of proteins that can be quantified as it is now possible to provide accurate statistical inference on proteins for which all peptides are missing in one condition.

The hurdle model is also highly promising to detect differential modification states in proteomics datasets. Indeed, many biological treatments cause changes in protein signaling cascades that result in changes in protein modifications. Such modifications are often completely absent before (or after) the treatment. Since most post-translational protein modifications are very low abundant and hence not easily detected in a typical shotgun proteomics experiment, this application of the hurdle model will be mainly suitable for studies that enrich for certain post-translational modifications.

11.4. Controlling the false discovery rate

I noticed that MSqRob is often rather liberal in its false discovery rate cut-offs. One reason affects all methods and has been explained in chapter 8 as being specific for the CPTAC dataset: the fact that high amounts of UPS1 proteins were spiked into a yeast background suppresses the yeast peptide ion intensities, which makes that, if two different spike-in conditions are compared, yeast proteins can be found as false positives with fold changes opposite to that of the spiked UPS1 proteins. We therefore advise researchers not to spike too much of an internal calibrant, and to deplete highly abundant proteins such as albumin from blood samples, if possible. A reason for MSqRob being rather liberal compared to other methods might be that peptide-level proteomics data is highly unbalanced due to many missing values. This makes it impossible to assign degrees of freedom to the moderated t-test. Indeed, the determination of a correct null distribution for contrasts based on finite samples in unbalanced mixed models is a long-standing unresolved issue. It is known that the null distribution of each contrast is asymptotically[36] normal, but with finite sample sizes, these distributions have fatter tails and are difficult to characterize. The Kenward-Roger [15] and Satterthwaite [16] approaches are well-known approximations in such cases.

It is however reasonable to assume that the upper bound on the degrees of freedom is determined by counting the number of peptides and subtracting the trace of the Hat-matrix. The Hat matrix $\boldsymbol{H}$ is defined as the matrix that transforms the fitted values $\hat{y}$ into the response values y:

$$\hat{y} = \boldsymbol{H}y \qquad (\text{Eq. } 11.1)$$

The trace (i.e. the sum of the diagonal elements) of $\boldsymbol{H}$ provides an indication for the effective degrees of freedom that are lost by fitting the model. When shrinkage estimators are used, the

[36] I.e. if the sample size approaches infinity.

loss of degrees of freedom will be lower than the number of parameters in the model. A reasonable lower bound on the degrees of freedom is the total number of samples minus the number of parameters corresponding to all effects that are unrelated to samples or peptides. This is the number of degrees of freedom that would be used if the data were summarized to the protein level and fitted with an ordinary least squares model. Initially, MSqRob used the total number of peptides minus the trace of the Hat matrix to calculate the degrees of freedom. However, for the hurdle model (chapter 10), we changed the degrees of freedom used for statistical inference to the lower bound, which makes MSqRob slightly less liberal. For example, in the HEART dataset, the difference between both approaches is 1946 significant proteins versus 1527 significant proteins at a 5% FDR cut-off level when comparing the atrial versus the ventricular regions. The difference is 74 vs. 12 significant proteins at 5% FDR when comparing the left vs. the right atrium and 45 vs. 0 when comparing the left vs. the right ventricle.

A second reason for MSqRob's perceived liberal performance on certain datasets is the fact that the models are often over-parameterized, especially for proteins with sparse peptide evidence. With shrinkage estimation, over-parameterization in the strict sense is not an issue, since inestimable model parameters will be shrunken to values very close to zero. However, due to the shrinkage of MS run effects and other potential batch effects towards values very close to 0, their contribution in the overall model variability will be grossly underestimated. This results in underestimated random effect variances and therefore too liberal results. Indeed, in the CPTAC dataset, the within-sample correlations are expected to be small because each run is a pure technical repeat: the same sample is measured over and over again.

11.5. MSqRob compared to other methods

MSqRob is not the only method that is used for differential protein abundance analysis. Competitors such as Perseus and MSstats were already widely used when MSqRob was first published in 2015, and over the years, new approaches for differential protein abundance analysis have emerged. In this section, I will give a brief overview of MSqRob's performance and user friendliness as compared to its main competitors and some new emerging approaches (summarized in table 11.1).

Table 11.1. Overview of the advantages and disadvantages of MSqRob's main competitors and some newly-published approaches.

Method name	Advantages	Disadvantages
Perseus [17]	User-friendly, comprehensive GUI	Inefficient summarization, only pairwise comparisons, arbitrary, arbitrary fudge factor
MSstats [18]	User-friendly, comprehensive R package, data-driven censoring threshold, GUI	No correct modeling of complex hierarchical designs, no continuous covariates
DanteR [19]	data-driven censoring threshold, GUI	Windows only
Koopmans *et al.* [20] approach	Data-driven censoring threshold, shared prior distributions: borrow strength across proteins	Not implemented, only pairwise comparisons, requires Markov Chain Monte Carlo (MCMC) sampling and is therefore likely rather slow
ProteoSign [21]	Convenient web application	Not methodologically innovative in terms of differential analysis

Proteus [22]	Convenient R package	Not methodologically innovative in terms of differential analysis
Diffacto [23]	Down-weighing or removal of incoherent peptides	Data structure not taken into account
MS-EmpiRe [24]	Relies on empirical distributions	Data structure not taken into account
CSNorm [25]	Non-parametric	Data structure not taken into account, incorrect permutation strategy, low power, too liberal FDR
DAPAR (pepa.test function) [26]	Includes shared peptides	Only pairwise comparisons
BayesENproteomics [27]	Includes peptide-treatment interactions, elastic net penalization, combines MCAR and MNAR imputation, weighs peptides by scores, pathway-level inference	Relies on Markov Chain Monte Carlo sampling and therefore likely rather slow

Perseus is widely used, but its appeal cannot be contributed to the performance of its default workflow, but rather to its versatility, user-friendliness to the non-expert and its seamless integration with the output of the popular, free search engine MaxQuant. Part its popularity might also be due to "inbreeding effects", whereby a protocol or data analysis procedure is repeated simply because it worked in the past or because it is the default approach and not because it provides the best possible results. This also applies for seemingly abandoned tools and home-brew scripts that are still used locally.

The default Perseus workflow is based on MaxLFQ summarization and although MaxLFQ does account for peptide-specific effects, it seems to perform slightly suboptimal compared to MSqRobSum summarization. More fundamentally, Perseus only allows pairwise t-tests, but no regression modeling. Compared to pairwise t-tests, regression models benefit from an increased power to estimate the residual variance because their variance estimate is based on all the observations in the model. This includes those observations that are not directly involved in the comparison at hand. Moreover, t-tests are unable to model additional sources of variability such as block effects. Finally, SAM is Perseus' only possibility to stabilize the t-tests' variance estimates. And, although the effects of SAM in Perseus are quite similar to those of limma's empirical Bayes variance estimation, SAM requires the user to set an arbitrary fudge factor s_0 which is added to the denominator of the t-test statistic [28], while limma's moderated t-test statistics are data-driven. This makes the SAM approach very arbitrary. The most recent Perseus version (1.6.2.3) allows two-way ANOVAs with post-hoc testing, which allows to take a blocking factor such as lab into account. However, Perseus' post-hoc tests are only capable of determining whether a protein passes a certain pre-specified significance threshold but fail to return any p-values. It is therefore impossible to produce rankings based on this analysis. For all these reasons, MSqRob shows a large and consistently superior performance compared to Perseus.

The R/Bioconductor proteomics quantification package MSstats can be seen as our most direct competitor. MSstats provides options for data visualization, statistical modeling and inference. MSstats' current workflow imputes missing values with an accelerated failure time model, summarizes the data to the protein level with Tukey's Median Polish and fits a linear mixed model to the summarized data [29]. However, imputation can also be omitted, and summarization can also be done based on linear regression modeling. MSstats allows the user

to specify one treatment effect (“Condition”) and one effect for biological replicates (“BioReplicate”). If multiple treatments were used, each treatment combination should be assigned to a different level of the “Condition” variable. Similarly, multiple blocking effects can be modeled in MSstats by assigning a different level to each combination of blocking factors in the “BioReplicate” input. However, this approach does not allow fitting additive models because it automatically assumes interaction effects. Moreover, due to the constraint of a single treatment and a single blocking effect, it is not possible to correctly model hierarchical effects in MSstats. Also, contrary to MSqRob, MSstats can only model discrete, but not continuous covariates. MSqRob’s ability to model both makes it much more flexible as it can correct for continuous covariates such as age. Furthermore, MSqRob outcompetes MSstats in terms of sensitivity and specificity because MSstats does not include ridge regression, empirical Bayes variance estimation and M-estimation. Indeed, we demonstrated the superior performance of MSqRob compared to the old MSstats mixed model in section 9.1 and to the current default MSstats pipeline in chapter 10.

Some papers seem to only discuss parts of MSqRob’s improvements. The Bioconductor package PECA, for instance, uses moderated t-tests with limma’s empirical Bayes procedure to estimate differential abundance of each peptide. Protein-level fold changes are then calculated by averaging over the fold changes of all peptides corresponding to a protein [30].

The censored regression approach of Karpievitch *et al.* (2009) [31] is a fruitful approach to deal with missing values because it assumes a combination of missingness completely at random and missingness due to low abundance. These authors developed the R package DanteR, where they use the Karpievitch *et al.* censored regression model for imputation [19]. They also suggested to summarize the data based on a peptide-based model, which should in principle result in unbiased summaries [32]. The development of DanteR has now been frozen and superseded by the InfernoRDN package, which has a GUI that can be installed on Windows systems.

Similarly, Koopmans *et al.* [20] proposed an "empirical Bayesian random censoring model" that assumes a censoring threshold for every protein. This model uses all protein data to estimate the different variance components in the model. However, separate models are fitted for each pairwise comparison. This approach is less powerful than modeling the data from all treatments together in a single model because residual variance estimate will be based on less data and will therefore be less stable. Furthermore, the model assumes a missing-by-low-abundance mechanism but does not account for missingness completely at random. Fitting the model is also computationally intensive because it requires Markov Chain Monte Carlo (MCMC) sampling.

Do note that lack of computational speed might not always an issue, especially considered in the light of a complete proteomics study, that can easily span over twelve months. For large studies, even very complex models can be run on high-performance computers within reasonable timeframes. Nevertheless, users often need some time to get used to new software and this process goes much easier if the analysis can be done in a short timeframe on their own computers. For simple routine analyses, users will therefore often prefer faster methods over slower ones.

Some other novel approaches do not seem to be methodologically innovative. For example, the workflow of the ProteoSign web application only provides model fitting and empirical Bayes variance estimation with limma on summarized protein values [21]. Similarly, the R package Proteus seems to be a proteomics wrapper around the limma package [22]. Even though such

approaches might contribute to user comfort, they seem rather redundant as the first application of limma on summarized protein values already dates back to 2009 [6]. Indeed, as shown in Fig. 9.5, differential protein abundance analysis with limma based on MaxLFQ values performs better compared to simple t-tests but does not attain the performance of peptide-based linear regression models. This stresses the importance of explaining why certain approaches perform better than others so that researchers can focus their efforts on improving the state-of-the-art rather than on reinventing the wheel.

The Python workflow Diffacto weighs down or removes peptides with signals that are incoherent with other peptides of a protein, under the assumption that these peptides have a high probability of being misidentified [23]. This might also be relevant to dampen the impact of differential modification statuses on protein quantification. Diffacto calculates protein fold changes as the sums of weighted peptide ratios between treatments. This enables factoring out the peptide effects. Protein significance is determined via an F-test whereby the assignments of the samples are permuted between treatments to generate a null distribution. However, Diffacto does not allow users to correct for additional covariates and does not account for within-sample correlation, which will result in biased quantifications. Inference is limited to ANOVA-like research hypotheses, i.e. testing whether the average protein abundance in at least one treatment is different from the average protein abundance in the other treatments. Furthermore, there should be enough samples per treatment to allow a sufficiently powerful permutation analysis. Indeed, the authors note that for pairwise comparisons, they require at least five samples in each condition to allow proper FDR control on the random permutation test. Therefore, Diffacto will likely perform poorly compared to MSqRob. Nonetheless, the exclusion of peptides that are deemed unreliable is a novelty that might increase its performance.

The MS-EmpiRe R package, which has only very recently been made available on bioRxiv, uses MSqRob as a state-of-the-art benchmark [24]. Its normalization approach uses single linkage clustering: runs are added one-by-one to a cluster whereby the log-transformed peptide intensities in each run undergo a shift in median relative to the runs that are already in the cluster. Hence, this normalization approach does not take distributional changes into account. However, their approach to protein quantification is very innovative. It is known that even after $\log_2$-transformation, there is often still a mean-variance relationship in the data. Therefore, MS-Empire assigns each peptide to a discrete bin based on its average $\log_2$-transformed intensity over all samples and treatments. For each bin, an empirical distribution of the peptide $\log_2$-fold changes for each pairwise sample combination within, but not across treatments is constructed. When comparing two treatments, the $\log_2$-fold change between those treatments for each peptide is compared to the empirical distribution of its corresponding bin to calculate a peptide-level z-value quantile. After correcting for outlying z-values, peptide-level z-values are summed to obtain protein-level z-values. These protein-level z-values are compared to a standard normal distribution to obtain protein-level p-values.

Based on a recently-published spike-in study [33], it was shown that MS-EmpiRe outperforms MSqRob in terms of sensitivity and specificity. Reasons for this superior performance might be that inference based on linear regression models like MSqRob rely on assumptions which we know to be rather questionable based on data exploration. Indeed, the residuals are not always normally distributed and since a small mean-variance structure can often be detected even after log-transformation, the homoscedasticity (i.e. equality of variances) assumption is also violated. MS-EmpiRe makes use of empirical fold change distributions to determine significance and is therefore not dependent on these assumptions. In my opinion, MS-EmpiRe

also performs rather well because its variability estimates for the $\log_2$ fold changes are based on many peptides. Since many proteins have very sparse peptide evidence, MS-EmpiRe relies on the assumption that peptides in the same bin behave alike, which is a way of borrowing strength across proteins. These authors also noted that MSqRob seems over-optimistic when scoring proteins with few peptide identifications, which might be an indication that our filtering is not stringent enough or that we need to improve our shrinkage estimation. Furthermore, it was noticed that MSqRob has difficulties in controlling the FDR, but, as the authors used a spike-in dataset without biological variability, this issue might be less severe in practice for reasons already outlined in section 11.1. MS-EmpiRe also has some shortcomings. Indeed, like Diffacto, MS-EmpiRe does not correct for within-sample correlation and it cannot correctly analyze complex designs as it does not allow the inclusion of additional covariates. Moreover, MS-EmpiRe only provides inference for pairwise comparisons.

CSNorm is new approach that uses non-parametric statistics to assess differential protein abundance [25]. In brief, each peptide's raw intensity is rescaled in a range from 0 to 1 and missing values are set to 0. For each protein, these values are summed per treatment over all its peptides in all replicates. The test statistic for differential abundance between two treatments is then the difference of these summed values. p-values are determined based on permutations of the treatment labels over the different peptides. Although CSNorm does not rely on any distributional assumption and therefore does not suffer from any deviations from normality or homoscedasticity, its permutation strategy is incorrect as it breaks the correlation structure of the peptides within a protein. Furthermore, by permuting the peptides without taking the within-run correlation into account, their results will be way too liberal. On top of that, CSNorm's true power will be rather low, especially for proteins with sparse peptide evidence, because it does not use any distributional information. From a practical point of view, methods to increase stability such as borrowing information across proteins are difficult to combine with a non-parametric method. CSNorm is also less favorable in terms of interpretability because its score is a computed value that is no longer linked to a protein's fold change. The authors used a spike-in experiment to compare their method to MSstats, amongst others. MSstats performed poorly on low spike-in concentrations because it cannot deal with missingness in one condition. For higher spike-in concentrations, however, MSstats outperformed CSNorm. This demonstrates that MSqRob will likely outperform CSNorm as well, especially when the hurdle extension is taken into account. Finally, it is remarkable to notice that all methods investigated by the authors have difficulties at controlling the FDR on the spike-in dataset, especially at high spike-in concentrations. This is again most likely due to ionization suppression of the background yeast proteins.

The pepa.test function in the DAPAR Bioconductor R package aims to improve protein quantification by including shared peptides [26]. Here, a single regression model was constructed that includes peptide effects and protein effects whereby peptides can map to multiple proteins. Protein significance is determined via likelihood ratio tests for each protein for the model under the assumption that a protein would be differentially abundant versus the assumption that a protein would not be differentially abundant. The residual variances are estimated with a shrinkage estimator and their likelihood ratio test statistic is proven to be asymptotically chi-squared distributed. To demonstrate superiority, pepa.test was compared to MSqRob and some other approaches by making use of two datasets. The first dataset is a simulation study with varying amounts of shared peptides. The second dataset is a spike-in dataset to which varying amounts of shared peptides were artificially added. The rationale for adding shared peptides is that it has been claimed that up to 50% of all peptides can be shared in complex datasets [34]. It is however not that surprising that DAPAR performs well on

datasets that were engineered to contain many shared peptides, as this is exactly the problem the authors tried to address. However, MSqRob also seems to show a somewhat poorer performance (in terms of precision vs. recall) on the simulated and spike-in datasets to which no shared peptides were added. This might be due to the fact that DAPAR also makes use of a shrinkage estimator for its likelihood ratio test statistic. However, as the pepa.test function can only test protein significance in pairwise comparisons, the authors only used datasets with two experimental conditions. Therefore, MSqRob cannot demonstrate the superior performance of linear regression models by estimating the residual variance based on many runs. Blocking effects or additional hierarchical levels are also not issues in these designs. Finally, the pepa.test function can also not be used to estimate protein fold changes.

BayesENproteomics is a Bayesian method that expands upon the ideas introduced by MSqRob [27]. It includes interactions of peptide effects with treatment effects to account for peptides that behave differently from other peptides mapping to the same protein. This theoretically allows the detection of differential modification statuses or splice variants. To cope with the increased model complexity, BayesENproteomics makes use of elastic net penalization instead of ridge regression. Elastic net is a combination of ridge with LASSO (Least Absolute Shrinkage and Selection Operator) [35]. The LASSO shrinkage penalty is proportional to the L1 norm $\|\boldsymbol{\beta}\|_1$ of the model parameters. The LASSO component allows parameters to be set exactly to 0. Such an automatic model selection will help to avoid too complex model structures. However, a disadvantage of LASSO compared to ridge is that it is less stable: other variables might be selected if the data is slightly perturbed. Elastic net regularization combines the "model selection property" of LASSO with the stability of ridge regression.

BayesENproteomics also assigns weights to peptides, thereby disfavoring not only outlying peptides (a similar goal as the M-estimation in MSqRob), but also peptides with low identification scores. BayesENproteomics imputes missing values with a multiple imputation strategy under the assumption of a combination of missingness completely at random and missingness not at random. In brief, BayesENproteomics constructs a logistic regression model for each protein that contains peptide and treatment effects to determine which missing values are missing completely at random versus not at random. Missing values that are deemed completely at random are imputed from multivariate normal distribution. Missing values that are deemed not at random are imputed from a truncated normal distribution with an upper limit equal to the percentile of the observed values corresponding to the fraction of values that were determined missing not at random by the logistic regression model. This imputation procedure is repeated with each iteration of the Gibbs sampler.

BayesENproteomics also provides a linear regression model for pathway analysis which can be fit after the main analysis:

$$y_{it} = \beta^0 + \beta_t^{\text{treatment}} + \beta_i^{\text{protein}} + \varepsilon_{it} \qquad (\text{Eq. } 11.2)$$

Herein, y_{it} is the $\log_2$ fold change of protein i in treatment t, β^0 the intercept that represents the pathway-level effect, $\beta_t^{\text{treatment}}$ the effect of treatment t, β_i^{protein} the effect of protein i and ε_{it} a random error term. The advantage of this approach over traditional enrichment strategies is that there is no arbitrary cut-off between differentially regulated and background pathways. Moreover, their approach includes information on the magnitudes and the directions of fold changes to give an overall estimate of each pathway's behavior.

BayesENproteomics outcompetes an ordinary least squares model and a robust ridge regression with Huber weights that was re-implemented in MatLab but did not include empirical Bayes variance estimation. For these models, the data were imputed with Perseus imputation, which assumes missingness due to low abundance. However, I would not be too surprised if this method would perform more or less on par or even outcompete the full MSqRob implementation with empirical Bayes variance estimation because of the shrinkage provided by the elastic net and the down-weighing of peptides with low identification FDRs. Moreover, the addition to detect peptides that deviate from their corresponding protein's fold change is definitely an asset. This again demonstrates the enormous potential of Bayesian methods to model extremely complex model structures. However, like all Bayesian methods, BayesENproteomics is computationally intensive and therefore slow. Running the model on a desktop computer took the researchers several hours for a dataset with only a few 100's of proteins. Present-day high-throughput shotgun mass spectrometric experiments typically contain however several thousands of proteins. Moreover, BayesENproteomics was demonstrated on a simple benchmark dataset with only three spike-in conditions. For more complex experimental designs, BayesENproteomics will be even much slower.

All of this demonstrates that MSqRob is rapidly becoming accepted as a valid method for differential protein analysis in the field. MSqRob is being challenged and used as a benchmark by several research groups and several of these new approaches are very promising. I personally see a bright future for Bayesian approaches. Indeed, even though MCMC sampling might currently still be somewhat too slow for routine high-throughput analyses, approaches such as plugging in a maximum a posteriori estimator, Bayes factors and variational Bayes make Bayesian approaches feasible as they are much faster than classic MCMC sampling.

11.6. The impact of technological and algorithmic innovations

The proteomics field is currently evolving extremely rapidly, and technical improvements are constantly pushing the boundaries of proteome coverage. In a future where mass spectrometers might one day be able to identify each and every peptide, one could question the usefulness of developing methods that deal with unbalancedness and missing data for proteome quantification. However, as this moment might still lie quite some years ahead, it is still important to strive for the best-possible analysis pipeline for data that is acquired with present-day technology (1). Moreover, even when these advancements would finally arrive, there would be massive amounts of historical public data containing a wealth of information ready to be queried (2).

Concerning point (1), technical improvements have indeed been massive over the years. Consider for example data-independent acquisition (DIA). Contrary to data-dependent acquisition (DDA), DIA fragments the complete range of the MS spectrum, often in sequential m/z windows. The main advantage of this method is that every peptide ion gets fragmented, so that at least in theory, every single peptide can be quantified, as long as it can be identified. At the start of my PhD in 2013, DIA was, although very promising, still considered a niche application as the quantification depth of DIA was rather limited due to the difficulties in deconvoluting the highly multiplexed MS^2 spectra. Nowadays, DIA is reported to detect more than twice the number of peptides of a comparable DDA run [36]. Modern DIA approaches also showed only 1 – 2% missing values compared to 51% in a DDA shotgun proteomics with a similar quantitation depth [37]. However, at present, it is still not completely clear how the protein discovery FDR should be controlled in DIA analyses. Moreover, DIA still needs one or more prior DDA runs of the same samples on the same instrument to generate spectral

libraries that can be used to identify peptides. Therefore, DIA is generally more suited for e.g. clinical proteomics and less for “hit-and-run” discovery analyses, although the Gevaert lab is currently experimenting with *in silico*-generated spectral libraries generated by the CompOmics lab, which would obviate the need for prior DDA runs. As the latest mass spectrometers can record over 40 spectra per second [38] and increases in machine speed allow DIA acquisition windows to become smaller and smaller [39], I expect DDA and DIA to converge more and more towards each other in the future.

Other new techniques such as differential ion mobility separation are also finding their way into practice. Thanks to differential ion mobility separation, co-eluting peptides can be separated, which drastically reduces the occurrence of chimeric spectra [40].

Moreover, protein identification methods are also becoming more advanced. For instance, the strong increase in analysis depth now allows the sporadic registration of some low-abundant or poorly detectable protein modifications in regular shotgun proteomics workflows. Such sub-stoichiometric modifications would be responsible for at least a third of all unassigned spectra [41]. This fact spurred the development of open modification search engines, which are capable of detecting mass shifts due to modifications that were not *a priori* specified [42, 43]. Dr. Sven Degroeve from the CompOmics lab at VIB-Ghent University recently developed Ionbot, an MS2PIP-based open modification search engine that is also capable of detecting amino acid substitutions [44, 45]. Furthermore, the ever increasing MS resolutions allow a higher MS^2 fragment ion coverage, which slowly paves the way for *de novo* peptide sequencing, the direct identification of peptides based on the MS^2 spectra without the need for a protein database, although these algorithms are still not accurate enough to compete with database searching [46].

However, despite all these exciting technical and algorithmic improvements that are brimming on the horizon, current routine DDA MS workflows still suffer from all the issues described in the introduction and even DIA data are not completely free from missing values. Moreover, in the coming years, technologies such as single cell proteomics are expected to start emerging [47]. Like current single-cell genomic sequencing technologies [48], such datasets can be expected to be plagued by many missing values. Therefore, the work that I have done is not only useful for present-day research but will also lay the foundations for future work on similar data types that will arise from new technologies. Naturally, one day, MSqRob in its current implementation will become obsolete for label-free proteomics. Thus, to remain a competitive method for protein quantification, we will need to keep innovating and improving (see chapter 12).

Concerning point (2), there is indeed a wealth of data available in public repositories, of which the potential is grossly underused. Many biological research questions could indeed be answered by re-analyzing existing publicly available datasets. Sharing data is extremely important for science as it allows others to uncover interesting biology that was not the primary focus of the original research group [49]. In a 2016 editorial in the New England Journal of Medicine, researchers who re-use public data have been termed “research parasites” [50], a derogatory term that has now been reappropriated, e.g. with the introduction of “The Parasites Awards” [51]. MSqRob can have an important contribution in the reanalysis of public data since it is much more sensitive than most of the tools that were used to quantify these datasets initially, yet fast enough to be used for high-throughput analyses. Important in this respect is interoperability between the output of proteomics search engines and MSqRob. Unfortunately, almost every search engine uses its own different file format. The Proteomics Standards Initiative (PSI) of the Human Proteome Organization (HUPO) has established some file format

standards, such as mzTab, which can be used, amongst others, to store peptide-level intensities [7]. However, the mzTab standard has not yet been widely adopted in the field. Therefore, we currently also provide support input from MaxQuant and moFF, as well as some basic support for Progenesis output.

Finally, I would like to stress that MSqRob will not be abandoned when I will leave StatOmics and the Gevaert groups. Indeed, other colleagues are actively working on MSqRob and I also plan to remain involved in the maintenance and development of MSqRob.

12. FUTURE RESEARCH PERSPECTIVES

MSqRob's extremely versatile regression framework provides ample opportunities for methodological improvements. Furthermore, future research should also be directed towards increasing MSqRob's computational performance and making it more broadly available to make sure that MSqRob remains competitive in the field.

As discussed, MSqRob might have too liberal FDR cut-offs (see section 11.4). However, part of the reason for too liberal FDR cut-offs in certain spike-in studies could be caused by the fact that the fold changes of the spiked proteins only go in one direction. Indeed, since the total amount of peptides (in µg) that is loaded onto the mass spectrometer is constant in each run, a higher abundance of a certain (spiked) protein in a first sample compared to a second sample must theoretically be compensated by a lower abundance of one or more (unspiked) proteins in the latter. However, the same holds true when a biological treatment results, for example, in an increase in the abundance of mitochondrial proteins. Such an increase in mitochondrial proteins will indeed imply loading a lower amount of all other proteins onto the mass spectrometer. Therefore, there is a need for better normalization methods that account for ion suppression effects.

Too liberal FDR cut-offs might also be particularly prevalent for datasets where biological variability is rather small or absent (e.g. classical spike-in studies such as CPTAC) because the MS run effects for proteins with sparse peptide evidence are shrunken to zero. This effect might be tempered in experimental datasets wherein true biological variability is included. Due to the much larger variability between biological replicates, it might be that MSqRob is capable of more correctly estimating the within-sample correlations in an experimental dataset with a biological research purpose. For this reason, it would be interesting if a spike-in dataset could be created that mimics this biological variability. This could for example be done by growing multiple human cultures under similar conditions and spiking each of them with yeast proteins and *E. coli* proteins obtained from cell cultures that were independently grown (i.e. creating biological repeats of the experiment proposed by Navarro *et al.* (2016) [52]). Spiking in low amounts of yeast and *E. coli* proteins would ensure that only the most abundant of these proteins would be detected, thus avoiding the rather unrealistic scenario of a whole proteome being differentially abundant. Moreover, spiking in low amounts would avoid a strong increase in sample complexity and prevent large ion suppression effects. By spiking in two different proteomes, it would be possible to mimic a realistic scenario where some proteins are more and others are less differentially abundant.

In terms of speed, MSqRob is computationally more demanding than ordinary least squares models as mixed models need to be fitted with an iterative procedure (the penalized least squares algorithm implemented in the lme4 package in R) [53]. These models then need to be iteratively updated during the M-estimation procedure. Also, as MSqRob is a peptide-based model, it does not provide protein-level summaries for each MS run by default. Such summaries are needed to obtain meaningful information at the protein level and are often used by experimenters, e.g. to create protein heat maps. And, although these summaries can be calculated based on MSqRob's parameter estimates, they are not meaningful if the MS run effects are shrunken to zero. This results in identical protein summaries for each run of the same treatment. Finally, MSqRob's relatively complicated model structure might be relatively difficult to understand for non-specialists.

To avoid issues with degrees of freedom and too liberal FDRs, we are developing a summarization-based approach for MSqRob, MSqRobSum. Indeed, it has been shown that mixed models can also be fitted with a two-stage approach [54]. For the first stage, we propose the following regression model:

$$y_{iptr} = \beta_{ip}^{\text{peptide}} + \beta_{itr}^{\text{run}} + \varepsilon_{iptr} \qquad (\text{Eq. } 12.1)$$

Here, y_{ipr} is the $\log_2$-transformed intensity for peptide p corresponding to protein i in MS run r. $\beta_{ip}^{\text{peptide}}$ and β_{itr}^{run} are the peptide and run effect, respectively and ε_{iptr} is a random error term. t is an indicator for treatment whereby each treatment spans multiple runs. In such a model, β_{itr}^{run} corresponds to the average intensity for protein i in run r after correcting for the peptide effects and can thus be used as a protein-level summary. We then proceed to model the protein-level summaries β_{itr}^{run} as follows:

$$\beta_{itr}^{\text{run}} = \beta_{i}^{0} + \beta_{it}^{\text{treatment}} + \varepsilon_{it} \qquad (\text{Eq. } 12.2)$$

Herein, β_i^0 is the intercept, which corresponds to the average effect of a certain reference treatment, $\beta_{it}^{\text{treatment}}$ is the effect of the t^{th} treatment relative to the reference treatment and ε_{it} is a random error term. Both stages are fitted with ridge regression and M-estimation, although convergence in the second stage already seems to occur after a single iteration. Due to the simpler model structures in both stages, MSqRobSum is much faster than MSqRob. Moreover, the summarization step can be performed automatically "under the hood", allowing the user to concentrate only on the stage 2 model specification.

However, MSqRobSum's performance is inevitably somewhat lower compared to MSqRob because information on the accuracy of each protein-level summary β_{itr}^{run} is lost. Indeed, MSqRobSum can no longer take the peptide-level correlation structure into account. However, the bias that plagues many other summarization-based methods due to the summarization of different types of peptides is avoided because the peptide-level effects are corrected for in the first stage. MSqRobSum thus trades a small drop in performance for a gain in simplicity, modularity and computational speed.

Measures to further improve variance stabilization have a strong potential to improve MSqRob's performance. A logical methodological improvement might be to extend the limma empirical Bayesian variance estimation to the random run effects in the model. Indeed, since MSqRob's mixed model encodes the run effects as random, their variance components can be estimated. It should therefore also be possible to borrow information across proteins with more abundant peptide evidence to improve random effect estimates for proteins with sparse peptide evidence.

In the future, MSqRob might take the step towards becoming fully Bayesian. Such an evolution would be natural because MSqRob's current implementation already contains quite some Bayesian flavors. Indeed, ridge regression is methodologically equivalent with assuming a normal prior, and the limma implementation uses a maximum a posteriori estimator after assuming a chi-squared prior on the inverse of the residual variances. As noted in section 11.5, plugging in a maximum a posteriori estimator, Bayes factors and variational Bayes are computationally fast, and we are currently looking into such approaches in collaboration with Mark Van de Wiel (Amsterdam University Medical Center). The main advantage of the Bayesian framework is its flexibility. Bayesian approaches would for example provide endless possibilities to include prior knowledge from previous experiments about each protein's

behavior. A disadvantage of the Bayesian framework is that inference becomes more complicated. Bayes factors, for instance, require fitting a complex and a simpler model for each contrast as the difference in parameterization between both models must correspond to each contrast.

In recent years, the focus in proteomics has shifted from pairwise comparisons to protein dynamics [55]. The Gevaert Lab also has various time series experiments [56] and many researchers are interested in how the abundances of proteins change over time, for example after a specific treatment. With classical algorithms it is only possible to make inferences about differences between discrete time points or linear changes over time [57, 58]. Such linear relationships can already be modeled with MSqRob but not with MSstats. However, linear relationships between treatment variables and protein abundances are very rare in practice, e.g. because of saturation effects. Moreover, relationships between treatment and protein abundance can often be non-linear in nature. For example, proteins related to the circadian rhythm have been shown to follow a sinusoidal abundance pattern [59]. Some authors have already applied non-linear regression models to proteomics data, for example to determine turnover rates [60], but these methods are limited by their prior assumptions. Additionally, state-of-the-art methods are not suitable for the analysis of longitudinal clustered data because they cannot correctly handle correlation structures.

Given the current version of MSqRob based on the mixed model framework, random effects can easily be added. This allows to exploit the link between penalization, splines and mixed models (see e.g. Wood (2004) [61]) for modeling non-linear trends in abundance patterns. In concrete terms, the method of Storey *et al.* (2005) [62] can easily be generalized from a spline-based analysis of microarray time series to the proteomics context.

Another possibility for expansion is the broadening of MSqRob's applicability. Until now, MSqRob has been promoted as a method for differential protein abundance analysis in label-free shotgun proteomics. However, MSqRob should be applicable on DIA data without any change in algorithm. If DIA features are identified and mapped to proteins prior to quantification, MSqRob can be applied as-is because the DIA data structure is then exactly the same as for label-free DDA data. If the goal is to quantify unidentified features, the analysis becomes simpler because the protein-level hierarchy is omitted. In that case, MSqRob's model structure can be simplified as follows:

$$y_{ftb} = \beta_f^0 + \beta_{ft}^{\text{treatment}} + \beta_{fb}^{\text{block}} + \varepsilon_{ftb} \qquad \text{(Eq. 12.3)}$$

Here, y_{ftb} is the $\log_2$-transformed intensity of feature f, β_f^0 the intercept, $\beta_{ft}^{\text{treatment}}$ the effect of treatment t, $\beta_{fb}^{\text{block}}$ the effect of block b (if present) and ε_{ftb} a random error term. Since DIA data has much fewer missing values, the count component of the hurdle model is likely superfluous.

MSqRob could also be expanded towards labeled data. The only important addition to the model would be an extra term to group differentially labeled peptides in the same run together, irrespective of whether peptides are quantified in the MS (e.g. SILAC labeling) or MS^2 spectrum (isobaric labeling). MSqRob's model would then appear as follows:

$$y_{itg} = \beta_i^0 + \beta_{it}^{\text{treatment}} + \beta_{ig}^{\text{group}} + \varepsilon_{itg} \qquad \text{(Eq. 12.4)}$$

In this model, β_i^0 is the intercept, the average $\log_2$-transformed intensity of protein i for the reference treatment in the reference group. $\beta_{it}^{\text{treatment}}$ is the effect of the tth treatment relative to the reference after correction for group and $\beta_{ig}^{\text{group}}$ the effect of group k relative to the reference after correction for treatment. $\beta_{ig}^{\text{group}}$ models the $g+1$ blocks of two or three (for duplex or triplex labeling, respectively) identical peptides sequences that are differentially labeled within the same run. Note that block effects and run effects are no longer explicitly included in the model because the inclusion of $\beta_{ig}^{\text{group}}$ guarantees the correction for all effects higher up in the hierarchy. Furthermore, an interaction between treatment and group effect can be added to account for differential fold changes for different peptides. It is already possible to encode such a model structure with MSqRob's current implementation, both in an R script as with the MSqRob Shiny App.

MSqRob can also be used for targeted proteomics data (e.g. SRM) or data from protein-protein interaction experiments. Difficulties here are more related to normalization. Indeed, both targeted proteomics experiments and interaction studies will mainly detect proteins that are differentially abundant. However, since most normalization methods assume that the bulk of the protein abundances remains unchanged, finding a correct normalization approach might be difficult for these types of experiments. Such experiments are sometimes normalized against cell counts, total amounts of proteins (determined by e.g. bicinchoninic acid (BCA) assay) or "housekeeping" proteins. However, such approaches are very crude and might sometimes even deteriorate the accuracy of the quantification [63]. Furthermore, testing MSqRob and the hurdle model on future single cell proteomics experiments will also be very exciting due to the many missing values that can be expected in those kinds of experiments [47].

Quantification in metabolomics is very similar to proteomics quantification [64]. It would therefore be straightforward to apply MSqRob to metabolomics data as well. Similar to the analysis of unidentified features in DIA data, metabolomics data has one less hierarchical level compared to label-free DDA data. Indeed, each metabolic peak typically corresponds to a single metabolite (or metabolic degradation product). Therefore, MSqRob's model can be simplified as follows:

$$y_{itb} = \beta_i^0 + \beta_{it}^{\text{treatment}} + \beta_{ib}^{\text{block}} + \varepsilon_{itb} \qquad \text{(Eq. 12.5)}$$

Here, y_{itb} is the $\log_2$-transformed intensity of metabolite i, β_i^0 the intercept, $\beta_{it}^{\text{treatment}}$ the effect of treatment t, $\beta_{ib}^{\text{block}}$ the effect of block b (if present) and ε_{itb} a random error term.

It will also be interesting to implement the ideas set forth by Mallikarjun *et al.* (2018) [27] in MSqRob, many of which we were already considering of implementing in the near future. However, since MSqRob does not make use of MCMC sampling, it will be much faster than BayesENproteomics. The idea of using elastic net regularization can indeed also be implemented in a frequentist context, although it remains to be seen if this will grant a boost in performance to MSqRob that is large enough for the added computational complexity. An elastic net penalty might mainly be interesting for the treatment parameters as elastic net can shrink treatment parameters exactly to zero, thus eliminating them from the model if there is not enough evidence in the data to justify inclusion in the model.

Note that MSqRob currently shrinks parameters towards zero. However, when two model parameters with the same sign are contrasted, the $\log_2$ fold changes are not necessarily being

penalized. Therefore, it would be interesting to explore fusion penalties, whereby the treatment parameters are shrunken towards each other, thus effectively penalizing the $\log_2$ fold changes [65].

It would also be interesting to integrate the uncertainty on the peptide identification into MSqRob's framework, or at least to pass on an uncertainty measurement (such as the peptide identification FDR) from the identification to the quantification step. I already briefly tried this in a very simple setting in the past, by comparing MSqRob without M-estimation to MSqRob with MaxQuant's posterior error probabilities (PEP) as weights, but this did not improve the performance on the CPTAC dataset. However, the heuristic of Mallikarjun *et al.* to combine peptide confidence scores with down-weighing outliers might be worthwhile to pursue in MSqRob as well.

Another obvious extension of MSqRob would be to increase its hierarchical levels. A first way to do this would be to start from the PSM level instead of from the peptide level. However, analysis on the lowest level of summarization does not guarantee a better performance if correlation is not considered [66]. Therefore, the peptide-level hierarchy should not be omitted because PSMs that map onto the same peptide are correlated to each other. Nonetheless, I suspect that the gain of modeling at the PSM level compared to the peptide level will be relatively small because most of the variation seems to manifest at the peptide level.

It is also possible to increase the hierarchy by adding a pathway level to the model. Indeed, in order to better understand the effects of a certain treatment, it is important to derive a functional interpretation from the proteomics data. This is not trivial since each of the thousands of detected proteins is part of a complex network of protein-protein interactions and thus can participate in multiple, related or unrelated pathways [67]. By introducing an extra hierarchy for the pathway level, correlations between proteins can be taken into account. The Kyoto Encyclopedia of Genes and Genomes (KEGG) [68] and Gene Ontology (GO) [69] are commonly used databases that assign proteins to different pathways. The main effect for mean differential abundance over all proteins will allow to prioritize modules with proteins that show differential abundance in the same direction. Using F-tests for protein-treatment interactions will allow to enrich for pathways with proteins that show differential abundance in a different direction, because the interaction term allows the differential abundance to be protein-dependent. Such a hierarchical approach has the advantage that effects that are too weak to be detected at the individual protein level can still be picked up at the pathway level on the basis of the abundance patterns of other proteins in the pathway.

Another option is modeling all proteins together in one big model, with or without an extra pathway-level hierarchy. This would make the inference even more powerful because pathways are also interconnected with each other. Moreover, modeling all proteins together could also enable MSqRob to take shared peptides into account. However, fitting such a big model would be computationally very demanding and therefore strongly decrease MSqRob's speed. Accounting for differential peptide usage is also a very promising road to explore. This would allow the detection of e.g. differential modification statuses or the production of alternative splice variants. For this purpose, some authors have proposed a peptide-centric approach (i.e. modeling each peptide separately) [70]. However, a much more powerful approach is to keep modeling the data at the protein level and adding a treatment-peptide interaction parameter to the model [27]. If the evidence is strong enough, this interaction effect will not be shrunken to zero. This interaction effect can be interpreted as the $\log_2$ fold change of a particular peptide after correction for a possible change in protein abundance and would allow to discern peptides that behave differently from the protein they originate from. This

would be helpful to detect splice variants or changes in a protein's modification status due to the treatment. Note that the hurdle model already provides a useful framework for detecting the complete presence or absence of (modified) peptides in shotgun proteomics datasets, as already discussed in section 11.3.

Together with my colleague Adriaan Sticker, I rewrote MSqRob's code to make it leaner and more performant. We also intend to work towards the publication of MSqRob as a Bioconductor software package. In addition, the inclusion of the tool in open source software suites for the analysis of MS-based proteomics will enable academics and companies to easily use state-of-the-art data analysis methods for differential proteomics in their daily research and development activities. To this aim, MSqRob will also be developed into a Perseus plug-in. Moreover, we will allow the MSqRob Shiny App to output the results of its differential analysis in a tab-delimited output format so that MSqRob's results can seamlessly be imported into Perseus.

Finally, the integration with moFF is also a very important point. MSqRob is already able to input moFF data. With the development of Ionbot by the CompOmics group, it will become possible to create an Ionbot – moFF – MSqRob workflow that handles all the data analysis from searching the raw spectra up to differential quantification.

13. REFERENCES PART III

1. Smith, R. *et al.*, *Proteomics, lipidomics, metabolomics: a mass spectrometry tutorial from a computer scientist's point of view.* BMC Bioinformatics, 2014. **15**(7): p. 1-14.
2. Stein, C. *Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution*. in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. 1956. Berkeley, Calif.: University of California Press.
3. Copas, J.B., *Regression, Prediction and Shrinkage.* Journal of the Royal Statistical Society. Series B (Methodological), 1983. **45**(3): p. 311-354.
4. Ahmed, S.E. and S.M.E. Raheem, *Shrinkage and absolute penalty estimation in linear regression models.* Wiley Interdisciplinary Reviews: Computational Statistics, 2012. **4**(6): p. 541-553.
5. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.
6. Ting, L. *et al.*, *Normalization and Statistical Analysis of Quantitative Proteomics Data Generated by Metabolic Labeling.* Molecular & Cellular Proteomics, 2009. **8**(10): p. 2227-2242.
7. Deutsch, E.W. *et al.*, *Proteomics Standards Initiative: Fifteen Years of Progress and Future Work.* Journal of Proteome Research, 2017. **16**(12): p. 4288-4298.
8. Argentini, A. *et al.*, *moFF: a robust and automated approach to extract peptide ion intensities.* Nature Methods, 2016. **13**(12): p. 964-966.
9. Paulovich, A.G. *et al.*, *Interlaboratory Study Characterizing a Yeast Performance Standard for Benchmarking LC-MS Platform Performance.* Molecular & Cellular Proteomics, 2010. **9**(2): p. 242-254.
10. Henao, R. *et al. Hierarchical factor modeling of proteomics data*. in *Computational Advances in Bio and Medical Sciences (ICCABS), 2012 IEEE 2nd International Conference on*. 2012.
11. Clough, T. *et al.*, *Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs.* BMC Bioinformatics, 2012. **13**(16): p. S6.
12. Guo, W. *et al.*, *Developing a CO2 bicarbonation absorber for promoting microalgal growth rates with an improved photosynthesis pathway.* RSC Advances, 2019. **9**(5): p. 2746-2755.
13. Fornecker, L.-M. *et al.*, *Multi-omics dataset to decipher the complexity of drug resistance in diffuse large B-cell lymphoma.* Scientific Reports, 2019. **9**(1): p. 895.
14. Van Quickelberghe, E. *et al.*, *Identification of Immune-Responsive Gene 1 (IRG1) as a Target of A20.* Journal of Proteome Research, 2018. **17**(6): p. 2182-2191.
15. Kenward, M.G. and J.H. Roger, *Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood.* Biometrics, 1997. **53**(3): p. 983-997.
16. Satterthwaite, F.E., *An approximate distribution of estimates of variance components.* Biometrics, 1946. **2**(6): p. 110-4.
17. Tyanova, S. *et al.*, *The Perseus computational platform for comprehensive analysis of (prote)omics data.* Nature Methods, 2016. **13**: p. 731.
18. Choi, M. *et al.*, *MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments.* Bioinformatics, 2014. **30**(17): p. 2524-2526.
19. Taverner, T. *et al.*, *DanteR: an extensible R-based tool for quantitative analysis of -omics data.* Bioinformatics, 2012. **28**(18): p. 2404-2406.
20. Koopmans, F. *et al.*, *Empirical Bayesian Random Censoring Threshold Model Improves Detection of Differentially Abundant Proteins.* Journal of Proteome Research, 2014.

21. Antonakis, A.N. *et al.*, *ProteoSign: an end-user online differential proteomics statistical analysis platform.* Nucleic Acids Research, 2017. **45**(W1): p. W300-W306.
22. Gierlinski, M. *et al.*, *Proteus: an R package for downstream analysis of MaxQuant output.* bioRxiv, 2018: p. 416511.
23. Zhang, B. *et al.*, *Covariation of Peptide Abundances Accurately Reflects Protein Concentration Differences.* Molecular & Cellular Proteomics, 2017. **16**(5): p. 936-948.
24. Ammar, C. *et al.*, *MS-EmpiRe utilizes peptide-level noise distributions for ultra sensitive detection of differentially abundant proteins.* bioRxiv, 2019: p. 514000.
25. Slama, P. *et al.*, *Robust determination of differential abundance in shotgun proteomics using nonparametric statistics.* Molecular Omics, 2018. **14**(6): p. 424-436.
26. Jacob, L., F. Combes, and T. Burger, *PEPA test: fast and powerful differential analysis from relative quantitative proteomics data using shared peptides.* Biostatistics, 2018.
27. Mallikarjun, V., S.M. Richardson, and J. Swift, *BayesENproteomics: Bayesian elastic nets for quantification of proteoforms in complex samples.* bioRxiv, 2018: p. 295527.
28. Giai Gianetto, Q. *et al.*, *Uses and misuses of the fudge factor in quantitative discovery proteomics.* Proteomics, 2016. **16**(14): p. 1955-1960.
29. Choi, M., *A flexible and versatile framework for statistical design and analysis of quantitative mass spectrometry-based proteomic experiments*. 2016, Purdue University: Open Access Dissertations.
30. Suomi, T. *et al.*, *Using Peptide-Level Proteomics Data for Detecting Differentially Expressed Proteins.* Journal of Proteome Research, 2015. **14**(11): p. 4564-4570.
31. Karpievitch, Y. *et al.*, *A statistical framework for protein quantitation in bottom-up MS-based proteomics.* Bioinformatics, 2009. **25**(16): p. 2028-2034.
32. Karpievitch, Y.V., A.R. Dabney, and R.D. Smith, *Normalization and missing value imputation for label-free LC-MS analysis.* BMC Bioinformatics, 2012. **13 Suppl 16**: p. S5.
33. O'Connell, J.D. *et al.*, *Proteome-Wide Evaluation of Two Common Protein Quantification Methods.* Journal of Proteome Research, 2018. **17**(5): p. 1934-1942.
34. Dost, B. *et al.*, *Accurate Mass Spectrometry Based Protein Quantification via Shared Peptides.* Journal of Computational Biology, 2012. **19**(4): p. 337-348.
35. Zou, H. and T. Hastie, *Regularization and Variable Selection via the Elastic Net.* Journal of the Royal Statistical Society. Series B (Statistical Methodology), 2005. **67**(2): p. 301-320.
36. Kelstrup, C.D. *et al.*, *Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics.* Journal of Proteome Research, 2018. **17**(1): p. 727-738.
37. Bruderer, R. *et al.*, *Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues.* Molecular & Cellular Proteomics, 2015. **14**(5): p. 1400-1410.
38. Shishkova, E., A.S. Hebert, and J.J. Coon, *Now, More Than Ever, Proteomics Needs Better Chromatography.* Cell Systems, 2016. **3**(4): p. 321-324.
39. Hu, A., W.S. Noble, and A. Wolf-Yadlin, *Technical advances in proteomics: new developments in data-independent acquisition.* F1000Research, 2016. **5**: p. F1000 Faculty Rev-419.
40. Winter, D.L., M.R. Wilkins, and W.A. Donald, *Differential Ion Mobility - Mass Spectrometry for Detailed Analysis of the Proteome.* Trends in Biotechnology, 2019. **37**(2): p. 198-213.
41. Chick, J.M. *et al.*, *A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides.* Nature Biotechnology, 2015. **33**(7): p. 743-749.
42. Kong, A.T. *et al.*, *MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics.* Nature methods, 2017. **14**(5): p. 513-520.

43. Devabhaktuni, A. *et al.*, *Measuring proteomes with long strings: A new, unconstrained paradigm in mass spectrum interpretation.* bioRxiv, 2018: p. 282624.
44. Degroeve, S. and L. Martens, *MS2PIP: a tool for MS/MS peak intensity prediction.* Bioinformatics, 2013. **29**(24): p. 3199-3203.
45. Degroeve, S. *Ionbot*. 2018 Acessed on: 30th January 2019. Available from: https://ionbot.cloud/.
46. Renard, B.Y. and T. Muth, *Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification?* Briefings in Bioinformatics, 2017. **19**(5): p. 954-970.
47. Angel, T.E. *et al.*, *Mass spectrometry-based proteomics: existing capabilities and future directions.* Chemical Society reviews, 2012. **41**(10): p. 3912-3928.
48. Van den Berge, K. *et al.*, *Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications.* Genome Biology, 2018. **19**(1): p. 24.
49. Martens, L. and J.A. Vizcaíno, *A Golden Age for Working with Public Proteomics Data.* Trends in Biochemical Sciences, 2017. **42**(5): p. 333-341.
50. Longo, D.L. and J.M. Drazen, *Data Sharing.* New England Journal of Medicine, 2016. **374**(3): p. 276-277.
51. *The Parasite Awards - Celebrating rigorous secondary data analysis*. Greene Laboratory. 2016-2017 Acessed on: 28 January 2019. Available from: http://researchparasite.com/.
52. Navarro, P. *et al.*, *A multicenter study benchmarks software tools for label-free proteome quantification.* Nature biotechnology, 2016. **34**(11): p. 1130-1136.
53. Bates, D. *et al.*, *Fitting Linear Mixed-Effects Models Using lme4.* Journal of Statistical Software; Vol 1, Issue 1 (2015), 2015.
54. Molenberghs, G. and G. Verbeke, *A review on linear mixed models for longitudinal data, possibly subject to dropout.* Statistical Modelling, 2001. **1**(4): p. 235-269.
55. Larance, M. and A.I. Lamond, *Multidimensional proteomics for cell biology.* Nature Reviews Molecular Cell Biology, 2015. **16**(5): p. 269-280.
56. Gawron, D. *et al.*, *Positional proteomics reveals differences in N-terminal proteoform stability.* Molecular Systems Biology, 2016. **12**(2).
57. Ali, A. *et al.*, *Quantitative proteomics and transcriptomics of potato in response to Phytophthora infestans in compatible and incompatible interactions.* BMC Genomics, 2014. **15**(1): p. 1-18.
58. Borirak, O. *et al.*, *Time-series analysis of the transcriptome and proteome of Escherichia coli upon glucose repression.* Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 2015. **1854**(10, Part A): p. 1269-1279.
59. Partch, C.L., C.B. Green, and J.S. Takahashi, *Molecular architecture of the mammalian circadian clock.* Trends in Cell Biology, 2014. **24**(2): p. 90-99.
60. Lau, E. *et al.*, *A large dataset of protein dynamics in the mammalian heart proteome.* Scientific Data, 2016. **3**: p. 160015.
61. Wood, S.N., *Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models.* Journal of the American Statistical Association, 2004. **99**(467): p. 673-686.
62. Storey, J.D. *et al.*, *Significance analysis of time course microarray experiments.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(36): p. 12837-12842.
63. Sandin, M. *et al.*, *Data processing methods and quality control strategies for label-free LC–MS protein quantification.* Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 2014. **1844**(1, Part A): p. 29-41.
64. Fischer, R., P. Bowness, and B.M. Kessler, *Two birds with one stone: doing metabolomics with your proteomics kit.* Proteomics, 2013. **13**(23-24): p. 3371-3386.
65. Witten, D.M., A. Shojaie, and F. Zhang, *The Cluster Elastic Net for High-Dimensional Regression With Unknown Variable Grouping.* Technometrics, 2014. **56**(1): p. 112-122.

66. Milac, T.I., T.W. Randolph, and P. Wang, *Analyzing LC-MS/MS data by spectral count and ion abundance: two case studies.* Statistics and Its Interface, 2012. **5**(1): p. 75-87.
67. Bessarabova, M. *et al.*, *Knowledge-based analysis of proteomics data.* BMC Bioinformatics, 2012. **13**(Suppl 16): p. S13.
68. Kanehisa, M. and S. Goto, *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Research, 2000. **28**(1): p. 27-30.
69. Ashburner, M. *et al.*, *Gene Ontology: tool for the unification of biology.* Nature Genetics, 2000. **25**(1): p. 25-29.
70. Ning, Z. *et al.*, *Peptide-Centric Approaches Provide an Alternative Perspective To Re-Examine Quantitative Proteomic Data.* Analytical Chemistry, 2016. **88**(4): p. 1973-1978.