

Leveraging shotgun proteomics for optimised interpretation of data-independent acquisition data: identification of diagnostic biomarkers for paediatric tuberculosis

By

Ashley Ehlers



Thesis presented in the fulfilment of the requirements for the degree of Master of Science in Medical Sciences in the Faculty of Medicine and Health Sciences at Stellenbosch University.

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Supervisor: Professor David Tabb
Co-supervisor: Professor Hanno Steen

December 2020

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Ashley Ehlers

December 2020

Abstract

Although diagnostic tests for paediatric tuberculosis (TB) are available, no specific test has been tailored to fit the diagnostic challenges children present as well as cater to limited resource settings. The high mortality rates recorded annually are associated with late diagnosis as well as insufficient household contact management (HCM). Further, urine has been identified as an attractive biofluid for urine protein biomarker discovery. Urine is non-invasive, easily attainable in large quantities and is associated with a low cost of collection. Improved data analysis approaches for protein and peptide identification and quantification has paved the way for the development of novel urine protein biomarkers for paediatric TB.

Data-dependent acquisition (DDA) is a powerful approach in discovery of possible urine protein markers. By leveraging the shotgun proteome capabilities of protein and peptide identification using database search algorithms, an optimized data-independent acquisition (DIA) analysis method was developed. In this study, prior to data analysis, the quality of the DDA and DIA approach was evaluated by identifying batch effects and assessing the dissimilarity to allow abnormal runs to be identified and subsequently excluded. It is hypothesized that the quantity of specific host proteins in urine is different for children with TB compared to symptomatic control children who do not have TB. Using an optimised DIA data analysis method leveraging DDA data will allow a statistical identification of differentially abundant proteins in comparative proteomics.

In this study, the *MSstats* R-package for protein-level abundance testing was employed to generate comparisons between two groups, TB cases and controls, for a South African human-immunodeficiency virus (HIV) negative cohort. Three human proteins, leucine-rich alpha-2-glycoprotein (A2GL), aggrecan core protein (PGCA) and cartilage intermediate layer protein 2 (CILP2) were identified as significantly different. The findings of this study support the hypothesis that using an optimised DIA data analysis method leveraging DDA data will identify the differential proteins, potentially leading to validation for use as discovery phase urine protein markers in the clinical setting.

Opsomming

Alhoewel diagnostiese toetse vir pediatriese tuberkulose (TB) beskikbaar is, is geen spesifieke toets aangepas om te pas by die diagnostiese uitdagings wat kinders bied nie asook om te voorsien na beperkte hulpbroninstellings. Die hoë sterftesyfers wat jaarliks aangeteken word, hou verband met laat diagnose sowel as onvoldoende huishoudelike kontakbestuur (HCM). Verder is urine geïdentifiseer as 'n aantreklike biovloeistof vir die ontdekking van proteïen biomerkers. Urine kolleksie is nie-indringend nie, dis maklik bereikbaar in groot hoeveelhede en hou verband met lae versamelingskoste. Verbeterde benaderings vir data-analise vir die identifisering en kwantifisering van proteïene en peptiede, het die weg gebaan vir die ontwikkeling van nuwe urienproteïen-biomerkers vir TB in kinders.

Data-afhanklike verkryging (DDA) is 'n kragtige benadering om moontlike urienproteïenmerkers te ontdek. Deur gebruik te maak van shotgun-proteoom se vermoëns om proteïen- en peptiedidentifikasie met behulp van databasis-soekalgoritmes te maak, is 'n geoptimaliseerde data-onafhanklike verkrygingsontledingsmetode (DIA) ontwikkel. In hierdie studie, voordat data-analise uitgevoer was, is die kwaliteit van die DDA- en DIA-benadering geëvalueer deur bondel-effekte te identifiseer en die verskille te beoordeel sodat abnormale monsters (uitskieters) geïdentifiseer en daarna uitgesluit kan word. Daar word veronderstel dat die hoeveelheid spesifieke proteïene in urine verskil vir kinders met TB in vergelyking met simptomatiese kontrole kinders wat nie TB het nie. Deur gebruik te maak van 'n geoptimaliseerde DIA-data-ontledingsmetode, wat gebruik maak van DDA-data, kan statistiese identifikasie van proteïene wat in verskillende mate in vergelykende proteomika bestaan, identifiseer word.

In hierdie studie is die *MSstats* R-pakket vir proteïenvlak-oorvloedtoetse gebruik om vergelykings tussen twee groepe, TB-gevallen en kontroles, te genereer vir 'n Suid-Afrikaanse mens-immuungebrek virus (MIV) negatiewe groep. Drie menslike proteïene, leucienryke alfa-2-glikoproteïen (A2GL), aggrecan-kernproteïen (PGCA) en kraakbeen-tussenlaagproteïen 2 (CILP2) is geïdentifiseer as beduidend verskillend. Die bevindinge van hierdie studie ondersteun die hipotese dat die gebruik van 'n geoptimaliseerde DIA-data-ontledingsmetode wat gebruik maak van DDA-data, die differensiële proteïene sal identifiseer, wat moontlik kan lei tot validering vir gebruik as ontdekkingsfase-urienproteïenmerkers in die kliniese omgewing.

Acknowledgements

I would like to sincerely say thank you to the following people that continued to be my ultimate supporters throughout this process:

- To my supervisor, Prof. David Tabb. Thank you for pushing me and asking the difficult questions to allow me to grow as a graduate student. Thank you for your soft skills and for always making me feel included in the decision-making process of a software name, a questionnaire, a presentation, and computer repairs just to name a few. Thank you for the network that you have helped me create with scientists in the field of proteomics who I admire. Thank you for your patience throughout my steep learning curve and for never being bothered when I ask the same question for the third time. Thank you for being the best example of what an advisor and mentor should be, I feel privileged to have been under your wing for the past two years. Thank you for believing in me, singing happy birthday to me and sharing your love for Natasha and Mango with me. I will always be grateful. A final thanks to you for coming all the way to South Africa and sharing your love for science with the world!
- To my co-supervisor, Prof. Hanno Steen. Thank you for allowing me to be part of this project. Thank you for always being there to help, advise and keeping me on my toes. Thank you for your patience and for giving me the tools to help me improve this ongoing project.
- To my mentor, Prof. Robert Husson. Thank you for initiating funding for this project and giving me the space to learn. Thank you for being there and reminding me to keep my eye on the ball. I look forward to continuing to work on this project.
- I would like to say thank you to proteomics software developers and mentors who guided me along the way, Marina Kriek (SwaMe), Brendan MacLean (Skyline), Dr. Lindsay Pino (data reproducibility) and Dr. Meena Choi (*MSstats*).
- To the South African Tuberculosis Bioinformatics Initiative team. Thank you for helping me solve my coding questions. Thank you for getting involved in this project and supporting me during my learning process. Thank you for all the laughs and stimulating philosophical discussions, I feel privileged to have been part of this team.

- To my friends, family, and my amazing mother. Thank you for being supportive and loving even when I am 400 km away. You were pivotal in getting me through this degree. Ek is onsettend lief vir julle!
- I would like to say thank you to my funders, the Thrasher Research Fund and the National Research Foundation (NRF) for their contribution.

Research Outputs

Oral Presentation

South African Society for Bioinformatics Online Symposium 2020

Ehlers, A; Tabb, D; Steen, H; Husson, R; Kriek, M. Leveraging shotgun proteomics for optimized interpretation of data-independent acquisition data. Online Zoom Platform, 4-6 August 2020.

List of abbreviations

A2GL	leucine-rich alpha-2-glycoprotein
BCG	Bacillus Calmette-Guerin
BH	Benjamini-Hochberg
CILP2	cartilage intermediate layer protein 2
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
ESRD	End-stage renal disease
FC	Fold change
FDR	False discovery rate
GUI	graphical user interface
HCM	Household contact management
HIV	Human immunodeficiency virus
IDE	Integrated Development Environment
IDFree	Identification free
iRT	indexed retention time
IQR	Inter-quartile range
LC-MS	Liquid chromatography mass spectrometry
LF-LAM	Lateral flow urine lipoarabinomannan
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
<i>M. tuberculosis</i>	<i>Mycobacterium tuberculosis</i>
PC	Principal component
PCA	Principal component analysis
PGCA	aggrecan core protein
PPD	Purified protein derivative
PSM	peptide spectrum match
QC	Quality control
RS	RAWs per spectrum
RT	Retention time
SWATH	Sequential Windowed Acquisition of All Theoretical Fragment Ion Mass Spectra
TB	Tuberculosis
TDA	Target-Decoy approach
TIC	Total Ion Chromatogram
TST	Tuberculin skin test

UniProtKB

UniProt Knowledge Base

WHO

World Health Organization

List of Equations

Equation 3.1	22
---------------------------	----

List of Figures

Chapter 1		
Figure 1.1	<i>Multiplexed versus conventional MS/MS.</i>	3

Chapter 2		
Figure 2.1	<i>Percentage of new and relapse TB cases that were children</i>	10
Figure 2.2	<i>Urinary proteome</i>	12
Figure 2.3	<i>Sample selection will give estimates for the whole population</i>	13

Chapter 3		
Figure 3.1	<i>An Ishikawa diagram (non-exhaustively) shows the major sources of variability in a typical LC-MS experiment</i>	21
Figure 3.2	<i>PCA of quality metrics</i>	23
Figure 3.3	<i>Loadings of the quality metrics under the Kaiser varimax criterion</i>	25
Figure 3.4	<i>Boxplots show the distribution of the medians of the Euclidean distances of experiments to PCA</i>	27
Figure 3.5	<i>PCAs with all samples labelled</i>	28

Chapter 4		
Figure 4.1	<i>Venn diagram shows the relationship between search engines</i>	34
Figure 4.2	<i>Three spectral libraries</i>	37
Figure 4.3	<i>The predicted and observed RT of 23 peptide sequences from the two most abundant human proteins</i>	39
Figure 4.4	<i>The residual plot shows the relationship between the SSRCalc and the measured RT</i>	40
Figure 4.5	<i>The isolation window scheme</i>	41

Chapter 5		
Figure 5.1	<i>Volcano plot illustrates significantly differentially abundant proteins</i>	46

Table of Contents

Declaration	i
Abstract	ii
Opsomming	iii
Acknowledgements	iv
Research Outputs	vi
List of abbreviations	vii
Chapter 1	1
General introduction	1
1.1. Background	1
1.2. Problem statement	3
1.3. Hypotheses	4
1.4. Aims and Objectives	4
Aim 1: To evaluate the quality metrics of DDA and DIA data and identify the outliers.....	4
Aim 2: To analyze the DIA data leveraging from prior knowledge from DDA.	4
Aim 3: To identify potential urine protein biomarkers for paediatric TB diagnosis.	5
1.5. Potential impact of the study	5
1.6. Thesis overview	5
Chapter 2: Literature review	5
Chapter 3: Evaluation of Quality Metrics for Data-Dependent Acquisition (DDA) discovery proteomics data and Data-Independent Acquisition (DIA) data acquired on a Q-Exactive (Thermo) instrument	5
Chapter 4: An Optimized Data-Independent Acquisition (DIA) analysis method leveraging from Shotgun Proteomics Using Open-source software	6
Chapter 5: Urine Proteome for the Identification of Biomarkers for Paediatric Tuberculosis	6
Chapter 6: Conclusion	6
Chapter 7: References	7
Chapter 2	8

Literature review	8
2.1. Abstract	8
2.2. Introduction	8
2.3. Paediatric Tuberculosis	10
2.4. Urine proteomics	12
2.5. Shotgun proteomics	15
2.6. Quality control in DDA	16
2.7. Data-independent acquisition mass spectrometry	17
2.8. Conclusion	18
Chapter 3	20
Evaluation of Quality Metrics for Data-Dependent Acquisition (DDA) discovery proteomics data and Data-Independent Acquisition (DIA) data acquired on a Q-Exactive mass spectrometer (Thermo Scientific)	20
3.1. Materials	20
3.2. Equipment	20
3.3. Experimental Section	20
3.3.1. Quality Metric Generation.....	20
3.3.2. Data Visualization and Explorative analysis	21
3.3.3. Dissimilarity.....	22
3.4. Results and Discussion	22
3.4.1. Comparison of DDA and DIA Performance Metrics	22
3.4.2. Dissimilarity Assessment.....	27
3.5. Conclusion	30
Chapter 4	32
An Optimised Data-Independent Acquisition (DIA) analysis method leveraging Shotgun Proteomes Using Open-source software	32
4.1. Materials	32
4.2. Equipment	33
4.3. Experimental Design	33
4.3.1. Identification of peptides by Sequence Database Searching	33

4.3.2. Creating the target Protein and Peptide List	34
4.3.3. Spectral Library Generation	34
4.3.4. Standard Calibration Peptides Selection	34
4.3.5. Targeted DIA data analysis	35
4.4. Results and Discussion	35
4.4.1. Sequence Database Searching.....	35
4.4.2. Target Peptide List Generation.....	36
4.4.3. Spectral Library Searching	36
4.4.4. Standard Calibrant Peptide Selection Process	40
4.4.5. DIA data analysis	42
4.5. Conclusion	43
Chapter 5	45
Urine Proteome for the Identification of Biomarkers for Paediatric Tuberculosis	45
5.1. Materials.....	45
5.2. Equipment.....	45
5.3. Experimental Procedure	45
5.3.1. Differential protein abundance testing	45
5.4. Results and Discussion	46
5.4.1. Discovery phase of premature urine protein biomarkers.....	46
5.4. Conclusion	48
Chapter 6	49
Conclusion, limitations, and future recommendations	49
Supplementary Material.....	51
Chapter 3.....	51
Chapter 4.....	55
Chapter 5.....	58
Chapter 7	59
References	59

Chapter 1

General introduction

The introductory chapter sets the tone for the rest of the thesis and highlights the need for a biomarker-based diagnostic test against paediatric Tuberculosis. Chapter 1 also states the problem, contains the hypotheses as well as the aims and objectives on how we intend to achieve our goal.

1.1. Background

Tuberculosis (TB), one of the oldest recorded human afflictions, remains the leading cause of death due to infectious disease. New diagnostic strategies are needed to provide early detection and treatment for this epidemic that globally kills two million people per year (MacLean *et al.*, 2020). Approximately 20% of annual TB case notification are children. Historically, child health has not been prioritised, mainly to due to perception that children are rarely infectious and do not develop severe active TB disease (Seddon and Shingadia, 2014). Besides the effects of under-reporting, children are unable to produce sputum on demand, resulting in smear-negative paediatric TB cases. Although it would be ideal to confirm TB cases using culturing, these facilities are often unavailable (Graham *et al.*, 2012).

About a decade ago, researchers began to use biological samples other than sputum such as urine, blood, and exhaled breath for TB diagnosis. Urine became an attractive biofluid for use in children due to availability, accessibility, processing and storage and the low risk to health workers during sample collection (Peter *et al.*, 2010). Excreted urine contains urinary proteins and peptides representing different stages of disease (Thongboonkerd, 2004). Therefore, the mechanisms of disease development and novel therapeutic targets could be discovered by urine proteomic approaches (Kalantari *et al.*, 2015a; Caterino *et al.*, 2018; Duangkumpha *et al.*, 2019). Mass spectrometry (MS)-based proteomics offers a highly parallel multiplexed platform that enables the quantification of large numbers of proteins and peptides (Lin *et al.*, 2018; Ding *et al.*, 2020).

In terms of MS-based proteomics analysis, data-dependent acquisition (DDA) remains the most accepted MS method for untargeted screening in discovery proteomics. With DDA methods, the most abundant ionized species from each precursor ion scan are selected for subsequent isolation, activation, and tandem mass analysis (Courchesne *et al.*, 1998). In

short, shotgun proteomics requires the identification of as many peptides as possible from complex protein mixtures (Michalski, Cox and Mann, 2011). The semi-stochastic nature of precursor ion selection and non-uniformity of scans, however, yield low peptide-level reproducibility, compromising the accuracy of the quantitative results (Kalli *et al.*, 2014). To circumvent the serial nature of DDA and increase the dynamic range, a multiplexed and data-independent acquisition (DIA) approach was developed. Multiplexed-data acquisition is based on more efficient parallel co-selection and co-dissociation of multiple precursor ions, the data from which contain chromatograms not only for individual peptides but also for their fragment ions.

Overall, the goals of DDA and DIA are similar; however, DIA does not select ions based on prior precursor ion scans. In DIA either all ions entering the MS get fragmented at every single point in chromatographic time or the mass-to-charge (m/z) range gets divided into smaller m/z ranges for fragmentation. This approach promises to improve the overall confidence of peptide identifications and relative protein quantification measurements (Chapman, Goodlett and Masselon, 2014). A comparison between conventional MS and multiplexed-data acquisition are shown in Figure 1.1. One such approach developed by AB SCIEX and the Aebersold laboratory is called SWATH-MS (MS^{ALL}), in which multiplexed tandem mass spectra are collected over predefined precursor ion windows termed swaths. The resulting signals are interpreted using prior knowledge from experimental fragmentation spectra (Gillet *et al.*, 2012). DDA and DIA techniques play complementary roles, yielding high-throughput, quantitative consistent, and traceable data that are suitable for proteomic biomarker discovery studies (Muntel *et al.*, 2015; Bruderer *et al.*, 2017).

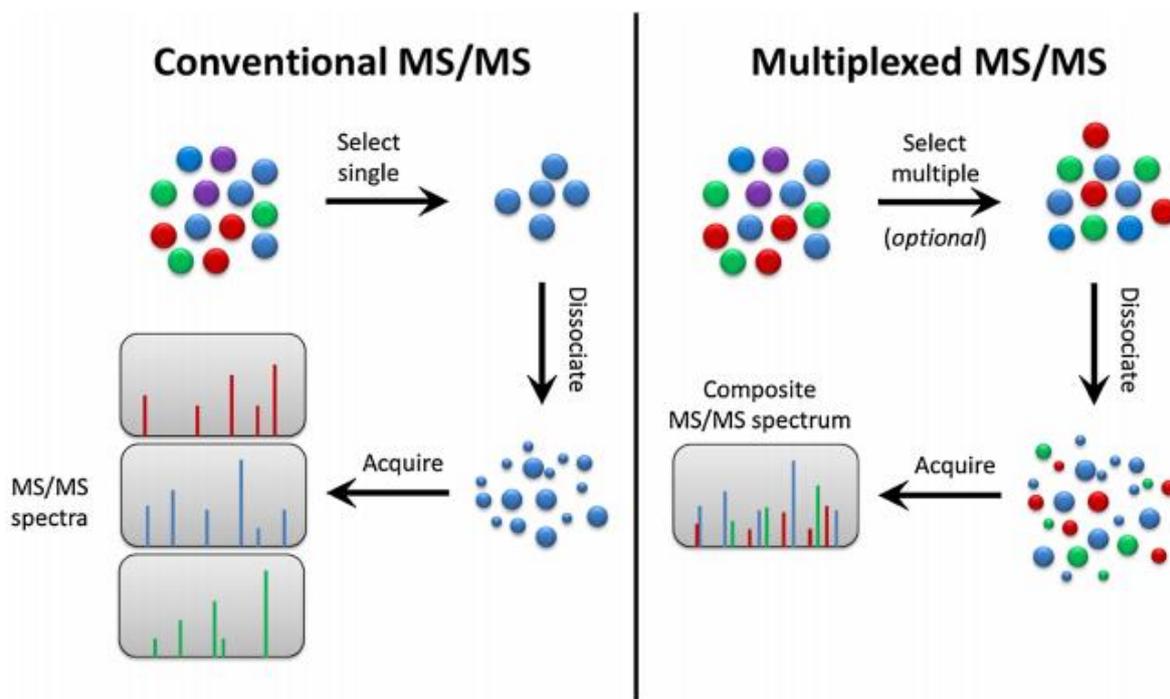


Figure 1.1. Multiplexed versus conventional MS/MS. While fragment selection and acquisition are sequential for the conventional mode, the multiplexed mode allows the acquisition of composite MS/MS spectra from multiple precursors at once. In the latter, the optional selection process can target contiguous or distant m/z ranges (Chapman, Goodlett and Masselon, 2014).

1.2. Problem statement

Children with tuberculosis (TB) have historically been neglected by clinicians, policy makers, academics, and advocates. This has largely been due to the perception that children are rarely infectious, and consequently contribute little to the spread of the epidemic, but also because of the perception that they rarely develop severe disease and because in many countries, children rarely have sputum smear-positive TB. However, children with TB are important. Not only is there a clinical imperative to identify, diagnose, and treat children for a disease that is curable, but by ignoring childhood TB, efforts at epidemic TB control will ultimately fail. This is because children could become the reservoir out of which future cases will develop. The lack of a sensitive and specific test for TB in children that can be performed in resource-limited settings, i.e. at low cost and with little or no laboratory infrastructure, is a major gap in our ability to diagnose and treat children. Excreted urine contains urinary proteins and peptides in different stages of disease. Data-independent acquisition

approaches feature high-throughput, quantitative consistency, and traceable data that are very suitable for urine proteomic biomarker discovery studies.

1.3. Hypotheses

It is hypothesized that the quantity of specific host proteins in urine is different for children with TB compared to symptomatic control children who do not have TB. Using an optimised DIA data analysis method leveraging DDA data will identify the abundantly differentially expressed proteins.

1.4. Aims and Objectives

Aim 1: To evaluate the quality metrics of DDA and DIA data and identify the outliers.

Quality metrics for DDA data will be generated using QuaMeter IDFree (Ma *et al.*, 2012). The SwaMe software (<https://github.com/PaulBrack/Yamato>) will be used to generate sets of quality metrics for the DIA data. The effect of the batches will be visualized using PCA, thereafter individual metrics causing variance will be identified. The outliers will be assessed using a dissimilarity approach.

Aim 2: To analyze the DIA data leveraging from prior knowledge from DDA.

The DIA data will be analyzed using Skyline software (MacLean *et al.*, 2010). The DIA workflow will first be optimized using prior knowledge from DDA data. The DDA data set will be subjected to database searching using the MS-GF+ search engine (Kim and Pevzner, 2014) against an Ensembl database and the MSFragger search engine (Kong *et al.*, 2017) against a UniProt Knowledgebase. The protein assembly will be handled by IDPicker (Ma *et al.*, 2009). To calibrate peptide retention time variation among DIA experiments, calibrant peptides will be selected from the two most abundant proteins commonly found in urine. For this, in-house spectral libraries will be made using the BiblioSpec software (Frewen and MacCoss, 2007) built into Skyline. The method will make use of a selected list of target peptides based on spectral counting. For the targeted DIA data analysis in Skyline, a comprehensive human urine assay library will be used.

Aim 3: To identify potential urine protein biomarkers for paediatric TB diagnosis.

The cohort will be analyzed to identify the effect of age on the incidence of TB. The resulting Skyline document will be converted to *MSstats* (Choi *et al.*, 2014) format using R. Protein-level testing for differential abundance will be performed using the *MSstats* R-package. Comparison tests between cases and controls will be done using *MSstats* to identify significant proteins.

1.5. Potential impact of the study

The goal of this study is to identify urine protein biomarkers for TB in children that could lead to the development of simpler and more accurate tests (such as ELISA) to diagnose TB in children. Improved diagnosis will lead to more appropriate treatment and better outcomes for children with TB. To achieve this goal DIA data were optimised for identification by leveraging prior knowledge from DDA data.

1.6. Thesis overview**Chapter 2: Literature review**

This Chapter represents a literature review which provides an overview of paediatric tuberculosis (TB). Furthermore, it addresses the shortcomings in research on early diagnosis for paediatric TB. The review highlights current knowledge of shotgun discovery proteomics and data-independent acquisition mass spectrometry as an approach to urine protein biomarker discovery.

Chapter 3: Evaluation of Quality Metrics for Data-Dependent Acquisition (DDA) discovery proteomics data and Data-Independent Acquisition (DIA) data acquired on a Q-Exactive (Thermo) instrument

Chapter 3 evaluates the quality metrics of DDA data, and comparative DIA data acquired on the same instrument in the same laboratory. In this study, QuaMeter “ID-Free” software and SwaMe software were employed to generate quality metrics for DDA and DIA human urine proteomics experiments, respectively. The use of these quality metrics identifies sources of

variability through factor analysis. The variability can impact the reproducibility of an experiment and mask the true biological conclusion within a batch effect, thereby undermining the search for paediatric tuberculosis diagnostic markers.

Chapter 4: An Optimized Data-Independent Acquisition (DIA) analysis method leveraging from Shotgun Proteomics Using Open-source software

Chapter 4 describes the use of a data independent acquisition (DIA) workflow on a Q-Exactive mass spectrometer for the detection and quantification of peptides using the Skyline Targeted Proteomics Environment. In this study, a targeted DIA data analysis method was optimized leveraging a comparative data dependent acquisition (DDA) data set. The DDA data set was subjected to database searching using the MS-GF+ and MSFragger search engines, respectively. The protein assembly was handled by the IDPicker algorithm. To allow for stable and accurate prediction of peptide retention times, the data analysis method employed a retention time database generated using standards representing the most abundant peptides commonly found in human urine proteomes. Experiment-based spectral libraries were created to support peptide standards selection, however, the targeted DIA data analysis method relied on a comprehensive human urine spectral library.

Chapter 5: Urine Proteome for the Identification of Biomarkers for Paediatric Tuberculosis

This Chapter evaluates the presence of urine protein biomarkers associated with paediatric TB. Data-independent acquisition (DIA) mass spectrometry was done by the Steen Laboratory as part of The Urine Proteomics Study conducted by Boston Children's Hospital. In contrast to the initial analysis in Steen Lab employing Spectronaut software, for this study the protein-level quantification and testing for differential abundance were performed on a Skyline document using the R-package, *MSstats* based on a linear mixed-effects model. These comparison tests identify the proteins with significantly different means in the data set by estimating fold changes and p values that have been adjusted to control the false discovery rate (FDR) at 0.05.

Chapter 6: Conclusion

In the final Chapter all the results are taken together and the significance of what was discovered is discussed, shortcomings highlighted, and prospects mentioned. Altogether, the results point to the potential of combining DDA and DIA for novel urine protein biomarker discovery, although future studies would be required to validate these findings.

Chapter 7: References

Chapter 2

Literature review

This Chapter represents a literature review which provides an overview of paediatric tuberculosis (TB). Furthermore, it addresses the shortcomings in research on early diagnosis for paediatric TB. The review highlights current knowledge of shotgun discovery proteomics and data-independent acquisition mass spectrometry as an approach to urine protein biomarker discovery.

2.1. Abstract

Tuberculosis (TB) caused by *Mycobacterium tuberculosis* remains a deadly infectious disease for people of all age groups; however, paediatric TB is often neglected due to the difficulty in diagnosis. The diagnostic challenges have led to limited research conducted in this field. Urine collection is non-invasive and easily attainable in large quantities. Therefore, urine is a clinically relevant biofluid for urine protein biomarker discovery for paediatric TB diagnosis. In this review the diagnostic challenges and current diagnostic tests for paediatric TB are highlighted. The urine proteome is then described as an ideal source for protein biomarker discovery. The two most common approaches used in proteomics, data-dependent acquisition (DDA) and data-independent acquisition (DIA), are then introduced. There is a clear need to intensify research efforts in this field, and novel urine protein biomarker that could be discovered DDA and DIA holds promise for paediatric TB diagnosis.

Keywords. data-dependent acquisition; data-independent acquisition; *Mycobacterium tuberculosis*; proteomics; urine protein biomarker

2.2. Introduction

Tuberculosis (TB) caused by *Mycobacterium tuberculosis* remains the most common cause of infection-related death worldwide. In 1993, the World Health Organization (WHO) declared TB to be a global public health emergency (Girardi and Ippolito, 2016). In 2017, about 10 million people became ill with TB and there were 1.6 million deaths caused this disease (WHO, 2017). People of all age groups are affected by TB with varying burden. The highest number of infections occurs in adults, whereas 11% of the total worldwide cases are attributed to children, with a male-to-female ratio close to 1. These annual statistics has

placed paediatric TB lower on the priority list in the end TB strategy at global and national levels (World Health Organization Executive Board, 2015). The need for children specific diagnostics was discussed in a meeting held in 2013 that outline key actions in the roadmap for addressing paediatric TB (World Health Organization (WHO), 2013).

There is a need to strengthen the evidence base that supports the need for early diagnosis in paediatric TB (World Health Organization (WHO), 2013). The lateral flow urine lipoarabinomannan (LF-LAM) assay Alere Determine™ TB LAM Ag detects a constituent of the cell wall of *Mycobacterium tuberculosis* in urine. LF-LAM is recommended by WHO to help detect active tuberculosis in HIV-positive people with severe HIV disease (World Health Organization (WHO), 2019). This biomarker-based assay has poor sensitivity and is deemed unreliable in children due to the lower bacterial burden of TB in children compared to adults (Marais and Pai, 2006). This does not mean that paediatric urine has no applicability in the clinical setting, instead it indicates that a specific diagnostic test for children are needed.

It is known that urine is a valuable source in biomarker discovery studies for diagnostic purposes. Urine contains several thousand proteins and is a less complex sample than plasma which contain more than 10 000 core proteins (Wasinger, Zeng and Yau, 2013). Liquid chromatography (LC) coupled to mass spectrometry (MS) remains the analytical technique of choice as it provides the best sensitivity, selectivity and identification capabilities for biofluids (Spahr *et al.*, 2001). MS enables direct identification of molecules based on the mass-to-charge ratio as well as fragmentation patterns. Thus, it fulfils the role of a qualitative analytical technique with high selectivity (Urban, 2016). The most widely used strategy of tandem LC-MS is known as shotgun or discovery proteomics. For this method, the MS instrument is operated in data-dependent acquisition (DDA) mode, where fragment ion (MS2) spectra for selected precursor ions detectable in a survey (MS1) scan are generated (Mann *et al.*, 2011). The resulting fragment ion spectra are then assigned to their corresponding peptide sequences by sequence database searching (Kapp and Schütz, 2007).

In data-independent acquisition (DIA) mass spectrometry (MS), the instrument deterministically fragments all precursor ions within a predefined mass-to-charge (m/z) range and acquires convoluted product ion spectra, containing the fragment ions of all concurrently fragmented precursors. By rapidly and recursively scanning through consecutive, adjacent precursor ion windows, termed swaths, the full precursor ion m/z range of tryptic peptides is covered and consequently, fragment ion spectra of all precursors within a user defined retention time (RT) versus m/z window are recorded over time. This results in a data set that is continuous in both fragment ion intensity and retention time dimensions and essentially

represents a digital recording of the protein sample analysed (Gillet *et al.*, 2012). The term, “DIA” was originally coined by Venable *et al.* in 2004 to contrast with DDA. Initially data generated by DIA was analysed using similar database search engines as for DDA data, but the multiplexing made it difficult to deconvolve the spectra (Venable *et al.*, 2004). DIA data analysis now rely on spectral library searching (Egertson *et al.*, 2013). Alternative approaches to spectral library searching exists, such as Walnut/PECAN (Searle *et al.*, 2018), Spectronaut (Bernhardt *et al.*, 2012) and DIA-NN (Demichev *et al.*, 2020). With the use of DIA-Umpire (Tsou *et al.*, 2015) it is also possible to generate pseudospectra.

The goal of this review is to introduce the reader to the current two most widely used analytical acquisition modes, DDA and DIA MS. We firstly discuss the diagnostic challenges in paediatric TB and the usefulness of urine as a source for protein biomarker discovery.

2.3. Paediatric Tuberculosis

The risk of children infected with active TB increases with exposure to adults with TB, age, human immunodeficiency virus (HIV) infection and undernourishment (Holmberg, Temesgen and Banerjee, 2019). A world map designed in 2018, shows the percentage of all new and relapse TB cases that occurred in children younger than 15 years of age. East Africa is observed to have the highest proportion of recorder TB cases (>10%) in children (Figure 2.1). Each region has different risk factors associated with the disease. For example, in South Africa the presence of maternal tuberculosis in combination with human immunodeficiency virus (HIV) is associated with a higher number of TB case notifications. In Kenya the problem of undernourishment contributes to a higher number of TB cases. Both these African countries are listed as high burden TB countries (World Health Organization (WHO), 2019).

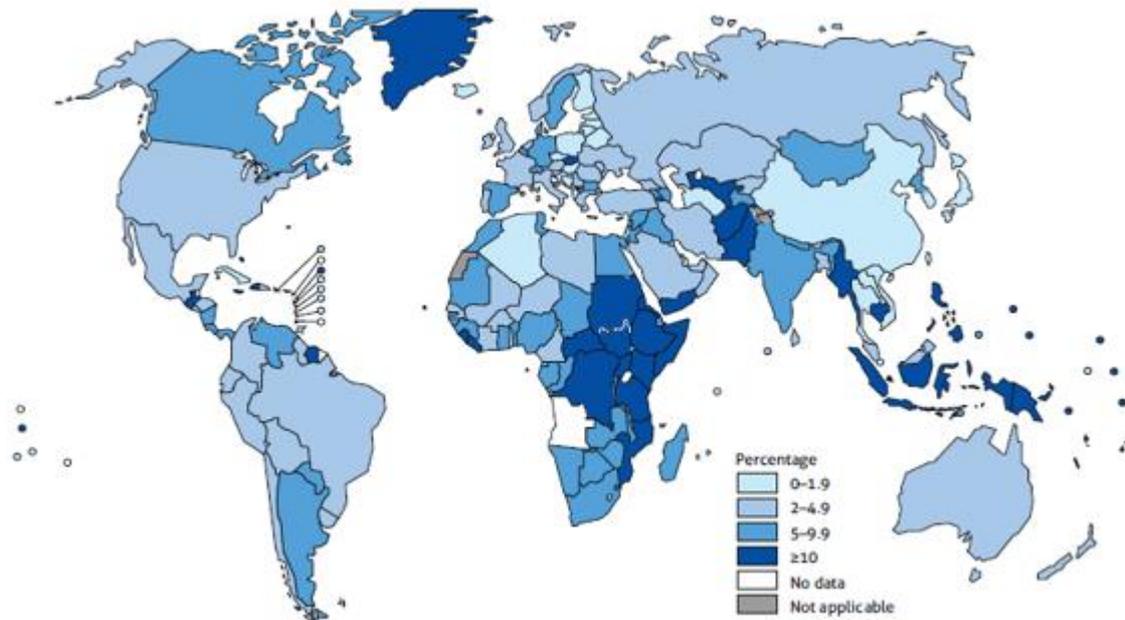


Figure 2.1. Percentage of new and relapse TB cases that were children (aged <15) (World Health Organization (WHO), 2019).

Whereas sputum is the specimen of choice for diagnosis of pulmonary TB in adults by microscopy, culture, or molecular methods, it is difficult to collect adequate respiratory specimens in young children (<7 years of age). This challenge, together with the paucibacillary nature of most pulmonary TB in children leads to a very low sensitivity currently available methods for TB in children (Marais and Pai, 2006). As a result, paediatric TB is often missed or overlooked due to non-specific symptoms (Schaaf *et al.*, 2010). The diagnostic challenge limits the ability to conduct research on TB in children (Newton *et al.*, 2008). Proper disease management will require development of affordable and sensitive diagnostic tests that are not sputum-based.

Currently TB-endemic countries rely on the TB skin test, also called the Mantoux tuberculin test (TST), together with symptoms and where available, chest radiographs, to diagnose TB in children. A TB skin test involves injecting tuberculin purified protein derivative (PPD) into the skin. The reaction is identified as palpable induration (hardness) at the site of injection, however, this response only indicates hypersensitivity and can be positive in persons with asymptomatic (latent) TB infection as well as those with TB disease. Furthermore, the TST can be negative in a TB infected child due to severe malnutrition, HIV infection and immunosuppressive drugs like high dose steroids (South African National Department of Health, 2013).

Besides the diagnostic challenges, paediatric TB faces an under-reporting challenge as well. The incidence of paediatric TB is higher than the number of TB case notifications provided from WHO estimates of tuberculosis prevalence in 2010, especially for children younger than 5 years of age (Dodd *et al.*, 2014). Household contact management (HCM) could substantially reduce childhood disease and death caused by tuberculosis globally. More children can be diagnosed earlier if we implement HCM (Dodd *et al.*, 2018). Results from this multi-cohort collaboration indicate that greater focus should be placed on the first 5 years of life as a period of high risk of progression from tuberculosis infection to disease. Despite the effectiveness of preventive therapy, most cases occurred within weeks of initiation of the contact investigation. Although contact tracing is a high yield means for early case detection, many children are reached too late to prevent disease. Earlier diagnosis of adult cases or community-wide screening approaches in children might be needed to improve prevention of tuberculosis in children (Martinez *et al.*, 2020) together with improved diagnostic testing.

2.4. Urine proteomics

Urine proteomics has become a popular subdiscipline of clinical proteomics because urine is an ideal source for the discovery of non-invasive disease biomarkers (Beasley-Green, 2016). The human kidney (Figure 2.2) is made of functional units called nephrons. The nephrons are divided into two compartments, the glomerulus that filters plasma yielding urine and the renal tubule that reabsorbs the urine. Therefore, urine may contain important information, not just about the kidneys and urinal tract but also about distant organs due to this glomerulus filtration. The analysis of the urinary proteome might therefore allow the identification of biomarkers of diseases, even diseases not related to renal dysfunction. Urine from a healthy individual contains a significant number of peptides and proteins (Decramer *et al.*, 2008). By contrast, blood serum has been the preferred biofluid due to the high abundance of information from blood serum for the discovery of biomarkers. However, the relatively high concentration of the most abundant serum proteins, as well as their wide range of protein concentrations, spanning at least nine orders of magnitude, often limit the study of serum biomarkers (Kentsis *et al.*, 2009). The use of MS methods has shown promising insights into the human physiology as it reflect the changes in a human body (Azarkan *et al.*, 2007; Decramer *et al.*, 2008).

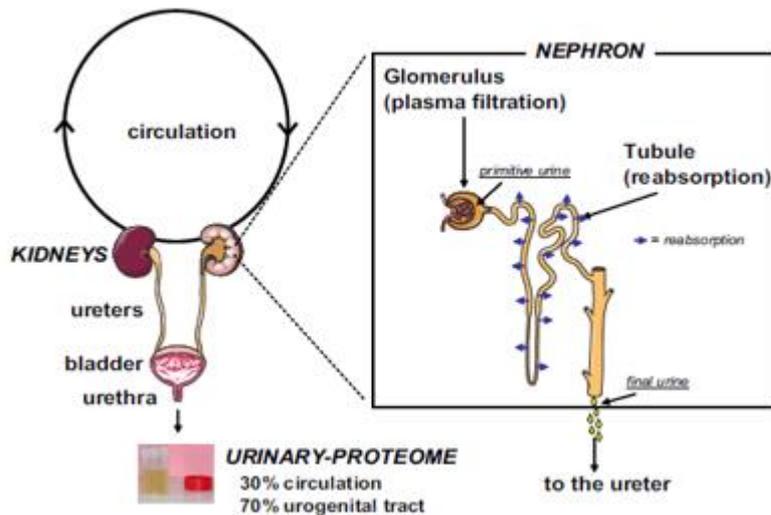


Figure 2.2. Urinary proteome. 70% of the urinary proteins and peptides originate from the kidney and the urinary tract, whereas the remaining 30% originates from the circulation.

Although these promising insights can be positively correlated to disease status, it is noteworthy to take into account the changes in protein and peptide concentrations due to the relationship between daily fluid intake and the intra-individual and inter-individual variability (Schaub *et al.*, 2004). Therefore, highly powered clinical proteomics study could have no significant meaning due to the challenge of reproducibility introduced by physiological changes. Inter- and intra-proteome variability is a major challenge, which is common in most proteomic studies of biological fluids. A later study discussed the standardisation of MS-method based on peptides generally observed within urine (Schiffer, Mischak and Novak, 2006). Over the years various MS technologies have been developed with varying degrees of analytical performance in terms of mass resolution, reproducibility, selectivity, and sensitivity (Beasley-Green, 2016). However, molecular biomarkers have not become practical in clinics yet, and extensive attempts have been devoted to validating these molecular markers (Kalantari *et al.*, 2015b). Once a clinical question has been identified, it is important that proper statistical methods are used to support the conclusion (Good *et al.*, 2007). It is expected that advances in analytical tools and software programs as well as accurate study design soon will improve sensitivity and specificity of available biomarkers (Kalantari *et al.*, 2015b). Many single biomarkers do not hold up during the validation phase. It was proposed that a multimarker panel may achieve high sensitivity and specificity with the same general criteria used for single markers (Fliser *et al.*, 2007; Barratt and Topham, 2007).

To develop urine protein biomarkers as diagnostic tools, clinical usefulness must be verified in a large sample sized study (Grewal *et al.*, 2015; Zak *et al.*, 2017). Many efforts have been

made to characterize more urinary proteins in recent years, but few have focused on the analysis throughput and detection reproducibility. In a study by Lin *et al.* the high abundance blood proteins in the plasma proteome had negative effects on the actual analytical depth. Only after extensive peptide fractionation and 10 hours of MS time could the proteome be mapped in-depth. This give blood a disadvantage in comparison to urine in clinical proteomics. Urine is less complex and it has been suggested to replicate the workflow in this study to extensive urine proteome profiling and clinical relevant biomarker discovery (Lin *et al.*, 2018). An adequate number of samples would allow for a generalisable result.

Often in case-control studies, discoveries start with a selection of cases (with disease) and controls (without disease). Selection bias can occur at biomarker development, sample acquisition and handling, participant selection, the assay process and at the statistical analysis and interpretation of results stage (Zheng, 2018). Therefore, a totally random selection of samples would be ideal, which allows us to assume that the variability observed in the selected samples represents the biological variability in the population and that the sample mean reflects the “intended use population” mean (Figure 2.3). Random selection is the ideal; however, it is often not possible to exclude all biases and confounding factors from samples. Therefore, these factors must be controlled for in sample selection and in data analysis. The clinical aim must be carefully determined. Formulate a “molecular hypothesis” and make sure the analysis method can quantify the essential molecules at the expected level. The primary objective should be defined considering sampling and performance of the prospective analytical and statistical methods (Forshed, 2017).

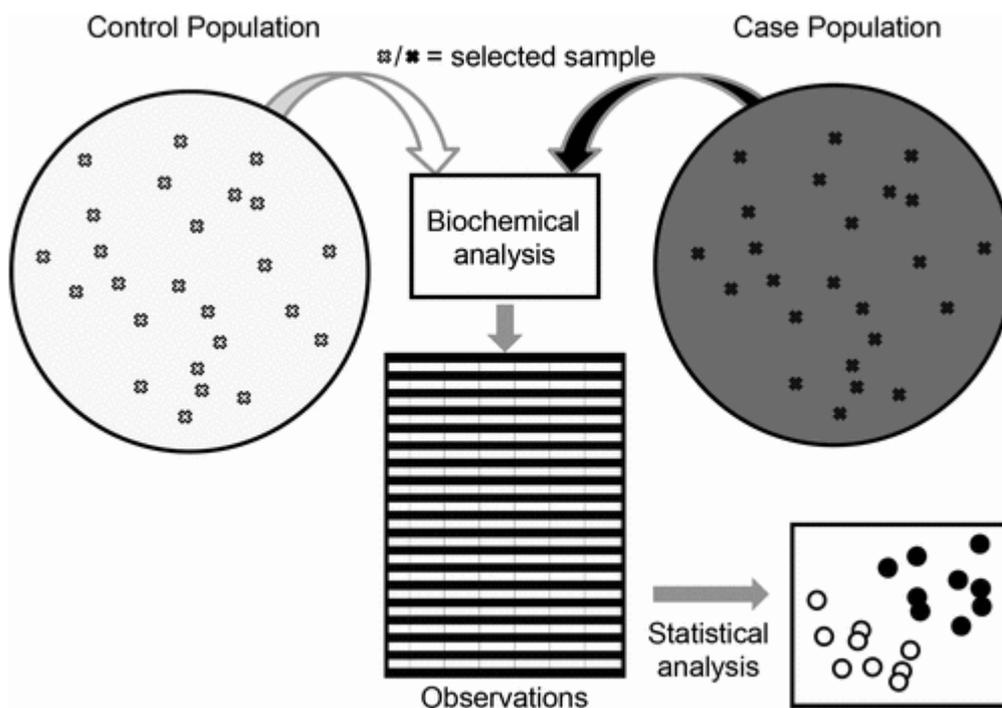


Figure 2.3. Sample selection will give estimates for the whole population.

2.5. Shotgun proteomics

Shotgun proteomics has developed as a robust and sensitive approach to identifying proteins in a complex biological sample. In this data dependent acquisition (DDA) approach, a sample for analysis is prepared by digesting a protein mixture with trypsin to yield a mixture of peptides. The peptides are then loaded on a liquid chromatography column in-line with a mass spectrometer (MS). The identity of thousands of peptides is provided during database searching, which have proven invaluable for automating the characterization of uninterpreted tandem mass spectra and facilitating high-throughput proteomics; however, there remains room for improvement. The database search algorithms make assumptions on where the peptide fragments and also which peptide bonds are most likely to break, allowing fragment intensity predictions. In this way most peptides are correctly identified, however, better prediction of peak intensities could optimise peptide identification. Protein modifications are identified by the algorithms looping over all possible combinations of modified and unmodified residues in a peptide sequence. The efficacy of this looping comes in question when multiple modifications are present in the sample, which often slows the

algorithms down. Search algorithms making use of previously characterized mass spectra could address both limitations (Frewen *et al.*, 2006a).

Nevertheless, sampling of complex proteomes by shotgun proteomics are incomplete and this is observed by assessment of protein and peptides by spectral counting approaches. One such approach is IDPicker, a GUI that stores query information from shotgun proteomes in a cross-platform SQLite file format. DDA remains to pose challenges when attempting to compare proteomes from different biological states based on spectral counting approaches (Li *et al.*, 2010). DDA is a powerful technique for the identification of proteins and peptides in a complex mixture; however, DDA become less effective in detecting all peptide in a mixture. To detect a peptide, the MS/MS data must be queried for the signal specific to that peptide. Because the MS/MS data in a DDA experiment is sampled stochastically, by which a different subset of the available precursor ions is sampled in each subsequent analysis (Domon and Aebersold, 2010), it is impossible to determine whether a peptide with no matching spectra is non-detectable, or detectable but not sampled by MS/MS (Egertson *et al.*, 2015). Due to this reason, a data independent acquisition (DIA) approach was later developed (Venable *et al.*, 2004).

2.6. Quality control in DDA

In any discovery proteomics experiment, as many spectra as possible are generated through tandem mass spectrometry (MS/MS). The generated spectral data is interpreted using various bioinformatics means. Unfortunately, despite the computational advances in the proteomics field, variability within results remains a challenge (Martens, 2013). The variability can stem from multiple sources. To achieve confidence in the obtained results and to ensure reproducible data, it is important to use quality control (QC) measures to monitor and control the existing sources of variability. This is especially important in long term multi-site projects (Bittremieux *et al.*, 2018). QuaMeter (Tabb, 2012) is an open-source tool that computes objective quality metrics for evaluating DDA experiments. The software accepts raw instrument data and identification data as inputs and outputs a tab-delimited file of quality metrics, which are interpreted using statistical methods. QuaMeter allows researchers to track sources of variability in routine practice in real time before critical samples are wasted (Ma *et al.*, 2012). The current QC tools available are limited to DDA discovery experiments. These quality metrics cannot be directly translated to DIA experiments (Bittremieux *et al.*, 2017). Efforts towards expanding QC to workflows such as DIA are being made (Kriek, unpublished), which will further add to the growing MS ecosystem.

2.7. Data-independent acquisition mass spectrometry

Until now, research has focused on the highly abundant urinary proteins and peptides. Analysis of the less abundant and naturally existing urinary proteins and peptides remains a challenge (Beasley-Green, 2016). Less than a decade ago sequential window acquired theoretical MS (SWATH-MS) was introduced to complement DDA approaches such as shotgun discovery proteomics. The SWATH-MS approach is a data independent acquisition (DIA) method performed on a high-resolution mass spectrometer that records a complete recording of all fragment ions of the detected peptide precursors over chromatographic time in a sample. The data analysis depends on a priori assays, derived from fragment ion spectra of the targeted peptides that are best generated in the same high-resolution instrument used for SWATH-MS acquisition (Gillet *et al.*, 2012). Using freely or commercially available software (OpenSWATH (Röst *et al.*, 2014) or Skyline (MacLean *et al.*, 2010) and a DDA-based library, SWATH-MS can be used to carry out protein quantification at performance metrics at a high throughput (Rosenberger *et al.*, 2014). DIA was observed to surpass the protein and peptide identification abilities of DDA. In two studies, DIA experiments doubled the number of proteins and peptides identified compared to DDA where the use of a type-specific spectral library was preferred (Muntel *et al.*, 2015; Bruderer *et al.*, 2017).

Library searching was first proposed as an alternative method to identify MS/MS spectra in 1998 (Yates *et al.*, 1998), but only recently has it been recognized as a means to interpret data-independent acquisition (DIA) data. DIA has emerged as a more reproducible quantitative strategy than data-dependent acquisition (DDA), (Muntel *et al.*, 2015; Li *et al.*, 2019) but it generally depends on acquisition of DDA data to build the spectral library. A common concern with the use of library searching is that peptide identification is limited to only the peptide spectra included in the library, so different approaches have emerged that combine library and database search results or that use more sophisticated library searches. Combining the effort of database searching and library searching in a dual search provides higher reproducibility of peptide identification and quantification without the need to generate new data for library searching (Fernández-Costa, Martínez-Bartolomé, D. McClatchy, *et al.*, 2020). High-throughput sequencing and protein prediction algorithms have provided adequate protein sequences for database searches. Likewise, spectral library searching was observed to improve identification and quantification using. The protein overlap of technical replicates in both DDA and DIA experiments was 30% higher with library-based identifications than with sequence database identifications (Fernández-Costa, Martínez-

Bartolomé, D. B. McClatchy, *et al.*, 2020). Most proteomics studies rely on project specific in-house spectral libraries. However, efforts towards creating a universal atlas database for DIA are being made (Tong *et al.*, 2019).

A recent study combined DDA and DIA in a single LC-MS/MS run and defined it as a data dependent and independent acquisition (DDIA) experiment. Using the retention time calibration curve from DDA data as a classifier for DIA extraction false-discovery rate (FDR) control more proteins could be detected with a smaller number of associated peptide compared to DDA and DIA methods (Guan *et al.*, 2020). The current DIA-MS methods normally cover a wide mass range, with the aim to target and identify as many peptides and proteins as possible and therefore frequently generate MS/MS spectra of high complexity. In a study by Li *et al.* smaller windows shortened the computational analysis time of DIA data while it directly improved quantitation precision. The window size prediction was made using prior knowledge about the biological sample (Li *et al.*, 2019). Collecting prior knowledge has shown to enhance the ability of DIA analysis and could lead to convergence of these methods. Including both the MS1 and MS2 level information has proven to increase the precision of the measurement for technical replicates. This also provides the ability to identify sources of technical variance by statistical modelling in turn improving the power of detecting differentially abundant proteins (Huang *et al.*, 2020).

2.8. Conclusion

To date, limited work has been performed on the diagnosis of paediatric TB. It is difficult to collect sputum on demand from children younger than 7 years of age. As a result, children are often overlooked as a risk group. The risk of children infected with TB increases with exposure to adults with TB. The currently available methods to diagnose TB in children have low sensitivity. Urine is non-invasive and easily attainable in large quantities. Therefore, urine presents as an ideal source for mass spectrometry (MS) - based protein biomarker discovery for paediatric TB diagnosis. MS methods has shown promising insights into the human physiology as it reflects the changes in a human body and could reflect disease status. In many studies single biomarkers do not hold up during the validation phase. It was proposed that a multimarker panel may achieve high sensitivity and specificity with the same general criteria used for single markers. DDA is a powerful and most used analytical method for discovery proteomics; however, the MS/MS data in a DDA experiment is sampled stochastically, by which a different subset of the available precursor ions is sampled in each subsequent analysis, it is impossible to determine whether a peptide with no matching

spectra is non-detectable, or detectable but not sampled by MS/MS. Due to this reason, a DIA approach was later developed. The ability to collect prior knowledge from DDAs ability to identify proteins and peptides has shown to enhance the ability of DIA analysis and could lead to the convergence of these technologies. Including both the MS1 and MS2 level information has proven to increase the precision of the measurement for technical replicates. This also provides the ability to identify sources of technical variances by statistical modelling in turn improving the power of detecting differentially abundant proteins.

Chapter 3

Evaluation of Quality Metrics for Data-Dependent Acquisition (DDA) discovery proteomics data and Data-Independent Acquisition (DIA) data acquired on a Q-Exactive mass spectrometer (Thermo Scientific)

Chapter 3 evaluates the quality metrics of DDA data, and comparative DIA data acquired on the same instrument in the same laboratory. In this study, QuaMeter “ID-Free” software and SwaMe software were employed to generate quality metrics for DDA and DIA human urine proteomics experiments, respectively. The use of these quality metrics identifies sources of variability through factor analysis. Technical variability can impact the reproducibility of an experiment and mask the true biological variability within a batch effect, thereby undermining the search for paediatric tuberculosis diagnostic markers.

3.1. Materials

- MSConvert GUI, part of the ProteoWizard library (<http://proteowizard.sourceforge.net/download.html>);
- QuaMeter “ID-Free” executable, part of the Bumpershoot project (<http://proteowizard.sourceforge.net/download.html>), scroll down and select the platform “Bumpershoot Windows 64-bit tar.bz2”;
- SwaMe Console executable, part of the Yamato framework (<https://github.com/PaulBrack/Yamato>);
- RStudio Desktop (<https://rstudio.com/products/rstudio/download/>);
- R version 4.0.2 (<https://cran.r-project.org/bin/windows/base/>).

3.2. Equipment

- A 64-bit computer with Windows 10 operating system, 8 GB of RAM, a quad-core i5 processor and more than 50 GB of free disk space.

3.3. Experimental Section

3.3.1. Quality Metric Generation

237 DDA and 225 DIA raw instrument files were converted to mzML format (Deutsch, 2010) using the MSConvert GUI, a tool built on the ProteoWizard library (Kessner *et al.*, 2008) with peak picking selected. The parameters used are provided in Table S3.1 of the Supplementary Material. For the DDA dataset, the “IDFree” QuaMeter software (Tabb *et al.*, 2014) was used to produce a list of quality metrics that are independent of identification success rates for MS/MS scans. The typical run time for one Q-Exactive raw DDA files less than 1 min per DDA file, largely consumed by the extraction of ion chromatograms. For the DIA dataset, a QuaMeter-based SWATH metric library called SwaMe software (Kriek, unpublished) was used to produce three different sets of metrics. The comprehensive set of metrics was used for further analysis. The typical run time for one Q-Exactive raw DIA file is less than 30 sec per DIA file.

3.3.2. Data Visualization and Explorative analysis

QuaMeter provides 44 identification-independent metrics to measure the performance in a single DDA experiment, while SwaMe produces three sets of metrics to measure the performance in a single DIA experiment. All metrics generated by QuaMeter IDFree are explained in the user manual (Tabb, 2012) and the metrics generated by SwaMe are defined in Table S3.2 of the Supplementary material. Robust principal component analysis (PCA) was used to visualize and explore the high dimensional metrics from QuaMeter and SwaMe, collectively (Ringnér, 2008) using R version 4.0.2 with the RStudio Integrated Development Environment (IDE). In each set of metrics, a dates column was added in the international standard format YYYY-MM-DD. All the fractions of MS/MS precursor charges and the fastest measured frequencies for MS1 and MS2 collected in any minute (in Hz) in the QuaMeter metrics and the retention time duration and swath size differences in the SwaMe metrics were omitted from PCA analysis due to the low variance contribution of these metrics. A subset consisting of only numeric values were used for PCA analysis.

PCA finds a linear combination of rescaled metric values that maximizes the amount of explained variability among the experiments as principal component one then it finds the linear combination of rescaled metric values that maximizes the amount of remaining variability explained as principal component two. The contribution to variance per component simplifies to nine PC scores, visualized with scree plots shown in Figure S3.1 of the Supplementary Material, where component 1 and 2 (PC1 and PC2) accounted for the most variance. PC1 and PC2 were visualized and samples were grouped by date to assess for batch effect. Factor analysis was used to identify the covariance relationship between the unobservable, latent variables (factors) and the observed quality metrics that explains the

individual sources of variability in the data. The factor analysis is carried out on the factor (correlation) matrix estimated using the robust quality metrics. The loading matrix rotates under the varimax criterion (Kaiser, 1958), where each factor successively accounts for the maximum variance of the squared loadings (squared correlations between variables and factors). This resulted in a high factor of loadings for a smaller number of variables and low factor loadings for the rest, which makes it easier to identify the variables (metrics) contributing to variance within the data.

3.3.3. Dissimilarity

The dissimilarity between a pair of DDA and a pair of DIA, respectively, was measured using the Euclidean distances between the PCA coordinates for a DDA. Mathematically, the dissimilarity between two dimensional coordinates x_1 and x_2 is

$$\sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + \dots (x_{1p} - x_{2p})^2}$$

Equation 3.1

Equation 3.1 was used to calculate the distance table along with a distance matrix was generated. The distance table and distance matrix were rearranged into one vector, and the median Euclidean distance value was calculated per run to indicate dissimilarity, where the larger the dissimilarity value was, the less similar the experimental run was to the other experiments. Boxplots visualized the number of outliers in the DDA and DIA dataset. All values above the top-whisker were labelled as outliers.

3.4. Results and Discussion

3.4.1. Comparison of DDA and DIA Performance Metrics

Despite the recent advances in proteomics technology, the results of the large-scale experiments were still subject to variability. The presence of source of variability placed emphasis on the performance of the instrument used for data acquisition. Variability can stem from multiple sources such as the computational interpretation, the stochastic nature of the different stages within an LC-MS experiment and the presence of contaminants. On the other hand, longitudinal variability arises from instrument drift and sample degradation (Bittremieux *et al.*, 2018). In Figure 3.1. the major sources of variability are shown. This

attention identifying sources of variability called quality control (QC) is an important preventive maintenance to give researchers confidence in their acquired results and to not suffer exaggerated claims.

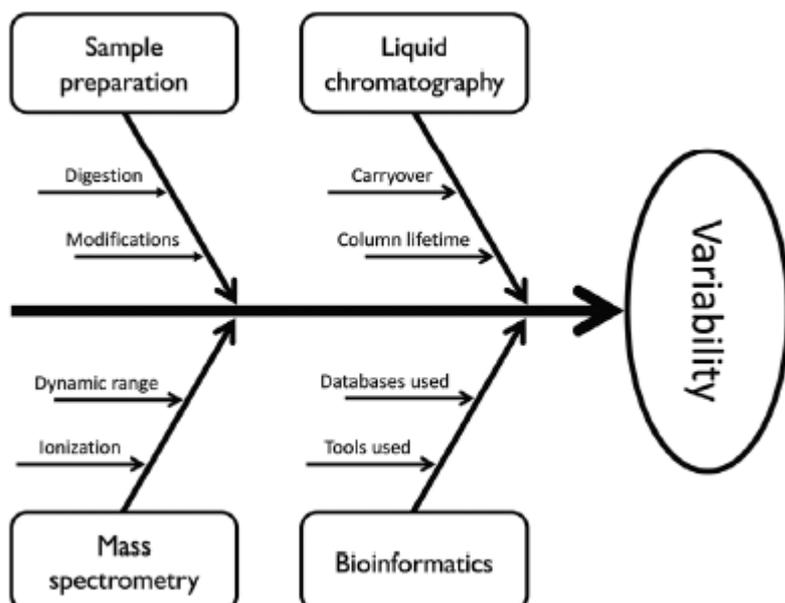


Figure 3.1. An Ishikawa diagram (non-exhaustively) shows the major sources of variability in a typical LC-MS experiment. These sources variability will impact the results and should be considered in a comprehensive DDA or DIA workflow (Bittremieux et al., 2018).

A need for freely available, open source, automated, and easy-to-use QC software tool was clear (Martens, 2013). In 2014, Tabb *et al.* developed an open-source software QuaMeter “IDFree” that can generate a set of quality metrics directly from raw spectral data. This allows for quality metrics to be generated within a few minutes of a DDA run being completed (Bittremieux *et al.*, 2017). Previously, identification free metrics were limited to DDA experiments, however, in 2019 QC efforts expanded using a QuaMeter-based SWATH metric library tool called SwaMe to generate quality metrics (Kriek, unpublished) for DIA workflows. Multivariate statistical methods are important in performance evaluation due to the complex processes of routine experiments. The quality metrics represents an integrated group of measures on these processes (Tabb *et al.*, 2014).

In this study, multivariate statistical methods were used to analyse the quality metrics. Robust multivariate statistical methods can produce insights on the laboratory, sample and experimental variability assessment of large-scale experiments (Tabb *et al.*, 2014). Principal Component Analysis (PCA) is a widely used mathematical algorithm that reduces the multidimensional inputs to a set of components, sorting them by fraction of variance

accounted for by each component (Ringnér, 2008). The first two components of the robust PCA (PC1 and PC2) for the QuaMeter IDFree metrics from DDA data and PC1 and PC2 for the PCA for the SwaMe undivided metrics from DIA data are visualised in Figure 3.2. The PCA plots represent a snapshot for data exploration between these two modes. The DDA experiments are grouped together by date for 13 consecutive days (Figure 3.2.A), whereas the DIA experiments are grouped by date for 14 consecutive days which were subjected to experiment eight months after the DDA experiments were done (Figure 3.2.B). During the eight-month period sources of biological variability such as degradation, modification and freeze thawing could have affected the sample stability over time.

The sample grouping; however, shows that there was no one day that separated from the rest of the experiments based on the quality metrics for DDA experiment and DIA experiments, but rather indicated that runs that appear to separate from the bulk of the data that were not run sequentially i.e. were random but not systematic events. One needs to also take the rest of the principal components into account to conclude which runs are to be considered samples not meeting the quality control criteria. In this case, for the DDA data, the first two principal components account for 54.0% and 12.5% of the variability in the QuaMeter IDFree metrics, respectively. For the DIA dataset, the first two components account for 92.4% and 4.7% of the variability in the SwaMe metrics, respectively. Adding a third component would describe 9.0% additional variability for the QuaMeter metrics and an additional 3.0% for the SwaMe metrics. Secondly, PC1 and PC2 for the SwaMe metrics account for a larger proportion of variability compared to PC1 and PC2 for the QuaMeter metrics, which may be due to SwaMe only producing the 10 comprehensive metrics compared to the 44 metrics produced by QuaMeter IDFree. Therefore, these ratios cannot be directly compared. To understand the underlying process that influences experimental performance, exploratory factor analysis was done (Tabb *et al.*, 2014). The technique evaluates the relationship between unobservable factors and a set of observed quality metric.

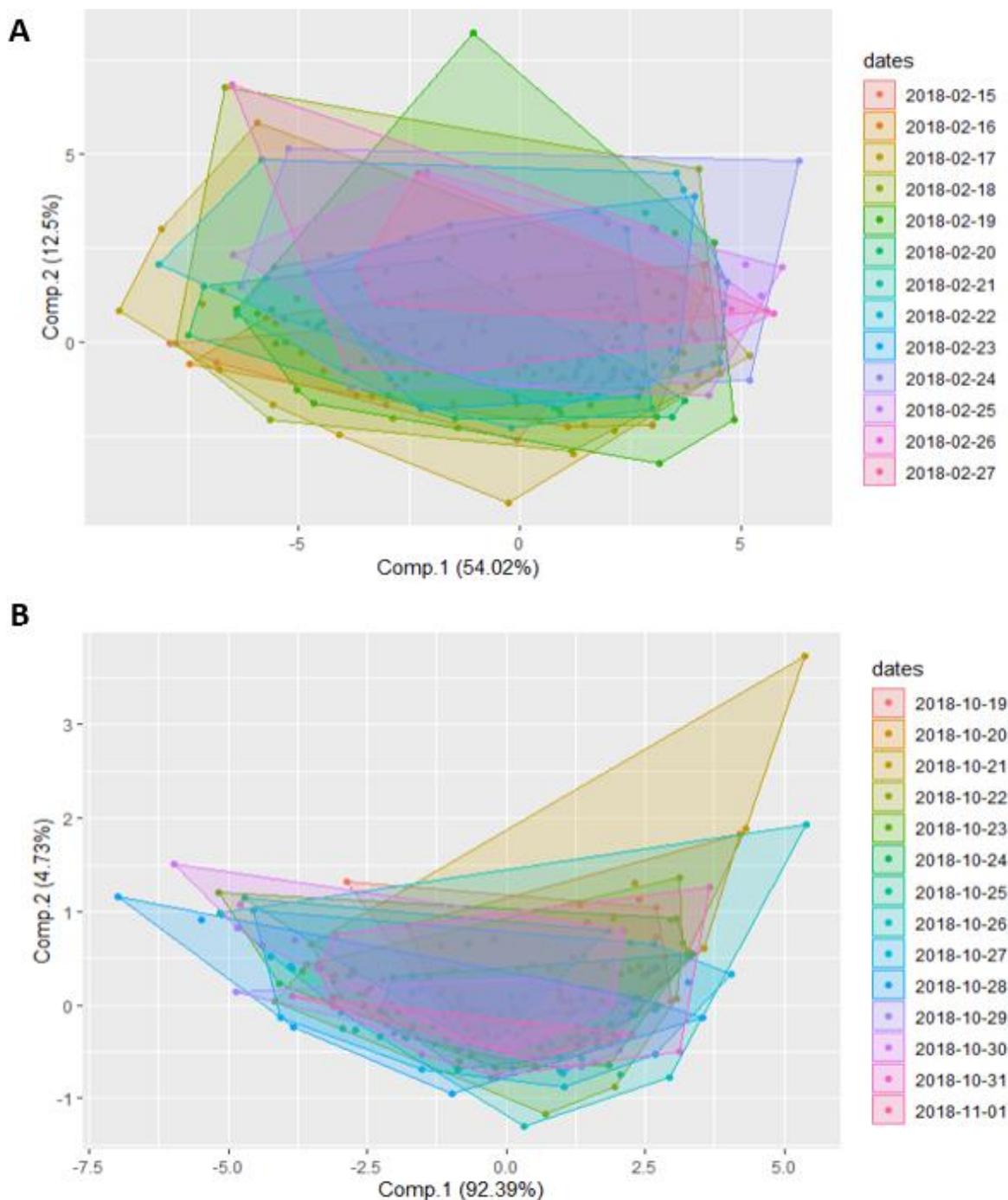


Figure 3.2. PCA of quality metrics. QuaMeter IDFree metrics were generated from DDA samples (A) and PCA of SwaMe undivided metrics were generated from DIA samples (B) collected on a Thermo Q-Exactive mass spectrometer. Each dot represents a sample. Samples were grouped by date denoted by a different colour.

In the DDA experiments in Figure 3.3.A. the loadings accounting for the greatest variability were associated with the number of MS1 scans collected (MS1.Count), the first quartile and the median of the MS1/MS1 scan peak counts (MS1.Density Q1 and Q2), the ratio of the TIC concentration for different quartiles (MS1.TIC.Q2 and Q3) and the change in TIC from

one scan to the next (MS1.TIC.Change.Q2 and Q3). In the DIA experiments in Figure 3.3. B. the loadings accounting for the maximum variability were associated with how many MS/MS scans were collected (MS2.Count) and the IQR for the number of ions detected in all MS2 scans (MS2DensityIQR). The distribution of density and the number of scans appeared important to the variability for both DIA and DDA, although in DDA they were all related to MS1, whereas in DIA they were all related to MS2. One might conclude that only DDA data are more influenced by MS1 signals than DIA, but investigation of the SwaMe metrics reveals that none of them pertain to MS1 TIC values. TIC metrics for SwaMe can be computed using the retention time (RT)-division calculator within the Yamato framework, but the RT-divided metrics were not considered for PCA. There is also no metric for MS1-Density changes in SwaMe, only the MS2-Density changes are observed. There is a metric for MS1 Count, yet in the DDA data it is observed as a source of variability while it is not in the DIA data. The examination of loadings revealed that combinations of key metrics can be used to monitor variability in the data, however, different combinations are observed for DDA data compared to DIA data. To understand what each quality metric signifies requires considerable domain knowledge.

the robust PCA coordinates for each pair of LC-MS/MS experiments. Euclidean distance is one metric to assess dissimilarity by comparison of only two experiments. The Euclidean metric is the most common and intuitive measure of distance. Thus, any abnormal experiment (outlier) will influence only the dissimilarity measures that include that experiment. Distance metrics characteristically yield less accurate estimates than actual measurements, but each metric provides a single model of travel over a given path. Euclidean distance tend to underestimate the distance while a measure such as Manhattan distance tend to overestimate the distance (Shahid *et al.*, 2009). Due to the limitations of distance metrics, unlike actual measurements, it can be directly used in spatial analytical modelling.

Figure 3.4. A and B visualise the distribution of the medians of the distances of experiments to PCA, where samples above the top-whisker were classified as outliers. In this case, the DDA dataset of 237 samples had 18 outliers with a distribution median of 4.4 and an inter-quartile range (IQR) of 1.8 and the DIA dataset of 225 samples had 14 outliers with a distribution median of 2.4 and an IQR of 1.0 (Supplementary Table S3.2) these outliers were highlighted in blue to visualise the points in space on the PCA plot for DDA and DIA experiments (Figure 3.5.A and Figure 3.5.B). The median values indicate that there are differences between groups and the larger IQR in the DDA dataset indicates that the data is more dispersed. The level and spread of the dissimilarity between experiments are consistent in the DDA dataset as well as in the DIA dataset. Due to the differences in the quality metrics generated, the QC of the two experiments cannot be directly compared. These results show that using the same instrument in the same laboratory greatly increase the reproducibility and repeatability of the experiments irrespective of the start time of the experiments. The samples identified as outliers were excluded from the biomarker discovery phase; however, researchers can decide to re-run these samples for inclusion in downstream analysis.

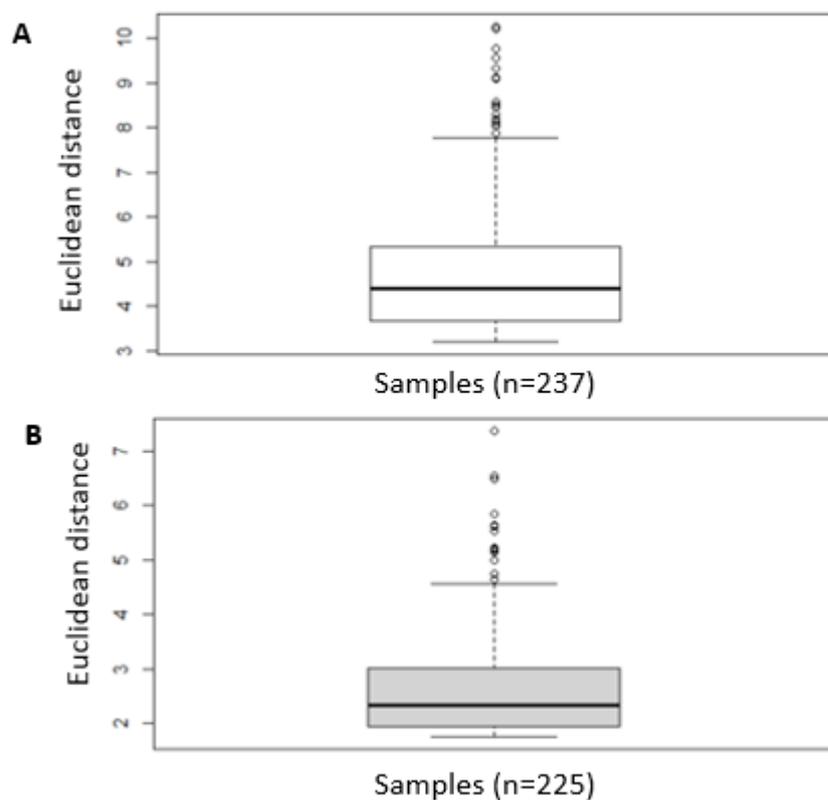


Figure 3.4. Boxplots show the distribution of the medians of the Euclidean distances of experiments to PCA. Outliers are observed in the DDA dataset (A) and in the DIA dataset (B).

Comparative samples were analysed on a Thermo Q-Exactive mass spectrometer in DDA and DIA mode. Quality metrics were successfully produced from the spectrum files in mzML format for all experiments. The QuaMeter “IDFree” metrics for DDA experiments and SwaMe undivided metrics for DIA experiments examines the signals recorded in the experiments that indicates variability. Researchers need to know if batch effects exist under either instrumental paradigm; if batch effects are prominent, the comparability within these experiments may be compromised. This study used Euclidean distances derived from PCA to recognize when a subset of experiments in a large set were distant from the main body of the data due to unusual instrument condition. The batch effects were minimal in the experiments. The individual sample point that appear furthest from the clustered sample points were not sequentially run, suggesting a common cause to their disparity.

The relationship between factors and a set of quality metrics to identify the sources of variability to explain the possible outliers. The distribution of density and the number of scans appeared important to the variability for both DIA and DDA, although in DDA they were all related to MS1, whereas in DIA they were all related to MS2. TIC impacted only the DDA data but not the DIA, this was possibly due to SwaMe comprehensive metrics not containing any TIC metrics. There was also no metric for MS1-Density changes in SwaMe, only the MS2-Density changes were observed. SwaMe has a metric for MS1 Count, yet in the DDA data it is observed as a source of variability while it is not reflected in the DIA data as well.

Traditional outlier analysis was done with dissimilarity assessment. The different median values indicate that there are differences between DDA and DIA samples. The samples identified as outliers in DDA data were not comparable with the samples identified as outliers in the DIA data. This not an unexpected result due to fewer quality metrics computed by SwaMe on DIA. The samples identified as outliers were excluded from the biomarker discovery phase (discussed in Chapter 5). Future recommendations would be to exclude the outliers and then redo the PCA to identify whether the method detects new possible outliers with the same sources of variability. It would also be useful to develop quality control models for dynamic performance monitoring that can allow for systematic incorporation of the insights of laboratory technicians.

Chapter 4

An Optimised Data-Independent Acquisition (DIA) analysis method leveraging Shotgun Proteomes Using Open-source software

Chapter 4 describes the use of a data independent acquisition (DIA) workflow on a Q-Exactive mass spectrometer for the detection and quantification of peptides using the Skyline Targeted Proteomics Environment (freely available on-line at MacCoss Lab Software <https://skyline.ms/project/home/software/Skyline/begin.view>). In this study, a targeted DIA data analysis method was optimized leveraging a comparative data dependent acquisition (DDA) data set. The DDA data set was subjected to database searching using the MSGFPlus and MSFragger search engines, respectively. The protein assembly was handled by the IDPicker algorithm. To allow for stable and accurate prediction of peptide retention times, the data analysis method employed a retention time database generated using standards representing the most abundant peptides commonly found in human urine proteomes. Experiment-based spectral libraries were created to support peptide standards selection, however, the targeted DIA data analysis method relied on a comprehensive human urine spectral library.

4.1. Materials

- Raw mass spectrometry data from 237 DDA proteomics experiments;
- Raw mass spectrometry data from 225 DIA proteomics experiments;
- Philosopher toolkit executable version 3.2.9 (<https://github.com/nesvilab/philosopher/releases/tag/v3.2.9>) scroll down and select “philosopher_v3.2.9_windows_amd64.zip”;
- Java Runtime Environment version 1.8 (required for MSFragger and MS-GF+ jar files) (<https://java.com/en/download/windows-64bit.jsp>);
- MSFragger Software, version 3.0 (<https://bit.ly/2z6dzXa>);
- MS-GF+ Software, release 2020.03.14 (<https://github.com/MSGFPlus/msgfplus/releases>); scroll down and select “MSGFPlus_v20200314.zip”
- Microsoft .NET Framework 4 (<https://www.microsoft.com/en-us/download/details.aspx?id=17851>);
- IDPicker 2.0 executable (<http://proteowizard.sourceforge.net/download.html>);

- Skyline Software version 20.1 (https://skyline.ms/wiki/home/software/Skyline/page.view?name=SkylineInstall_64_20-1);
- RStudio Desktop (<https://rstudio.com/products/rstudio/download/#download>);
- R version 4.0.2 for Windows (<https://cran.r-project.org/bin/windows/base/>).

4.2. Equipment

- A 64-bit computer with Windows 10 operating system, 8 GB of RAM, a quad-core i5 processor, and more than 50 GB of free disk space.

4.3. Experimental Design

4.3.1. Identification of peptides by Sequence Database Searching

In classic shotgun proteomics, the data are recorded using data dependent acquisition (DDA). Traditional database searching identified peptides from 237 DDA raw instrument files. Database searching matches the tandem mass spectra with a database of all possible peptide sequences predicted within a proteome sequence and provides scores for each assignment (Kertesz-Farkas *et al.*, 2012). For the first search, MS-GFPlus algorithm (Kim and Pevzner, 2014) was used to generate spectral vectors, and in this case, search the Ensembl protein database to compute rigorous scores for peptide-spectrum-matches (PSMs) and estimate false discovery rate (FDRs). In a second re-search, the Philosopher tool was employed through the command line to download protein sequences from UniProt Knowledgebase (UniProtKB). This generated a human UniProtKB database with common contaminants and decoys added. The ultrafast search engine, MSFragger (Kong *et al.*, 2017) was used here. The command line executions for the search engines are provided in Table S4.1 of the Supporting Material.

To further boost the performance, IDPicker (Ma *et al.*, 2009) was used to graphically interpret the search results creating one protein assembly file (idpDB) returning identification rates. This tool uses the Target-Decoy Approach (TDA) to impose a peptide spectrum match false discovery rate at a pre-determined threshold of 5%. It is expected that at most 5% of the returned identifications would be false positives (Jeong, Kim and Bandeira, 2012). This inference tool is useful in producing protein and peptide pivot tables which allows the peptide spectra identified to be statistically compared by spectral counts across different laboratories or methods.

4.3.2. Creating the target Protein and Peptide List

For the next step, the R programming language was used to read the protein table from the ultrafast search with MSFragger. For the proteins identified, a count of how many spectrum files matched a protein at least once was computed as a score defined as the RS-score using a for loop. The *tidyverse* R library provided functions to sort proteins by their RS-score in descending order, breaking ties by sorting by descending numbers of filtered spectra. The top protein identifiers based on RS-score greater than N divided by two from this list were kept. Thereafter the non-redundant peptide table from IDPicker was read into R, a second for loop was used to compute the RS-score for each peptide. The peptides were sorted by RS-score in descending order, breaking ties by sorting by descending numbers of filtered spectra. For each protein identifier in the list, a set of five peptide sequences matching each protein identifier were pulled out. The final list of the top protein identifiers with their top five matching peptides sequences based on a high RS-score were used as the targets in the targeted DIA analysis method.

4.3.3. Spectral Library Generation

BiblioSpec software package (Frewen *et al.*, 2006b) is a built-in tool in the Skyline software (MacLean *et al.*, 2010) to create and search experiment-based tandem MS spectral libraries. The tool was used to create two custom spectral libraries using the 237 DDA raw files as input along with the MS-GF+ search outputs and MSFragger search outputs, respectively. Spectral library information was provided from the Steen laboratory who generated a comprehensive human urine protein library using the closed-source software, Spectronaut (Bernhardt *et al.*, 2012). From this pre-assembled library, an assay library in the BiblioSpec format was generated.

4.3.4. Standard Calibration Peptides Selection

Using the tables produced by IDPicker, the two most abundant human urinary proteins were selected. For each protein, twelve peptides sequences were pulled out based on their RS-score. These peptide sequences were imported in Skyline along with the spectral library generated using the 237 DDA raw instrument files and the outputs from the MSFragger search. An initial integration was done using all DIA raw instrument files to extract chromatograms for the 24 peptide sequences across all files. Skyline predicted retention times based on the sequence-specific retention calculator (SSRCalc) as well as observed retention times for the 24 peptide sequences across all DIA raw files. To select only twelve final peptides, a criterion for inclusion was generated. Predicted retention times had to be

spread across a wide RT range while the observed retention times should not surpass an one-minute variability interval across all DIA raw files (Escher *et al.*, 2012). Appropriate charts were made using *ggplot2* and *ggpubr* in R to visualize the predicted and observed retention times. A second selection criterion was that a selected peptide must have intensity values across all DIA files. For this, the R-package MSstats was used to generate intensities for the peptide sequence across all DIA files. The final list of twelve peptides were used to create an iRT database file.

4.3.5. Targeted DIA data analysis

The methods and results described in the preceding sections were all used to evaluate the tools to be used in the analysis of DIA data leveraging from DDA data. The final method used a target peptide list that was created using the ultrafast MSFragger database search engine with the commonly used UniProtKB. Subsequent protein inference was done with the IDPicker algorithm. The comprehensive human urine spectral library was used to inform DIA chromatogram integration. To ensure accurate and stable RT prediction in the DIA analysis, twelve calibrant peptide standards were generated. The Skyline settings are provided in Table S4.3 of the Supporting Information. All DIA raw data file was imported into Skyline to extract chromatograms for quantification.

4.4. Results and Discussion

4.4.1. Sequence Database Searching

The search with MS-GF+ yielded 52 186 peptide spectra while the search with MSFragger yielded 44 770 peptide spectra with 39 289 (68.1%) peptides shared in the shotgun experiment (Figure 4.1). These search results emphasize that when different combinations of tools are employed the capability of data analysis vary widely due to the different ways in how these algorithms score the peptide-spectrum matches (PSMs) which could inherently be a ranking problem (Frank, 2009). Using the MS-GF+ search engine, IDPicker could infer more peptides compared to the MSFragger search strategy. MS-GF+ uses a probabilistic model and computes rigorous E-values (score) giving the tool the ability to be used on diverse types of spectra, instruments and methods (Kim, Gupta and Pevzner, 2008, Kim and Pevzner, 2014). On the other hand, MSFragger has proven to be particularly useful for large-scale unbiased detection of post translational modifications (Kong *et al.*, 2017).

The fragment-ion indexing allowed MSFragger to identify the PSMs in significantly less time compared to MS-GF+. However, the scoring system used by MS-GF+ could be giving the tool the identification advantage over the X!Tandem scoring system (Bjornson *et al.*, 2008)

used by MSFragger. For the MSGF+ search, the Ensembl database was used; however, UniProtKB provides annotations with a minimal level of redundancy through human input or integration with other databases making it the predominantly used database for researchers (Xu and Xu, 2004, Chen, Huang and Wu, 2017). The UniProtKB was used to match the MSFragger search results, which provides the accessions most used between laboratories. The UniProt database matches the database employed by the Steen laboratory. For comparable accessions, these search results were used for further analysis.

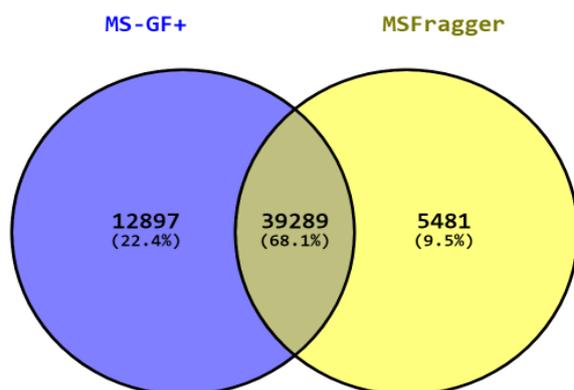


Figure 4.1. Venn diagram shows the relationship between search engines. The MS-GF+ search engine identified more peptides than the MSFragger search.

4.4.2. Target Peptide List Generation

The DIA data analysis method used zooms in on the protein and peptide targets selected post data collection. The DIA data analysis method used here rely on the ability of shotgun proteomics (DDA-based) to detect all component peptides for targeted data analysis. From the protein table of 2245 proteins, a set of the top protein identifiers was selected. The number of samples containing spectra identified to each protein was computed. The *tidyverse* programming paradigm for crafting code in R was installed to sort the protein list by RS-score in descending order, breaking ties by sorting by descending numbers of overall spectra. A total of 590 proteins were selected. From the non-redundant peptide list, a set of 5 peptides were pulled out that match to the 590 proteins resulting in 2950 peptides. Proteins with less than 3 peptides were removed, resulting in 514 proteins and 2662 peptides in the target list. There is an intrinsic limitation caused by the stochastic nature of shotgun proteomics (Faria *et al.*, 2017), therefore employing a refined data analysis approach could increase the sensitivity of the search for diagnostic markers.

4.4.3. Spectral Library Searching

Traditional database searching algorithms are built on the assumption that each MS/MS represents the fragments of an individual peptide rather than containing the fragments of many peptides that were co-fragmented, therefore it cannot deconvolve the multiplexed peptide spectra from DIA (Noble and MacCoss, 2012). Alternative approaches such as DIA-Umpire or PECAN exist to query MS data acquired in DIA mode without the need of a spectral library, but studies have been proven them to be less sensitive compared to using a spectral library (Navarro *et al.*, 2017, Ting *et al.*, 2017, Tsou *et al.*, 2015). Thus, DIA data analysis relies on spectral libraries instead to collect previously identified spectra, then create a reference library to identify spectra. The query spectra are compared to references in the library to find the ones that are most similar. A dot product metric is used to measure the degree of similarity, which distinguishes correct from incorrect identifications.

In this study, the BiblioSpec software package (Frewen *et al.*, 2006b) was used to create two experiment-based spectral libraries using the 237 DDA raw instrument files and the MS-GF+ search results and the MSFragger search results as inputs, respectively. Compared to the results from IDPicker, where the MS-GF+ database search identified more peptides than MSFragger database search, the spectral library creation search detected 35 981 unique peptides through the MS-GF+ search and 40 404 unique peptides through the MSFragger search. Although MSFragger identified more unique peptides, the peptides came from a smaller set of peptides overall. A study by Frewen *et al.* (2006) confirms this finding by showing that spectral library searching increases the number of peptides that can be identified compared to database searching. The study also points out that a large spectrum library has a very small reduction in sensitivity over a small spectrum library. It is known that with database searching the use of a different size FASTA files can affect the false discovery rate results. However it is currently unknown whether the size of a spectral library can result in a variance of the true positive rate in SWATH extraction (Zi *et al.*, 2014). The FDR is an estimate from the best peak group per peptide query within one run, independent from the other runs acquire during the same experiment. It has been demonstrated that the FDR should thus be controlled at the peptide query-, peptide- and protein-level in peptide-centric scoring workflows applying comprehensive spectral libraries (Rosenberger *et al.*, 2018).

One set of MS/MS spectra for the peptide, AAAATGTIFTFR++ is shown in Figure 4.2 to visualise the intensity differences of the fragment ions between libraries. The quality and coverage of the spectral library is critical for peptide identifications, however the availability of open spectral libraries that can be used to process SWATH data is limited, therefore most users create their project specific libraries in-house. Spectral information was provided by the Steen laboratory who created a comprehensive human urine spectral library using closed-source tool, Spectronaut. This library was formatted into an open-source BiblioSpec

format by Brendan MacLean (developer of Skyline MS). The assay library contained 56 697 unique peptides detected with higher precursor ion intensities compared to the experiment-based libraries. In Figure 4.2.A we can observe how the library does not measure the intensities for the b-ions in the lower m/z range, however higher intensities are recorded for the y-ions that are detected. This is because the Spectronaut KIT spectrum is normalised so that the biggest peak is automatically at 100 on the intensity scale.

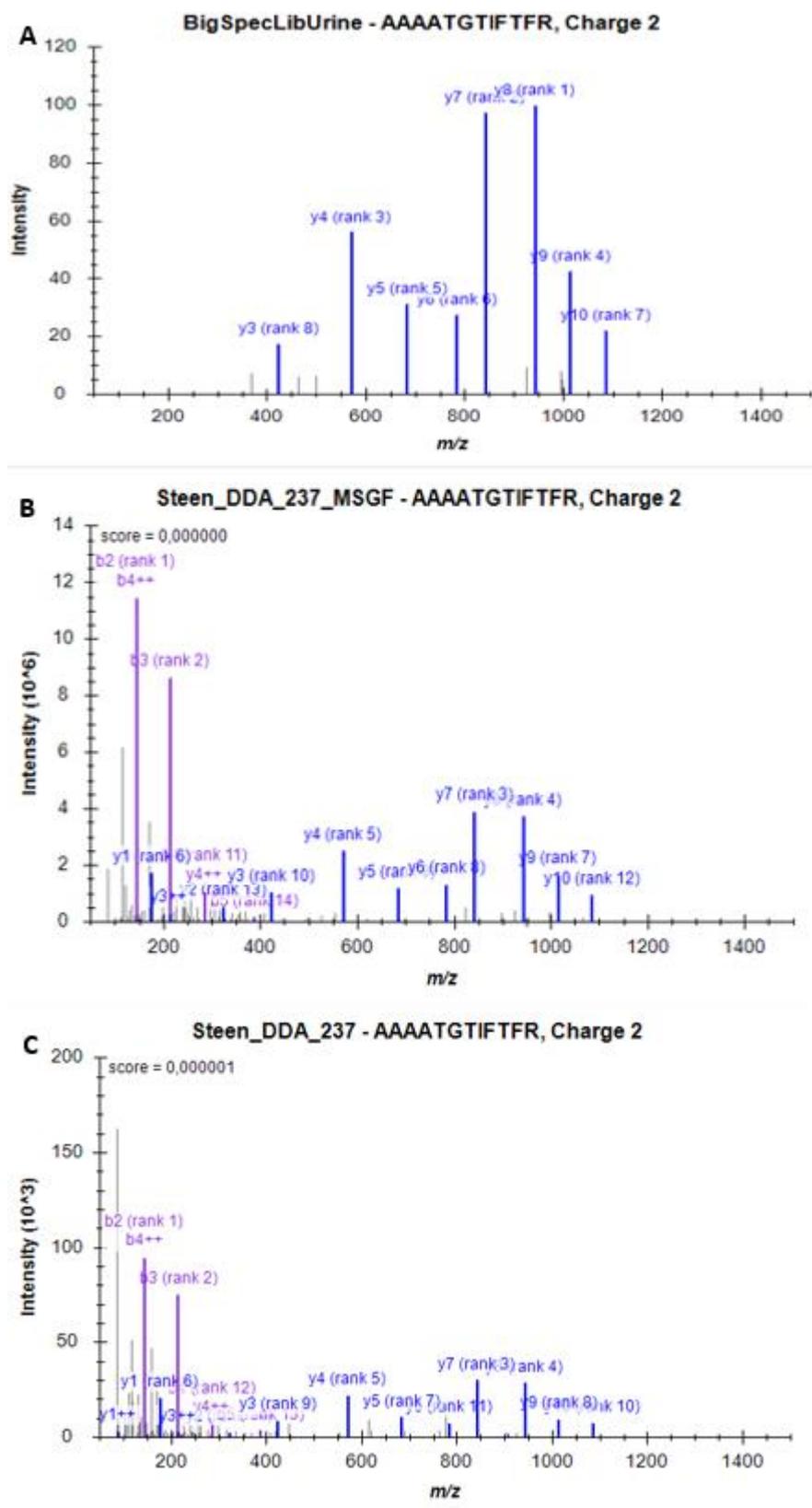


Figure 4.2. Three spectral libraries. The pepXML files from the MSFragger search along with the DDA instrument files (A) and the mzIdentML files from the MS-GF+ search along with the DDA instrument files (B) was used to generate a spectral library. A large

Spectronaut library generated using pooled urine samples was formatted to a BiblioSpec format (C).

4.4.4. Standard Calibrant Peptide Selection Process

Targeted analysis of data-independent acquisition (DIA) data is a powerful mass spectrometric approach for comprehensive, reproducible, and precise proteome quantification. It requires a spectral library, which contains all considered peptide precursor ions with empirically determined fragment ion intensities and their predicted retention time (RT). RTs, however, are not comparable on an absolute scale, especially if heterogeneous measurements are combined, therefore index-based retention times (iRTs) have been employed (Bruderer *et al.*, 2016). iRT is a technique for storing calibrated, empirically measured peptide retention times in a library for future use in retention time prediction for scheduled acquisition and peak identity validation. The iRT concept has been implemented in Skyline using the sequence-specific retention calculator (SSRCalc) with 100-Å pore size C₁₈ columns (Krokhin, 2006). In this study, commercially available synthetic standard iRT peptides were not spiked into samples during preparation. Thus, standard calibrant peptides were selected post data collection.

A method based on lookups of previous registered retention times were employed for the DIA analysis method. From the IDPicker protein pivot table, the two most abundant proteins, human serum albumin and uromodulin were selected based on their RS-score. A total of 24 peptides, 12 per protein were pulled out from the peptide pivot table based on a high RS-score using R programming language. The sum of 24 peptides were imported into Skyline as target peptides while only 23 peptides match to the spectral library. All DIA raw file were integrated into Skyline and the peptide retention results were exported. Using the R-package *MSstats*, the intensities for each fragment ion for each peptide were extracted. This information provided a criterion to select the 12 standard calibration peptides from. The criteria require a wide retention time range with minimal retention time variability and consistent intensities across all DIA raw files (Escher *et al.*, 2012).

Each peptide sequence was assigned to a number label shown in Table S4.2 of the Supporting Material to make visualisation easier. The dot chart shows the wide range of predicted retention times for the 23 peptides. Peptide 15 from uromodulin and peptide 8 from albumin represents the start and end point for RT prediction (Figure 4.3.A). The boxplots show the variability of the observed RTs across all DIA files. Peptides 13 and 14 are certainly the ones we want to include in the database due to low variability seen in their observed RTs whereas peptide 1 has a four-minute variability range and was excluded from

further consideration (Figure 4.3.B). The R-package, *MSstats* (Choi *et al.*, 2014) was also used to obtain a list of intensities for all peptides. Intensities were ranked and peptides with zero intensities were excluded. Peptides that were eventually included in the iRT database are 15, 13, 5, 14, 22, 3, 19, 18, 9, 4, 6 and 8. These peptides have a high degree of correlation ($R=0.9889$) between the predicted and observed retention time across all DIA raw files (Figure 4.4). Having standards is useful for the characterisation and identification of peptides (Veenaaas, Linusson and Haglund, 2018). The retention time of the last chromatographic peak indicated that the chromatographic run was less than 50 minutes long.

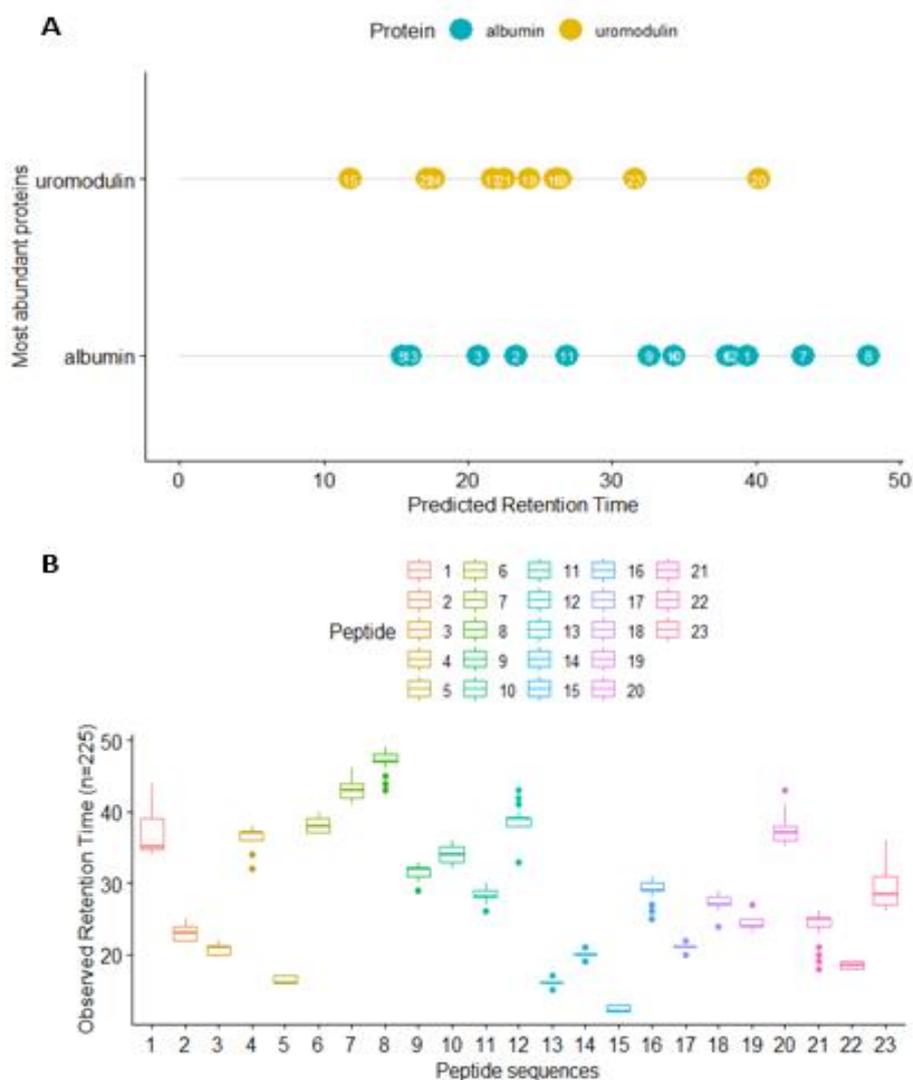


Figure 4.3. The predicted and observed RT of 23 peptide sequences from the two most abundant human proteins. The dot chart shows the wide range of RTs predicted by Skyline for the 23 peptides (A). The boxplot shows the variability of the observed RTs by Skyline for the 23 peptides across all 225 DIA RAWs (B).

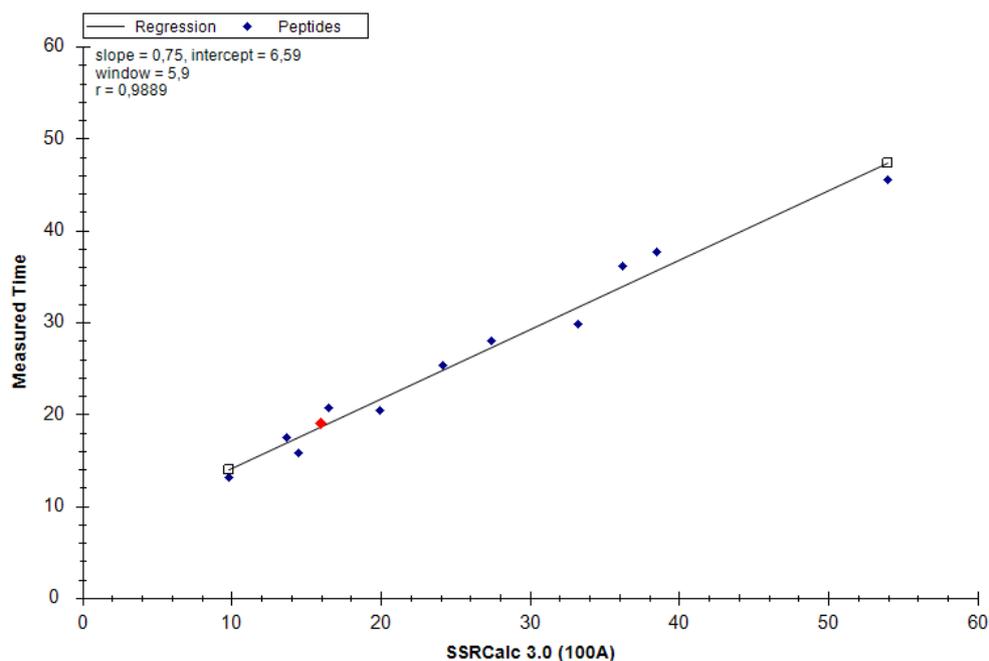


Figure 4.4. The residual plot shows the relationship between the SSRCalc and the measured RT. The relationship shows a high degree of correlation ($R=0.99$).

4.4.5. DIA data analysis

In DDA, individual precursors are selected for fragmentation in a semi-stochastic nature in shotgun proteomics favouring the most intense peaks, as a result a low percentage of the detectable peptides get fragmented and an even lower number are reliably identified (Michalski, Cox and Mann, 2011, Schubert *et al.*, 2015). This is the key aspect to DDA, the protein identifiability is much more reproducible than at the peptide level. Since only the selected precursors are fragmented, a large proportion of the information is lost. In DIA, in contrast, all precursors are fragmented and MS2 data is acquired for all fragment ions. This allows to record MS1 and MS2 data from virtually all peptides in a sample without loss of information (Gillet *et al.*, 2012, Egertson *et al.*, 2013).

In this study, the aim is to combine DDA's strength in discovery proteomics with the sensitivity and reproducibility of untargeted DIA (Rauniyar, 2015) to enhance the extraction of peptide-specific information from highly multiplexed DIA data in a targeted fashion. DIA implementations use wide pre-defined precursor fragmentation windows to acquire comprehensive fragment ion data (Ludwig *et al.*, 2018). In this study, all peptides within a defined variable mass-to-charge (m/z) window were fragmented (Figure 4.5). This will place window boundaries in regions where peptides do not generally occur. The use of variable

windows outperforms fixed windows in terms of quantification and identification (Zhang *et al.*, 2015). For the resulting highly multiplexed ion spectra for all detectable peptides over the elution time, an elaborate data processing tool is required (Tujin *et al.*, 2016). Skyline software platform (MacLean *et al.*, 2010) was used with settings provided in Table S4.3 of the Supporting Material. Skyline is a tool used for many instrument methods for interpretation of results. This vendor-neutral, user friendly and active support of Skyline made it an attractive option for processing the DIA data in a targeted fashion compared to the OpenSWATH software platform (Röst *et al.*, 2014). The quantitative metrics produced for 529 proteins, 2601 peptides and 3511 precursors. DIA integrations in Skyline were ran twice, first the retention time tolerance windows were set to 5-min and then with a 2-min window, respectively (Boswell, Abate-Pella and Hewitt, 2015). Fewer peptides were expected for the 2-min window Skyline document; however, there was no difference in the numbers of peptides or the retention times of the peptides between these two Skyline documents using these time windows.

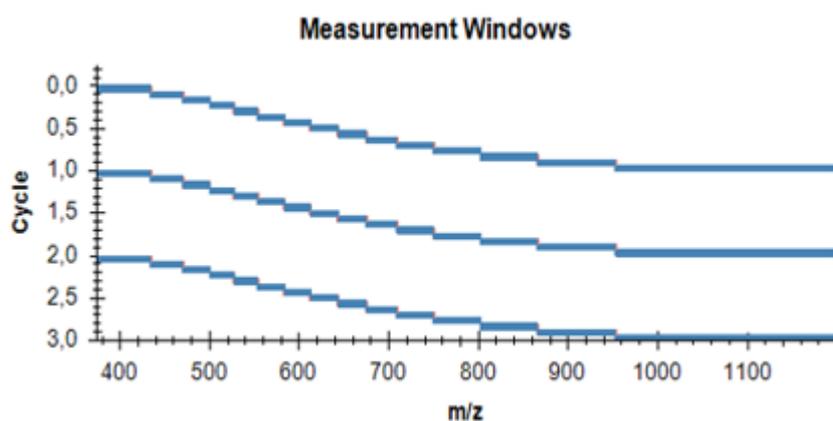


Figure 4.5. The isolation window scheme. The measurement windows derived from the DIA spectrum files used a variable isolation window scheme for the targeted DIA data analysis method.

4.5. Conclusion

Using the DDA workflow individual precursors are selected for fragmentation. This approach could limit the reliability of peptide identification. DDA data were subjected to traditional database searching with the MS-GF+ and MSFragger search engines employed, respectively. The protein assembly was then handled with a statistical tool, IDPicker using a 1% PSM FDR to achieve meaningful comparisons. The universal MS-GF+ search was

observed to identify more peptides than the ultrafast MSFragger search. The MSFragger search was done with a UniProtKB which provides protein accession with minimal redundancy due to human input and the integration with other databases. Therefore, these results were used to create a protein list containing the top protein identifiers of interest based on RS-score and filtered spectra. In the DIA workflow all precursors are fragmented and MS2 data were acquired for all fragment ions. Due to the complexity of the MS2 level data, database searching was not used, instead spectral library searching offers the ability to extract peptide-centric information from the multiplexed DIA data.

In this study, we leverage from DDA data to create a target protein and peptide list to achieve DIA data analysis in a targeted fashion. The quality and coverage of the spectral library is critical for peptide identifications. Two DDA-based spectral libraries were created using MS-GF+ and MSFragger search outputs, respectively. Here, we found that the MSFragger search outputs created a larger library compared to the MS-GF+ search outputs. The MSFragger DDA-based library was used to help create authentic standards derived from human albumin and human uromodulin peptides which prove here to be more reliable than using retention projection. The availability of open spectral libraries for DIA analysis are limited, a comprehensive human spectral library was created by a different laboratory. This library was used in the targeted DIA data analysis method using the Skyline software platform described here.

Chapter 5

Urine Proteome for the Identification of Biomarkers for Paediatric Tuberculosis

This Chapter evaluates the presence of urine protein biomarkers associated with paediatric TB. Data-independent acquisition (DIA) mass spectrometry was done by the Steen Laboratory as part of The Urine Proteomics Study conducted by Boston Children's Hospital. Unlike the initial analysis in Steen Lab employing Spectronaut software, the protein-level quantification and testing for differential abundance were performed on a Skyline document using the R-package, *MSstats* based on a linear mixed-effects model. These comparison tests identify the proteins with significantly different means in the data set by estimating fold changes and *p* values that have been adjusted to control the false discovery rate (FDR) at 0.05.

5.1. Materials

- RStudio Desktop (<https://rstudio.com/products/rstudio/download/#download>);
- R version 4.0.2 for Windows (<https://cran.r-project.org/bin/windows/base/>).

5.2. Equipment

- A 64-bit computer with Windows 10 operating system, 8 GB of RAM, a quad-core i5 processor, and more than 50 GB of free disk space.

5.3. Experimental Procedure

5.3.1. Differential protein abundance testing

A subset of South African human-immunodeficiency virus (HIV) negative (N=80) samples were selected from the integrated Skyline DIA document (see Chapter 4) using *dplyr* in R. The subjects who had active TB (N = 31, Age = 1.5 months – 14 years) were compared with controls who had no TB (N=49, Age = 1.5 months – 14 years) (matched for age, sex and HIV status) in a randomised blocked study design. The resulting Skyline document was converted to *MSstats* format (Choi *et al.*, 2014) using R. *MSstats* is suitable for researchers with a limited statistics and programming background. Peptides that were associated with more than one protein sequence were removed. The *dataProcess* function of *MSstats* was

used to calculate the abundance values for peptides using the median normalisation. The quality of the normalisation was visualised using boxplots. To find an additively fit model for the data, Tukey median polish was performed. For the comparisons between TB cases and controls, *MSstats* provided model-based estimates of fold changes as well as *p* values that were adjusted to control the FDR at the cut-off 0.05 (Benjamini and Hochberg, 1995). A volcano plot was used to visualize the differentially abundant proteins.

5.4. Results and Discussion

5.4.1. Discovery phase of premature urine protein biomarkers

Epidemiologic studies in industrialized countries and resource-limited countries with high burdens of TB found that younger populations, females, and persons of African or Asian origins had a higher risk for TB (Rée, 1997; Yang *et al.*, 2004). Even with these statistics, paediatric TB has not been prioritised, mainly due to the difficulty in diagnosis. This study aims to use an DIA data that has been optimised leveraging the protein and peptide identification abilities from DDA data. Generally, two-sample t-tests are used for significance testing, in which the relative or absolute abundances for each peptide or protein across the condition of interest would be compared. However, in a typically proteomics study samples sizes are often too small, which results in uncertainty of the true sample variability. A small sample size can exhibit a small observed fold change (Kammers *et al.*, 2015) which is biologically meaningless (Mccarthy and Smyth, 2009; Colquhoun, 2014). In a study by Clough *et al.* two MS samples were used to demonstrate that a simultaneous statistical model of all the relevant features and conditions yields a higher sensitivity of protein quantification as compared to commonly employed alternatives (Clough *et al.*, 2012). Later an open-source R-package, *MSstats* (Choi *et al.*, 2014) was developed to test for differential abundance on the protein-level using a linear mixed-effects model.

Overall, 529 proteins were detected with more than 2 peptides in Skyline. For statistical analysis, 491 proteins with valid intensity values in all samples were further considered. In this study, a subset of the Skyline document (see Chapter 4) containing fragment ion level information (Selevsek *et al.*, 2015) representing a South African HIV negative cohort consisting of 31 cases and 49 controls. The *MSstats dataProcess* function summarizes the quantitative experimental information from a set of fragmented peptides as well as the intensities, abundances for one condition (case versus control). Peptides that were used in more than one protein were removed. Intensities with a detection *q*-value below 0.01 were

replaced with zeros. The q-value reflects the aggregate error required to include an intensity value. Many normalisation methods have been adapted from the DNA microarray techniques for use on proteomics data (Callister *et al.*, 2006). *MSstats* uses median normalisation based on the assumption that samples are separated by a constant (Kultima *et al.*, 2009). The effects of this step for all proteins are visualised in Figure S5.1 of the Supporting Material. Data generated by the MS analysis are prone to biases, which can be accounted for with normalisation resulting in more reliable downstream analysis. Due to peptide-to-peptide variability, *MSstats* provides a nonparametric test on normalised peptide values, thus minimising the number of free parameters, as well as for measuring significance with permutation testing (Slama *et al.*, 2018).

For each comparison, a contrast matrix was created. Meena Choi (developer of the *MSstats* R-package) assisted in the development of the contrast matrix. The *groupComparison* function uses the matrix to compute fold changes and *p* values that were adjusted to control the FDR at the cut-off 0.05 (Benjamini and Hochberg, 1995). A volcano plot was used to illustrate the significantly expressed proteins (Figure 5.1). The comparisons between the two groups, cases and controls were graphically represented to display the quantitative data. Three human proteins, leucine-rich alpha-2-glycoprotein (A2GL), aggrecan core protein (PGCA) and cartilage intermediate layer protein 2 (CILP2) were identified as statistically significant (Figure 5.1). A2GL were upregulated with a 2-fold increase, while PGCA and CILP2 were downregulated with FCs between 0 and -1. All proteins detected were quantified by group, A2GL was more abundant in TB case samples, while PGCA and CILP2 was more abundant in control samples.

A2GL plays a role in transforming growth factor beta receptor binding. This protein is also involved in positive regulation of endothelial cell proliferation and the biological response to bacterium. In a study by Yang *et al.* plasma leucine-rich alpha-2-glycoprotein was identified as an innovative biomarker for inflammation in association with the progression of End-stage renal disease (ESRD) (Yang *et al.*, 2020). PGCA has a binding role for carbohydrates, the extracellular matrix structure, hyaluronic acid and metal ions. This protein is also involved in cartilage condensation, cell adhesion, the development of the central nervous system, heart development and skeletal system development. Aggrecan cleavage is a sign of cartilage degeneration in disease such as rheumatoid arthritis and osteoarthritis (Hsueh, Kraus and Önnarfjord, 2017; Suna *et al.*, 2018). Limited Gene Ontology (GO) annotations has been made for CILP-2; however, it has a related function to PGCA in cartilage. CILP-2 has also been linked to cartilage degenerative diseases (Bernardo *et al.*, 2011).

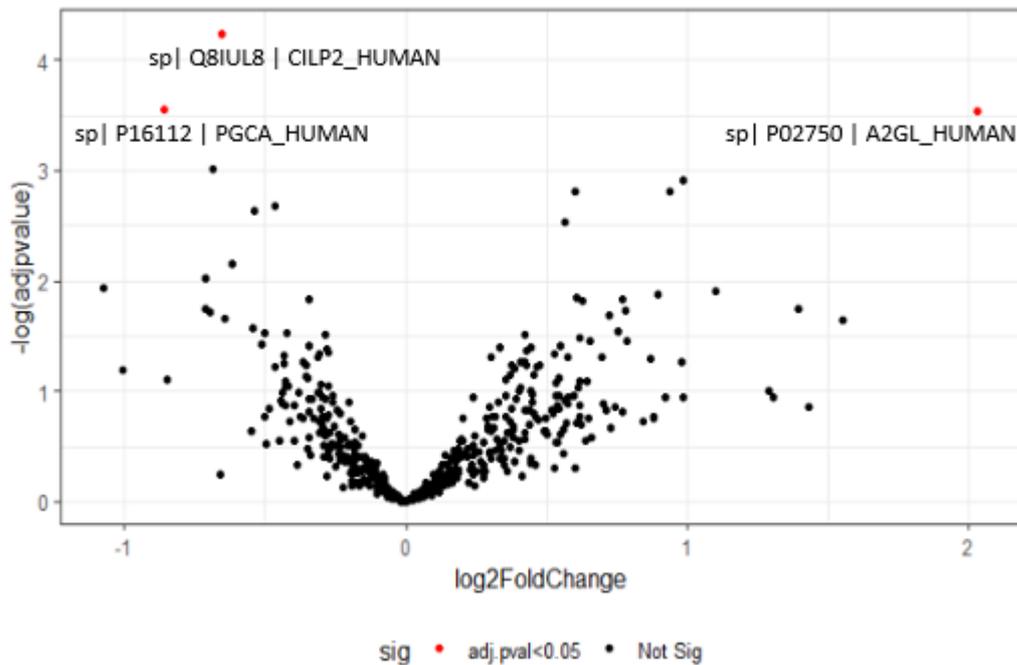


Figure 5.1. Volcano plot illustrates significantly differentially abundant proteins. The $-\log_{10}$ (Benjamini-Hochberg corrected P value) is plotted against the \log_2 (fold change).

5.4. Conclusion

Liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) is widely used for such quantitative proteomic investigations. The typical output of such studies is a list of identified and quantified peptides from a tool such as Skyline. The biological and clinical interest is, however, usually focused on quantitative conclusions at the protein level. *MSstats* was employed to generate comparisons between the two groups, cases and controls for the South African TB cohort were graphically represented to display the quantitative data. Three human proteins, leucine-rich alpha-2-glycoprotein (A2GL), aggrecan core protein (PGCA) and cartilage intermediate layer protein 2 (CILP2) were identified as significantly different. A2GL were upregulated with a 2-fold increase and PGCA and CILP-2 were downregulated with FCs between 0 and -1. Both PGCA and CILP-2 plays a role in cartilage degeneration, while A2GL is an important biomarker for inflammation associated with the progression of ESRD. Further investigations should be done to validate these findings. This study generated knowledge that could have important clinical applications on urine protein biomarker discovery for paediatric TB.

Chapter 6

Conclusion, limitations, and future recommendations

Quality metrics were successfully generated using QuaMeter “IDFree” software on the DDA data and SwaMe software on the DIA data. The PCA showed that the batch effects were minimal. The distribution of density and the number of scans appeared important to the variability for both DIA and DDA, although in DDA they were all related to MS1, whereas in DIA they were all related to MS2. TIC impacted only the DDA data but not the DIA, possibly because SwaMe comprehensive metrics do not contain any TIC measures. There was also no metric for MS1-Density changes in SwaMe; only MS2-Density changes are observed. SwaMe does produce a metric for MS1 Count, yet it only appeared a significant source of variability in the DDA data. For further studies, the additional metrics generated for SwaMe should be included in PCA analysis. A broader diversity of metrics should enable a closer look at DIA variability.

Dissimilarity assessments were performed by calculating Euclidean median distances in the major components and comparing run-to-run. Boxplots were used to identify outliers as the runs above the top whisker. There were 18 outliers in the DDA experiment and 14 different outliers in the DIA experiment, which were excluded during the urine protein biomarker discovery phase. Future recommendations would be to exclude the outliers and then redo the PCA to identify whether the method detects new possible outliers with the same sources of variability. It would also be useful to develop quality control models for dynamic performance monitoring that can allow for systematic incorporation of the insights of laboratory technicians. The decision-making of whether to re-run those samples is open to the researcher, potentially driving experimental costs higher.

The DDA workflow selected individual precursors for fragmentation. DDA data were subjected to traditional database searching via the MS-GF+ and MSFragger search engines. Protein assembly was then handled with a statistical tool, IDPicker using a 1% PSM FDR to achieve meaningful comparisons. The MS-GF+ search was observed to identify more peptides than the ultrafast MSFragger search. The MSFragger search was done with a UniProtKB which provides protein accession with minimal redundancy due to human input and the integration with other databases. Therefore, these results were used to create a protein list containing the top protein identifiers of interest based on reproducibility of detection and the number of PSMs observed per protein. In the DIA workflow all precursors are fragmented and MS2 data were acquired for all fragment ions. Due to the complexity of

the MS2 level data, database searching was not used, instead spectral library searching offers the ability to extract peptide-centric information from the multiplexed DIA data.

The DIA data analysis method leveraged DDA data to create lists of quantifiable proteins and peptides. This optimisation allows for greater sensitivity in peptide quantification during DIA analysis in the Skyline environment. Although MS-GF+ identified more peptides by database searching, MSFragger detected more distinct peptides by spectral library creation from a smaller set of identified peptides. The magnitude of sensitivity differences between the search engines, however, was not large. Future studies may include PSM filtering software that employs distributional analysis or machine learning classifiers to boost overall sensitivity.

In this study, commercially available iRT peptides were not spiked into samples during the experiment. Standard calibrant peptides were successfully selected from the two most abundant proteins commonly found in human urine: serum albumin and uromodulin. This allows the Skyline software to predict more accurately when a peptide should appear in a 50-minute DIA experiment. The output of the DIA analysis method was a list of identified and quantified peptides of interest. The biological and clinical interest, however, focusses on the quantitative conclusion made at the protein level. *MSstats* was employed to generate comparisons between the two groups, cases and controls for the South African TB cohort were graphically represented to display the quantitative data. Three human proteins, leucine-rich alpha-2-glycoprotein (A2GL), aggrecan core protein (PGCA) and cartilage intermediate layer protein 2 (CILP2) were identified as significantly different. A2GL were upregulated with a 2-fold increase and PGCA and CILP-2 were downregulated with FCs between 0 and -1. Both PGCA and CILP-2 plays a role in cartilage degeneration, while A2GL is an important biomarker for inflammation associated with the progression of ESRD. Further investigations using new cohorts of patient data or with alternative methods such as ELISA may validate these proteins as significantly different. This study generated knowledge that could have important clinical applications on urine protein biomarker discovery for paediatric TB.

We hypothesized that the quantity of specific host proteins in urine is different for children with TB compared to symptomatic control children who do not have TB. In this study, we showed that by using an optimised DIA data analysis method leveraging DDA data we can identify abundantly differentially expressed proteins. Further investigation is required to determine the impact of the significant proteins, and whether they would hold up during the validation phase. This study highlights the ability of the advances in MS technology to discover novel protein biomarkers in urine for paediatric TB diagnosis.

Supplementary Material**Chapter 3****Table S3.1.** *QuaMeter IDfree and SwaMe metrics and execution*

The QuaMeter software was executed with this command line:

```
quameter.exe *. mzML -cpus 1 -MetricsType idfree -OutputFilepath metrics.tsv
```

The SwaMe software was executed with this command line:

```
Yamato.Console.exe *. mzML -d 10 -dir true
```

The quameter.cfg file used for the orbitrap configuration included the following:

```
ChromatogramMzLowerOffset = ".05mz"
```

```
ChromatogramMzUpperOffset = ".05mz"
```

```
Instrument = "Orbi"
```

Table S3.2. *The undivided quality metrics generated by SwaMe explained (Kriek, unpublished)*

Undivided metrics	
Metric	Definition
MissingScans	The number of scans where there was not a single ion detected.
RTDuration	Difference between the first scan start time and the last scan start time
SwathSizeDifference	Difference between the largest swath and the smallest swath
MS2Count	Number of MS2 scans in the entire run
MS1Count	Number of MS1 scans in the entire run
SwathsPerCycle	Number of swaths in the same cycle
TotalMS2IonCount	Number of ions detected in all MS2scans across the run
TotalMS1IonCount	Number of ions detected in all MS1scans across the run
MS2Density50	The median number of ions in all MS2 scans
MS2DensityIQR	The IQR for the number of ions detected in all MS2 Scans

Table S3.3. Medians and interquartile ranges (IQR) of dissimilarity

	Samples	Median	IQR
Figure 3.3A: DDA experiments from the Thermo Q-Exactive MS (in the same lab)	237	4.4	1.8
Figure 3.3B: DIA experiments from the Thermo Q-Exactive MS (in the same lab)	225	2.4	1.0

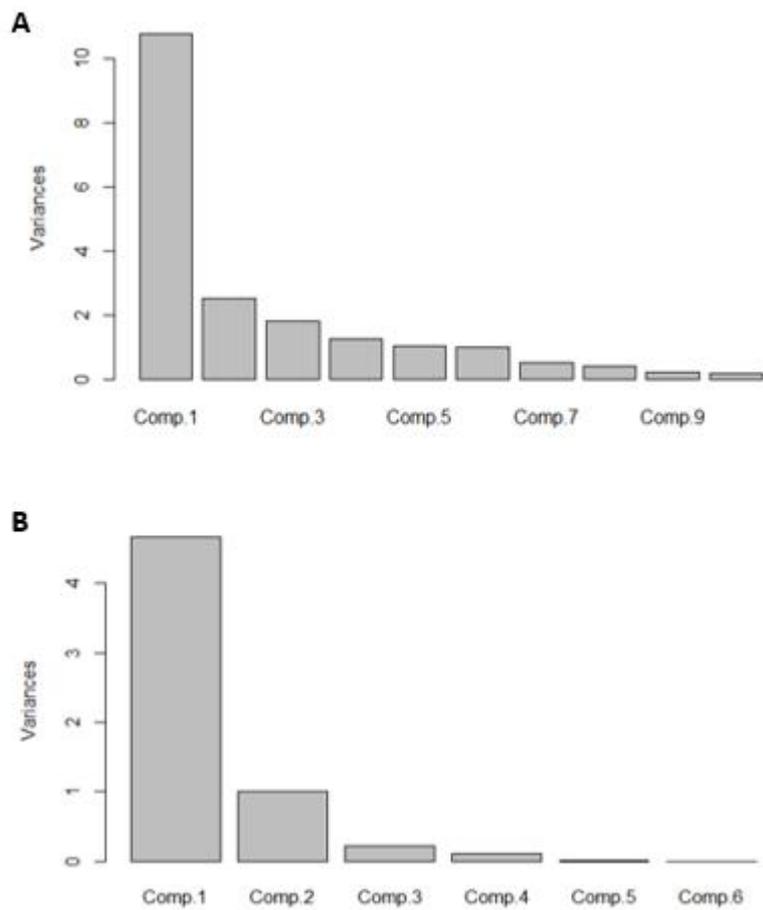


Figure S3.1. Scree plot shows the contribution of variance per component, where component 1 and 2 (PC1 and PC2) accounted for the most variance in the DDA dataset (**A**) and in the DIA dataset (**B**).

Chapter 4

Table S4.1. Database search engine executions

Run Philosopher from the command line to download protein sequences from UniProt.

Execute the following two commands:

```
philosopher_windows_amd64.exe workspace --init
```

```
philosopher_windows_amd64.exe database --reviewed --contam --id UP000005640
```

MSFragger Database Searching command:

```
java -d64 -Xmx8G -jar R:\Research\MSFragger-2.3\MSFragger-2.3.jar 20200407-  
Tryptic.params raws\*.raw
```

MS-GF+ Database Searching command:

```
java -Xmx8G -jar MSGFPlus.jar -d 20180308-Ensembl-Mtb-Human-Cntms.fasta -t 20ppm  
-e 1 -m 3 -inst 1 -ntt 1 -thread 2 -tda 1 -ti 0,1 -n 1 -maxLength 50 -mod modifications.txt
```

Table S4.2. 23 peptides for standard calibrant peptide selection. The number label assigned to each peptide sequence from the two most abundant human proteins, albumin, and uromodulin.

Number Label	Peptide	Protein
1	LVRPEVDVMCTAFHDNEETFLK	albumin
2	SLHTLFGDK	albumin
3	LCTVATLR	albumin
4	AVMDDFAAFVEK	albumin
5	ETYGEMADCCAK	albumin
6	VFDEFKPLVEEPQNLIK	albumin
7	HPYFYAPELFFAK	albumin
8	DVFLGMFLYEYAR	albumin
9	RPCFSALEVDETYVPK	albumin
10	HPDYSVLLLR	albumin
11	QTALVELVK	albumin
12	SHCIAEVENDEMPADLPSLAADFVESK	albumin
13	YICENQDSISSK	albumin
14	DSTIQVVENGESSQGR	uromodulin
15	LECGANDMK	uromodulin
16	DWVSVVTPAR	uromodulin
17	STEYGEGYACDTDLR	uromodulin
18	VFMYLSDSR	uromodulin
19	MAETCVPVLR	uromodulin
20	FAGNYDLVYLHCEVYLCDTMNEK	uromodulin
21	VGGTGMFTVR	uromodulin
22	DGPCGTVLTR	uromodulin
23	ACAHWSGHCCCLWDASVQVK	uromodulin

Table S4.3. Settings used in the Skyline proteomics environment for optimised DIA analysis.

Transition settings	
Filter	Peptide precursor charges: 2, 3
	Ion charges: 1
	Ion types: y, b
	Product ion selection: From m/z>precursor to 3 ions
Library	Ion match tolerance: 0.5 m/z
	Pick 3 product ions
	Select from filtered ion charges and types
Full-Scan	MS/MS filtering
	DIA Acquisition method
	Orbitrap product mass analyser
	Isolation scheme
	Resolving power 60 000 at 400 m/z
	Retention time filtering
	Use only scans within 5 minutes of MS/MS IDs
Peptide settings	
Prediction	Retention time predictor
	5 min time window
Library	Spectronaut Library
	Pick peptides matching library
	Rank peptides by picked intensity
Modifications	None

Chapter 5

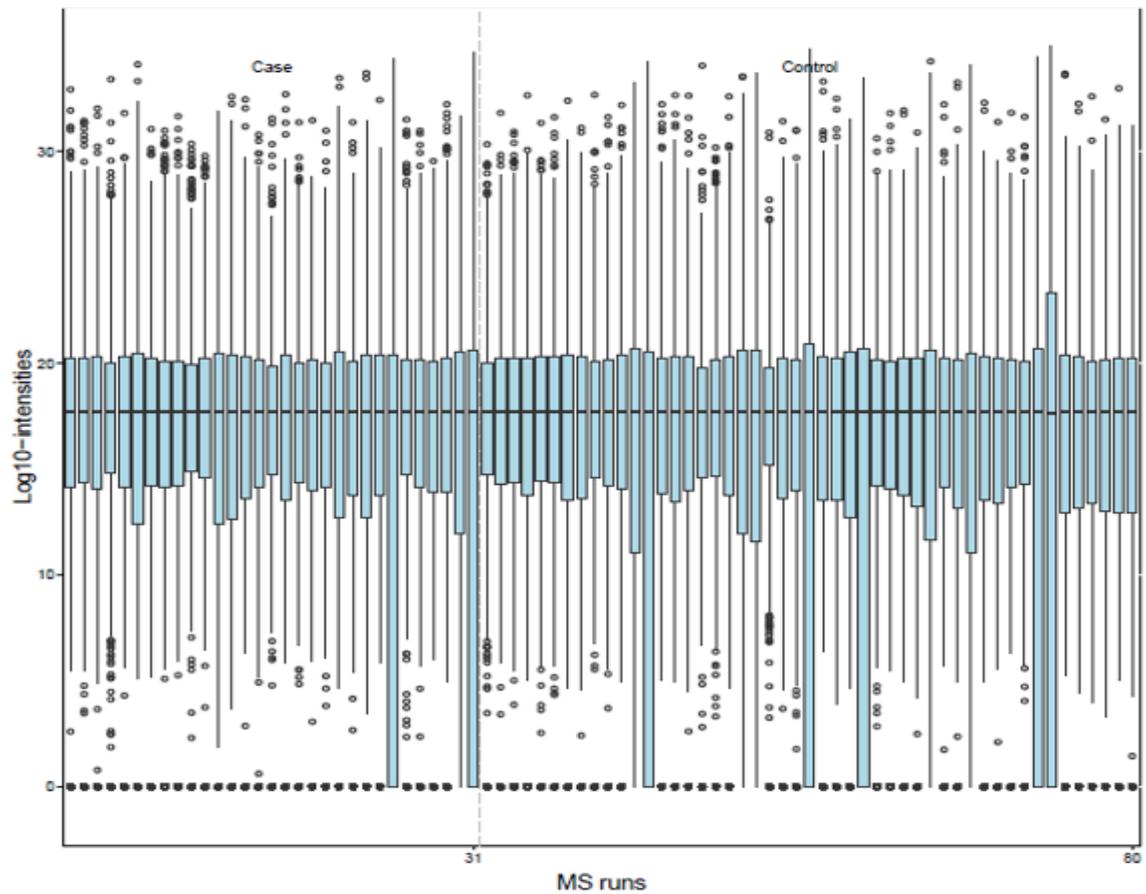


Figure S5.1. A quality control plot for all endogenous proteins. This plot was generated using MSstats assessing the effects of the normalisation step.

Chapter 7

References

1. Azarkan, M. *et al.* (2007) 'Affinity chromatography: A useful tool in proteomics studies', *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, 849(1–2), pp. 81–90. doi: 10.1016/j.jchromb.2006.10.056.
2. Barratt, J. and Topham, P. (2007) 'Urine proteomics: The present and future of measuring urinary protein components in disease', *CMAJ*, 177(4), pp. 361–368. doi: 10.1503/cmaj.061590.
3. Beasley-Green, A. (2016) 'Urine proteomics in the era of mass spectrometry', *International Neurourology Journal*. Korean Continence Society, 20(2), pp. 70–75. doi: 10.5213/inj.1612720.360.
4. Benjamini, Y. and Hochberg, Y. (1995) 'Benjamini-1995.pdf', *Journal of the Royal Statistical Society B*, 57(1), pp. 289–300. doi: 10.2307/2346101.
5. Bernardo, B. C. *et al.* (2011) 'Cartilage intermediate layer protein 2 (CILP-2) is expressed in articular and meniscal cartilage and down-regulated in experimental osteoarthritis', *Journal of Biological Chemistry*, 286(43), pp. 37758–37767. doi: 10.1074/jbc.M111.248039.
6. Bernhardt, O. M. *et al.* (2012) 'Spectronaut', *Biognosys*, p. 2012. Available at: <https://biognosys.com/media.ashx/spectronaut-a-fast-and-efficient-algorithm-for-mrm-like-swath-processing.pdf>.
7. Bittremieux, W. *et al.* (2017) 'Computational quality control tools for mass spectrometry proteomics', *Proteomics*, 17(3–4), pp. 3–4. doi: 10.1002/pmic.201600159.
8. Bittremieux, W. *et al.* (2018) 'Quality control in mass spectrometry-based proteomics', *Mass Spectrometry Reviews*, 37(5), pp. 697–711. doi: 10.1002/mas.21544.
9. Bjornson, R. D. *et al.* (2008) 'X!Tandem, an Improved Method for Running X!Tandem in Parallel on Collections of Commodity Computers', *Journal of Proteome Research*, 7(1), pp. 293–299. doi: 10.1021/pr0701198.
10. Boswell, P. G., Abate-Pella, D. and Hewitt, J. T. (2015) 'Calculation of retention time tolerance windows with absolute confidence from shared liquid chromatographic retention data', *Journal of Chromatography A*, 1412(3), pp. 52–58. doi: 10.1016/j.chroma.2015.07.113.
11. Bruderer, R. *et al.* (2016) 'High-precision iRT prediction in the targeted analysis of data-independent acquisition and its impact on identification and quantitation',

- Proteomics*, 16(15–16), pp. 2246–2256. doi: 10.1002/pmic.201500488.
12. Bruderer, R. *et al.* (2017) 'Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results', *Molecular and Cellular Proteomics*, 16(12), pp. 2296–2309. doi: 10.1074/mcp.RA117.000314.
 13. Callister, S. J. *et al.* (2006) 'Normalization Approaches for Removing Systematic Biases Associated with Mass Spectrometry and Label-Free Proteomics', *Journal of Proteome Research*, 5(2), pp. 277–286. doi: 10.1021/pr050300l.
 14. Caterino, M. *et al.* (2018) 'Urine proteomics revealed a significant correlation between urine-Fibronectin abundance and estimated-GFR decline in patients with bardet-Biedl syndrome', *Kidney and Blood Pressure Research*, 43(2), pp. 389–405. doi: 10.1159/000488096.
 15. Chapman, J. D., Goodlett, D. R. and Masselon, C. D. (2014) 'Multiplexed and data-independent tandem mass spectrometry for global proteome profiling', *Mass Spectrometry Reviews*, 33(6), pp. 452–470. doi: 10.1002/mas.21400.
 16. Chen, C., Huang, H. and Wu, C. H. (2017) 'Protein Bioinformatics Databases and Resources', in *Methods in Molecular Biology*, 1558, pp. 3–39. doi: 10.1007/978-1-4939-6783-4_1.
 17. Choi, M. *et al.* (2014) 'MSstats: An R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments', *Bioinformatics*, 30(17), pp. 2524–2526. doi: 10.1093/bioinformatics/btu305.
 18. Clough, T. *et al.* (2012) 'Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs.', *BMC bioinformatics*, 13 (16), pp. 16. doi: 10.1186/1471-2105-13-S16-S6.
 19. Colquhoun, D. (2014) 'An investigation of the false discovery rate and the misinterpretation of p-values', *Royal Society Open Science*, 1(3). p. 9. doi: 10.1098/rsos.140216.
 20. Courchesne, P. L. *et al.* (1998) 'Optimization of capillary chromatography ion trap-mass spectrometry for identification of gel-separated proteins', *Electrophoresis*, 19(6), pp. 956–967. doi: 10.1002/elps.1150190611.
 21. Decramer, S. *et al.* (2008) 'Urine in clinical proteomics', *Molecular and Cellular Proteomics*, 7(10), pp. 1850–1862. doi: 10.1074/mcp.R800001-MCP200.
 22. Demichev, V. *et al.* (2020) 'DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput', *Nature Methods*, 17(1), pp. 41–44. doi: 10.1038/s41592-019-0638-x.
 23. Deutsch, E. W. (2010) 'Mass Spectrometer Output File Format mzML', *Methods in molecular biology (Clifton, N.J.)*, 604(August), pp. 319–331. doi: 10.1007/978-1-

60761-444-9_22.

24. Ding, H. *et al.* (2020) 'Urine Proteomics: Evaluation of Different Sample Preparation Workflows for Quantitative, Reproducible, and Improved Depth of Analysis', *Journal of Proteome Research*, 19(4), pp. 1857–1862. doi: 10.1021/acs.jproteome.9b00772.
25. Dodd, P. J. *et al.* (2014) 'Burden of childhood tuberculosis in 22 high-burden countries: A mathematical modelling study', *The Lancet Global Health*. Dodd *et al.* Open Access article distributed under the terms of CC BY, 2(8), pp. e453–e459. doi: 10.1016/S2214-109X(14)70245-1.
26. Dodd, P. J. *et al.* (2018) 'Potential effect of household contact management on childhood tuberculosis: a mathematical modelling study', *The Lancet Global Health*. The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license, 6(12), pp. e1329–e1338. doi: 10.1016/S2214-109X(18)30401-7.
27. Domon, B. and Aebersold, R. (2010) 'Options and considerations when selecting a quantitative proteomics strategy', *Nature Biotechnology*, 28(7), pp. 710–721. doi: 10.1038/nbt.1661.
28. Duangkumpha, K. *et al.* (2019) 'Urine proteomics study reveals potential biomarkers for the differential diagnosis of cholangiocarcinoma and periductal fibrosis', *PLoS ONE*, 14(8), pp. 1–17. doi: 10.1371/journal.pone.0221024.
29. Egertson, J. D. *et al.* (2013) 'Multiplexed MS/MS for improved data-independent acquisition', *Nature Methods*, 10(8), pp. 744–746. doi: 10.1038/nmeth.2528.
30. Egertson, J. D. *et al.* (2015) 'Multiplexed peptide analysis using data-independent acquisition and Skyline', *Nature Protocols*, 10(6), pp. 887–903. doi: 10.1038/nprot.2015.055.
31. Escher, C. *et al.* (2012) 'Using iRT, a normalized retention time for more targeted measurement of peptides', *PROTEOMICS*, 12(8), pp. 1111–1121. doi: 10.1002/pmic.201100463.
32. Faria, S. S. *et al.* (2017) 'A Timely shift from shotgun to targeted proteomics and how it can be groundbreaking for cancer research', *Frontiers in Oncology*, 7(2), pp. 13. doi: 10.3389/fonc.2017.00013.
33. Fernández-Costa, C., Martínez-Bartolomé, S., McClatchy, D. B., *et al.* (2020) 'Impact of the Identification Strategy on the Reproducibility of the DDA and DIA Results', *Journal of Proteome Research*, 19(8), pp. 3153–3161. doi: 10.1021/acs.jproteome.0c00153.
34. Fernández-Costa, C., Martínez-Bartolomé, S., McClatchy, D., *et al.* (2020) 'Improving Proteomics Data Reproducibility with a Dual-Search Strategy', *Analytical Chemistry*, 92(2), pp. 1697–1701. doi: 10.1021/acs.analchem.9b04955.
35. Fliser, D. *et al.* (2007) 'Advances in urinary proteome analysis and biomarker

- discovery', *Journal of the American Society of Nephrology*, 18(4), pp. 1057–1071. doi: 10.1681/ASN.2006090956.
36. Forshed, J. (2017) 'Experimental Design in Clinical 'Omics Biomarker Discovery', *Journal of Proteome Research*, 16(11), pp. 3954–3960. doi: 10.1021/acs.jproteome.7b00418.
37. Frank, A. M. (2009) 'A ranking-based scoring function for peptide-spectrum matches', *Journal of Proteome Research*, 8(5), pp. 2241–2252. doi: 10.1021/pr800678b.
38. Frewen, B. E. *et al.* (2006) 'Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries', *Analytical Chemistry*, 78(16), pp. 5678–5684. doi: 10.1021/ac060279n.
39. Frewen, B. and MacCoss, M. J. (2007) 'Using BiblioSpec for Creating and Searching Tandem MS Peptide Libraries', *Current Protocols in Bioinformatics*. Wiley, 20(1), pp. 7-13. doi: 10.1002/0471250953.bi1307s20.
40. Gillet, L. C. *et al.* (2012) 'Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis', *Molecular and Cellular Proteomics*. American Society for Biochemistry and Molecular Biology Inc., 11(6), p. 111. doi: 10.1074/mcp.O111.016717.
41. Girardi, E. and Ippolito, G. (2016) 'A new era in the control of tuberculosis', *Infectious Disease Reports*, 8(2), p. 57. doi: 10.4081/idr.2016.6644.
42. Good, D. M. *et al.* (2007) 'Body fluid proteomics for biomarker discovery: Lessons from the past hold the key to success in the future', *Journal of Proteome Research*, 6(12), pp. 4549–4555. doi: 10.1021/pr070529w.
43. Graham, S. M. *et al.* (2012) 'Evaluation of tuberculosis diagnostics in children: 1. Proposed clinical case definitions for classification of intrathoracic tuberculosis disease. Consensus from an expert panel', *Journal of Infectious Diseases*, 205(2), pp. 199–208. doi: 10.1093/infdis/jis008.
44. Grewal, R. *et al.* (2015) 'Biomarker discovery for diagnosis and treatment of tuberculosis: a role for biobanking?', *Journal of Biorepository Science for Applied Medicine*, 3(1), pp. 47-56. doi: 10.2147/bsam.s64571.
45. Guan, S. *et al.* (2020) 'Data Dependent-Independent Acquisition (DDIA) Proteomics', *Journal of Proteome Research*. 19(8), pp. 3230-3237. doi: 10.1021/acs.jproteome.0c00186.
46. Holmberg, P. J., Temesgen, Z. and Banerjee, R. (2019) 'Tuberculosis in children', *Pediatrics in Review*, 40(4), pp. 168–178. doi: 10.1542/pir.2018-0093.
47. Hsueh, M.-F., Kraus, V. and Önnérjörd, P. (2017) 'Cartilage matrix remodelling differs by disease state and joint type', *European Cells and Materials*, 34(1), pp. 70–

82. doi: 10.22203/eCM.v034a05.
48. Huang, T. *et al.* (2020) 'Combining precursor and fragment information for improved detection of differential abundance in data independent acquisitions', *Molecular and Cellular Proteomics*, 19(2), pp. 421–430. doi: 10.1074/mcp.RA119.001705.
49. Jeong, K., Kim, S. and Bandeira, N. (2012) 'False discovery rates in spectral identification.', *BMC bioinformatics*. BioMed Central Ltd, 13(16), p. S2. doi: 10.1186/1471-2105-13-S16-S2.
50. Kaiser, H. F. (1958) 'The varimax criterion for analytic rotation in factor analysis', *Psychometrika*, 23(3), pp. 187–200. doi: 10.1007/BF02289233.
51. Kalantari, S. *et al.* (2015) 'Human Urine Proteomics: Analytical Techniques and Clinical Applications in Renal Diseases', *International Journal of Proteomics*. Hindawi Publishing Corporation, 11, pp. 221-232. doi: 10.1155/2015/782798.
52. Kalli, A. *et al.* (2014) 'Evaluation and optimization of mass spectrometric mode: Focus on LTQ-Orbitrap Mass analyzers', *Journal of Proteome Research*, 12(7), pp. 3071–3086. doi: 10.1021/pr3011588.Evaluation.
53. Kammers, K. *et al.* (2015) 'Detecting significant changes in protein abundance', *EuPA Open Proteomics*, 7, pp. 11–19. doi: 10.1016/j.euprot.2015.02.002.
54. Kapp, E. and Schütz, F. (2007) 'Overview of tandem mass spectrometry (MS/MS) database search algorithms.', *Current protocols in protein science / editorial board, John E. Coligan ... [et al.]*, 25(8), pp. 1–19. doi: 10.1002/0471140864.ps2502s49.
55. Kentsis, A. *et al.* (2009) 'Urine proteomics for profiling of human disease using high accuracy mass spectrometry', *PROTEOMICS - CLINICAL APPLICATIONS*, 3(9), pp. 1052–1061. doi: 10.1002/prca.200900008.
56. Kertesz-Farkas, A. *et al.* (2012) 'Database Searching in Mass Spectrometry Based Proteomics', *Current Bioinformatics*, 7(2), pp. 221–230. doi: 10.2174/157489312800604354.
57. Kessner, D. *et al.* (2008) 'ProteoWizard: Open source software for rapid proteomics tools development', *Bioinformatics*, 24(21), pp. 2534–2536. doi: 10.1093/bioinformatics/btn323.
58. Kim, S., Gupta, N. and Pevzner, P. A. (2008) 'Spectral Probabilities and Generating Functions of Tandem Mass Spectra: A Strike against Decoy Databases', *Journal of Proteome Research*, 7(8), pp. 3354–3363. doi: 10.1021/pr8001244.
59. Kim, S. and Pevzner, P. A. (2014) 'Universal database search tool for proteomics.', *Nature communications*, 5, p. 5277. doi: 10.1038/ncomms6277.
60. Kong, A. T. *et al.* (2017) 'MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics', *Nature Methods*, 14(5), pp. 513-520. doi: 10.1038/nmeth.4256.

61. Krokhin, O. V. (2006) 'Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: Application to 300- and 100-Å pore size C18 sorbents', *Analytical Chemistry*, 78(22), pp. 7785–7795. doi: 10.1021/ac060777w.
62. Kultima, K. *et al.* (2009) 'Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides', *Molecular and Cellular Proteomics*, 8(10), pp. 2285–2295. doi: 10.1074/mcp.M800514-MCP200.
63. Li, M. *et al.* (2010) 'Comparative shotgun proteomics using spectral count data and quasi-likelihood modeling', *Journal of Proteome Research*, 9(8), pp. 4295–4305. doi: 10.1021/pr100527g.
64. Li, W. *et al.* (2019) 'Assessing the Relationship Between Mass Window Width and Retention Time Scheduling on Protein Coverage for Data-Independent Acquisition', *Journal of the American Society for Mass Spectrometry*. *Journal of The American Society for Mass Spectrometry*, 30(8), pp. 1396–1405. doi: 10.1007/s13361-019-02243-1.
65. Lin, L. *et al.* (2018) 'Fast quantitative urinary proteomic profiling workflow for biomarker discovery in kidney cancer', *Clinical Proteomics*. BioMed Central, 15(1), pp. 1–12. doi: 10.1186/s12014-018-9220-2.
66. Ludwig, C. *et al.* (2018) 'Data-independent acquisition-based SWATH - MS for quantitative proteomics: a tutorial', *Molecular Systems Biology*. EMBO, 14(8). p. 58. doi: 10.15252/msb.20178126.
67. Ma, Z. Q. *et al.* (2009) 'IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering', *Journal of Proteome Research*, 8(8), pp. 3872–3881. doi: 10.1021/pr900360j.
68. Ma, Z. Q. *et al.* (2012) 'QuaMeter: Multivendor performance metrics for LC-MS/MS proteomics instrumentation', *Analytical Chemistry*, 84(14), pp. 5845–5850. doi: 10.1021/ac300629p.
69. MacLean, B. *et al.* (2010) 'Skyline: An open source document editor for creating and analyzing targeted proteomics experiments', *Bioinformatics*, 26(7), pp. 966–968. doi: 10.1093/bioinformatics/btq054.
70. MacLean, E. *et al.* (2020) 'Advances in Molecular Diagnosis of Tuberculosis', *Journal of Clinical Microbiology*. Edited by C. Suzanne Kraft, 58(10), pp. 1–13. doi: 10.1128/JCM.01582-19.
71. Mann, M. *et al.* (2011) 'Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer', *Molecular & Cellular Proteomics*, 10(9), p. M111.011015. doi: 10.1074/mcp.m111.011015.
72. Marais, B. J. and Pai, M. (2006) 'Specimen collection methods in the diagnosis of

- childhood tuberculosis', 24(4), pp. 249–252. doi: 10.4103/0255-0857.29381.
73. Martens, L. (2013) 'Bringing proteomics into the clinic: The need for the field to finally take itself seriously', *Proteomics - Clinical Applications*, 7(5–6), pp. 388–391. doi: 10.1002/prca.201300020.
74. Martinez, L. *et al.* (2020) 'The risk of tuberculosis in children after close exposure: a systematic review and individual-participant meta-analysis', *The Lancet*, 395(10228), pp. 973–984. doi: 10.1016/S0140-6736(20)30166-5.
75. Mccarthy, D. J. and Smyth, G. K. (2009) 'Testing significance relative to a fold-change threshold is a TREAT', *Bioinformatics*, 25(6), pp. 765–771. doi: 10.1093/bioinformatics/btp053.
76. Michalski, A., Cox, J. and Mann, M. (2011) '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*, 10(4), pp. 1785–1793. doi: 10.1021/pr101060v.
77. Muntel, J. *et al.* (2015) 'Advancing urinary protein biomarker discovery by data-independent acquisition on a quadrupole-orbitrap mass spectrometer', *Journal of Proteome Research*, 14(11), pp. 4752–4762. doi: 10.1021/acs.jproteome.5b00826.
78. Navarro, P. *et al.* (2017) 'Europe PMC: A multi-center study benchmarks software tools for label-free proteome quantification', 34(11), pp. 1130–1136. doi: 10.1038/nbt.3685.A.
79. Newton, S. *et al.* (2008) 'Paediatric Tuberculosis (UKPMC Funders Group)', *Lancet Infect. Dis.*, 8(8), pp. 498–510. doi: 10.1016/S1473-3099(08)70182-8.Paediatric.
80. Noble, W. S. and MacCoss, M. J. (2012) 'Computational and statistical analysis of protein mass spectrometry data', *PLoS Computational Biology*, 8(1). doi: 10.1371/journal.pcbi.1002296.
81. Peter, J. *et al.* (2010) 'Urine for the diagnosis of tuberculosis: current approaches, clinical applicability, and new developments', *Current Opinion in Pulmonary Medicine*, 16(3), pp. 262–270. doi: 10.1097/MCP.0b013e328337f23a.
82. Rauniyar, N. (2015) 'Parallel reaction monitoring: A targeted experiment performed using high resolution and high mass accuracy mass spectrometry', *International Journal of Molecular Sciences*, 16(12), pp. 28566–28581. doi: 10.3390/ijms161226120.
83. Rée, H. (1997) 'TB/HIV: A clinical manual', *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 91(6), pp. 731–732. doi: 10.1016/s0035-9203(97)90551-4.
84. Ringnér, M. (2008) 'What is principal component analysis?', *Nature Biotechnology*, 26(3), pp. 303–304. doi: 10.1038/nbt0308-303.

85. Rosenberger, G. *et al.* (2014) 'A repository of assays to quantify 10,000 human proteins by SWATH-MS', *Scientific Data*, 1, pp. 1–15. doi: 10.1038/sdata.2014.31.
86. Rosenberger, G. *et al.* (2018) 'Europe PMC Funders Group Statistical control of peptide and protein error rates in large- scale targeted DIA analyses', 14(9), pp. 921–927. doi: 10.1038/nmeth.4398.Statistical.
87. Röst, H. L. *et al.* (2014) 'OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data', *Nature Biotechnology*, 32(3), pp. 219–223. doi: 10.1038/nbt.2841.
88. Schaaf, H. S. *et al.* (2010) 'Tuberculosis at extremes of age', *Respirology*, 15(5), pp. 747–763. doi: 10.1111/j.1440-1843.2010.01784.x.
89. Schaub, S. *et al.* (2004) 'Urine protein profiling with surface-enhanced laser-desorption/ionization time-of-flight mass spectrometry', *Kidney International*, 65(1), pp. 323–332. doi: 10.1111/j.1523-1755.2004.00352.x.
90. Schiffer, E., Mischak, H. and Novak, J. (2006) 'High resolution proteome/peptidome analysis of body fluids by capillary electrophoresis coupled with MS', *Proteomics*, 6(20), pp. 5615–5627. doi: 10.1002/pmic.200600230.
91. Schubert, O. T. *et al.* (2015) 'Building high-quality assay libraries for targeted analysis of SWATH MS data', *Nature Protocols*, 10(3), pp. 426–441. doi: 10.1038/nprot.2015.015.
92. Searle, B. C. *et al.* (2018) 'Comprehensive peptide quantification for data independent acquisition mass spectrometry using chromatogram libraries', *bioRxiv*, 51(1), p. 51. doi: 10.1101/277822v1.
93. Seddon, J. A. and Shingadia, D. (2014) 'Epidemiology and disease burden of tuberculosis in children: A global perspective', *Infection and Drug Resistance*, 7, pp. 153–165. doi: 10.2147/IDR.S45090.
94. Selevsek, N. *et al.* (2015) 'Reproducible and consistent quantification of the *saccharomyces cerevisiae* proteome by SWATH-mass spectrometry', *Molecular and Cellular Proteomics*, 14(3), pp. 739–749. doi: 10.1074/mcp.M113.035550.
95. Shahid, R. *et al.* (2009) 'Comparison of distance measures in spatial analytical modeling for health service planning', *BMC Health Services Research*, 9, pp. 1–14. doi: 10.1186/1472-6963-9-200.
96. Slama, P. *et al.* (2018) 'Robust determination of differential abundance in shotgun proteomics using nonparametric statistics', *Molecular Omics*, 14(6), pp. 424–436. doi: 10.1039/c8mo00077h.
97. South African National Department of Health (2013) 'Guidelines for the Management of Tuberculosis', *Clinical Infectious Diseases*, 24, p. 11-53. Available: <http://www.kznhealth.gov.za/family/National-Childhood-TB-Guidelines-2013>.

98. Spahr, C. *et al.* (2001) 'Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry I. Profiling an unfractionated tryptic digest', *Proteomics*, 1(1), pp. 93–107. doi: 10.1002/1615-9861.
99. Suna, G. *et al.* (2018) 'Extracellular matrix proteomics reveals interplay of aggrecan and aggrecanases in vascular remodeling of stented coronary arteries', *Circulation*, 137(2), pp. 166–183. doi: 10.1161/CIRCULATIONAHA.116.023381.
100. Tabb, D. L. (2012) 'QuaMeter " IDFree " Manual'.
101. Tabb, D. L. *et al.* (2014) 'QC Metrics from CPTAC Raw LC-MS/MS Data Interpreted through Multivariate Statistics', *Analytical Chemistry*, 86(5), pp. 2497–2509. doi: 10.1021/ac4034455.
102. Thongboonkerd, V. (2004) 'Proteomics in nephrology: Current status and future directions', *American Journal of Nephrology*, 24(3), pp. 360–378. doi: 10.1159/000079148.
103. Ting, Y. S. *et al.* (2017) 'PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data', *Nature Methods*, 14(9), pp. 903–908. doi: 10.1038/nmeth.4390.
104. Tong, Y. *et al.* (2019) 'Data-independent acquisition-based quantitative proteomic analysis reveals differences in host immune response of peripheral blood mononuclear cells to sepsis', *Scandinavian Journal of Immunology*. Blackwell Publishing Ltd, 89(4), p. 12. doi: 10.1111/sji.12748.
105. Tsou, C. *et al.* (2015) 'DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics', *Nature Methods*, 12(3), pp. 258–264. doi: 10.1038/nmeth.3255.
106. Tujin, S. *et al.* (2016) 'Advances in targeted proteomics and applications to biomedical research', *Proteomics*, 16(15–16), pp. 2160–2182. doi: 10.1002/pmic.201500449.Advances.
107. Urban, P. L. (2016) 'Quantitative mass spectrometry: An overview', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2079), p. 11. doi: 10.1098/rsta.2015.0382.
108. Veenaas, C., Linusson, A. and Haglund, P. (2018) 'Retention-time prediction in comprehensive two-dimensional gas chromatography to aid identification of unknown contaminants', *Analytical and Bioanalytical Chemistry*. Analytical and Bioanalytical Chemistry, 410(30), pp. 7931–7941. doi: 10.1007/s00216-018-1415-x.
109. Venable, J. D. *et al.* (2004) 'Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra', *Nature Methods*, 1(1), pp. 39–45. doi: 10.1038/nmeth705.
110. Wasinger, V. C., Zeng, M. and Yau, Y. (2013) 'Current Status and Advances

- in Quantitative Proteomic Mass Spectrometry', *International Journal of Proteomics*, 2013, 7(1), pp. 1–12. doi: 10.1155/2013/180605.
111. World Health Organization. (2017) 'Global Tuberculosis Report 2017: Leave no one behind - Unite to end TB', *JAMA*, 727, pp. 1393-1394. doi: 10.1001/jama.2014.11450.
112. World Health Organization. (2013) 'Roadmap for childhood tuberculosis', p. 39. Available: <https://www.who.int/tb/publications/tb-childhoodroadmap/en/>.
113. World Health Organization. (2019) '*Global tuberculosis report*'.
114. World Health Organization Executive Board. (2015) 'Global strategy and targets for tuberculosis prevention, care and control after', 134, pp. 1–23. Available: <https://apps.who.int/iris/handle/10665/172828>.
115. Xu, D. and Xu, Y. (2004) 'Protein databases on the internet.', *Current protocols in molecular biology*, 68(1), p. 19. doi: 10.1002/0471142727.mb1904s68.
116. Yang, F. J. *et al.* (2020) 'Plasma Leucine-Rich α -2-Glycoprotein 1 Predicts Cardiovascular Disease Risk in End-Stage Renal Disease', *Scientific Reports*, 10(1), pp. 1–9. doi: 10.1038/s41598-020-62989-7.
117. Yang, Z. *et al.* (2004) 'Identification of Risk Factors for Extrapulmonary Tuberculosis', *Clinical Infectious Diseases*, 38(2), pp. 199–205. doi: 10.1086/380644.
118. Yates, J. R. *et al.* (1998) 'Method to Compare Collision-Induced Dissociation Spectra of Peptides: Potential for Library Searching and Subtractive Analysis', *Analytical Chemistry*, 70(17), pp. 3557–3565. doi: 10.1021/ac980122y.
119. Zak, D. E. *et al.* (2017) 'A prospective blood RNA signature for tuberculosis disease risk', *The Lancet*, 387(10035), pp. 2312–2322. doi: 10.1016/S0140-6736(15)01316-1.This.
120. Zhang, Y. *et al.* (2015) 'The Use of Variable Q1 Isolation Windows Improves Selectivity in LC-SWATH-MS Acquisition', *Journal of Proteome Research*, 14(10), pp. 4359–4371. doi: 10.1021/acs.jproteome.5b00543.
121. Zheng, Y. (2018) 'Study Design Considerations for Cancer Biomarker Discoveries', *The Journal of Applied Laboratory Medicine*, 3(2), pp. 282–289. doi: 10.1373/jalm.2017.025809.
122. Zi, J. *et al.* (2014) 'Expansion of the ion library for mining SWATH-MS data through fractionation proteomics', *Analytical Chemistry*, 86(15), pp. 7242–7246. doi: 10.1021/ac501828a.