Towards Precision Oncology

Advancing Multiomic Biomarker Discovery with Mass Spectrometry Proteomics

Mosquim Junior, Sergio

2024

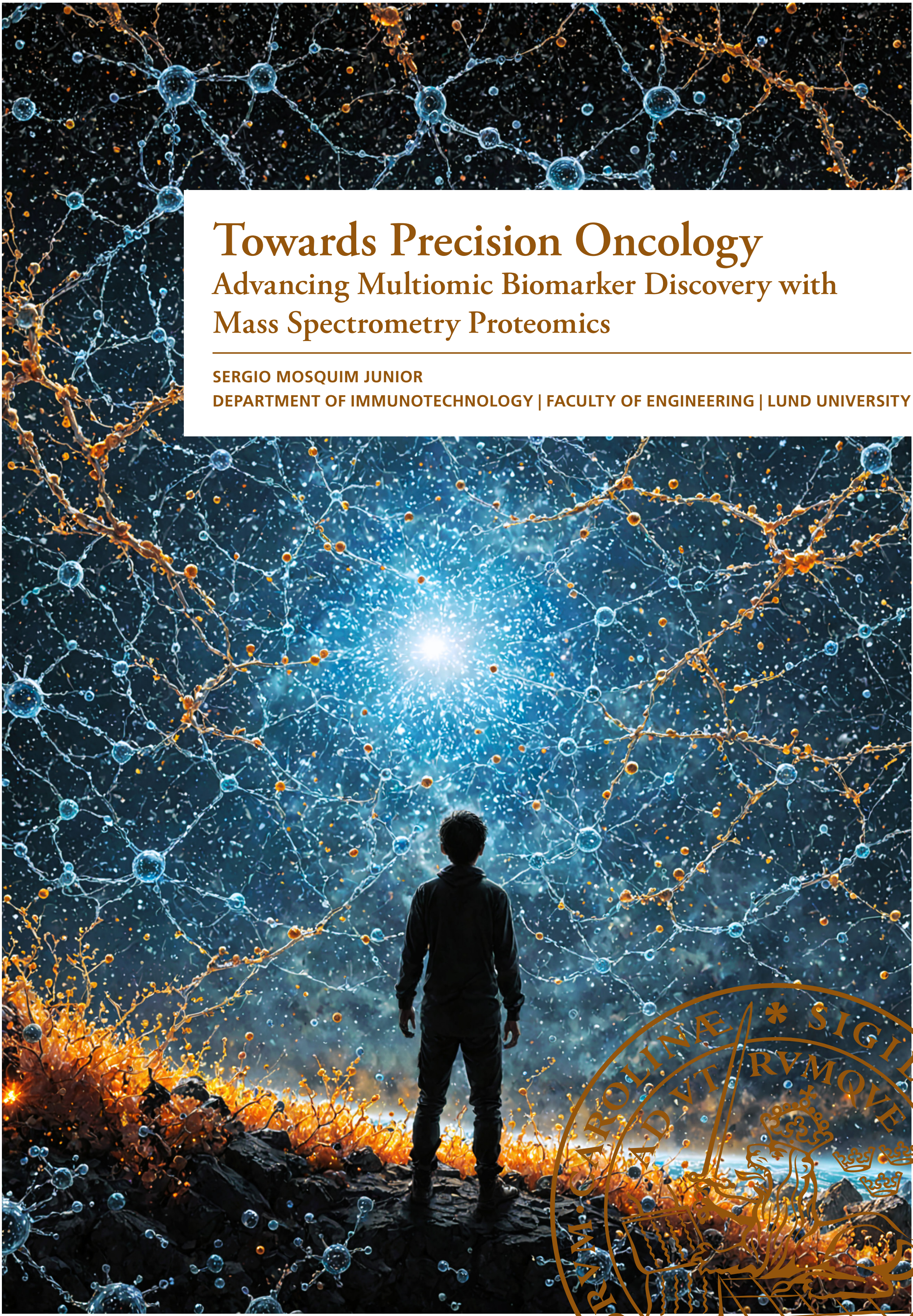
LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Towards Precision Oncology

Advancing Multiomic Biomarker Discovery with Mass Spectrometry Proteomics

SERGIO MOSQUIM JUNIOR
DEPARTMENT OF IMMUNOTECHNOLOGY | FACULTY OF ENGINEERING | LUND UNIVERSITY



Towards Precision Oncology:

Advancing Multiomic Biomarker Discovery with Mass Spectrometry Proteomics

Sergio Mosquim Junior

DOCTORAL DISSERTATION

Doctoral dissertation for the degree of Doctor of Philosophy (PhD) at the Faculty of Engineering at Lund University to be publicly defended on 13th of September at 09.00 in Auditorium, Medicon Village, Scheeletorget 1, 223 63, Lund

Faculty opponent
Prof. Dr. Connie R. Jimenez, Translational OncoProteomics, Department of Medical Oncology, Amsterdam University Medical Center, Amsterdam, The Netherlands

Organisation: LUND UNIVERSITY

Document name: Doctoral Dissertation

Date of issue: 13th September 2024

Author(s): Sergio Mosquim Junior

Sponsoring organisation:

Title and subtitle: Towards Precision Oncology: Advancing Multiomic Biomarker Discovery with Mass Spectrometry Proteomics

Abstract:

Over the past couple of decades, considerable advances in automation and high-throughput omics technologies have contributed to the generation of unprecedented amounts of molecular data, challenging the traditional "one size fits all" approach in favour of personalised medicine, where the individual needs of patients are addressed.

However, the speed at which such molecular data is being acquired does not translate in a proportional number of clinical implementations. One way in which the molecular data can contribute to individualised treatment is by the discovery of novel and more robust biomarkers.

Proteins are interesting biomarker candidates, as they mediate most processes in living organisms, and are by far the most utilised molecules in clinical use. For these reasons, global analysis of proteins could contribute to the development of better biomarkers.

By combining mass spectrometry-based proteomics with automation solutions, workflows that can potentially improve biomarker discovery are discussed and showcased in this thesis. Such workflows are exemplified in (i) the multiplexed enrichment of blood plasma, allowing higher throughput and a unique platform for automation of affinity enrichment, (ii) the acquisition of matching multiomics data from a large number of breast cancer patient samples, allowing the optimisation of data acquisition strategies and utilisation of such data for functional analyses of the intrinsic subtypes, (iii) and the multiomics data analysis of patient samples, contributing to the most comprehensive molecular profile of metastatic processes in oestrogen receptor-positive breast cancer utilising transcriptomics, proteomics, phosphoproteomics and immune infiltration data acquired from the same tumour samples.

In summary, the work presented in this thesis highlights strategies incorporating mass spectrometry-based proteomics and automation, allowing a greater number of samples to be analysed, more data to be extracted from samples and making better use of these data, which would hopefully improve biomarkers discovery, contributing to the field of personalised medicine.

Key words: Multi-Omics, Data Integration, Proteomics, Affinity Enrichment, Phosphoproteomics, Mass Spectrometry, Breast Cancer, Blood Plasma, Deconvolution, Automation

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language: English

ISSN and key title:

ISBN:

978-91-8104-115-6 (print)

978-91-8104-116-3 (electronic)

Recipient's notes

Number of pages: 94

Price

Security classification

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature

Date 2024-07-29

Towards Precision Oncology:

Advancing Multiomic Biomarker Discovery with Mass Spectrometry Proteomics

Sergio Mosquim Junior

LUND
UNIVERSITY

Coverphoto by Sergio Mosquim Junior

Faculty of Engineering

Department of Immunotechnology

ISBN 978-91-8104-115-6 (print)

ISBN 978-91-8104-116-3 (electronic)

Printed in Sweden by Media-Tryck, Lund University

Lund 2024

"The world is indeed full of peril, and in it there are many dark places; but still there is much that is fair, and though in all lands love is now mingled with grief, it grows perhaps the greater."

J. R. R. Tolkien, The Fellowship of the Ring

Table of Contents

Abstract

Over the past couple of decades, considerable advances in automation and high-throughput omics technologies have contributed to the generation of unprecedented amounts of molecular data, challenging the traditional "one size fits all" approach in favour of personalised medicine, where the individual needs of patients are addressed.

However, the speed at which such molecular data is being acquired does not translate in a proportional number of clinical implementations. One way in which the molecular data can contribute to individualised treatment is by the discovery of novel and more robust biomarkers.

Proteins are interesting biomarker candidates, as they mediate most processes in living organisms, and are by far the most utilised molecules in clinical use. For these reasons, global analysis of proteins could contribute to the development of better biomarkers.

By combining mass spectrometry-based proteomics with automation solutions, workflows that can potentially improve biomarker discovery are discussed and showcased in this thesis. Such workflows are exemplified in (i) the multiplexed enrichment of blood plasma, allowing higher throughput and a unique platform for automation of affinity enrichment, (ii) the acquisition of matching multiomics data from a large number of breast cancer patient samples, allowing the optimisation of data acquisition strategies and utilisation of such data for functional analyses of the intrinsic subtypes, (iii) and the multiomics data analysis of patient samples, contributing to the most comprehensive molecular profile of metastatic processes in oestrogen receptor-positive breast cancer utilising transcriptomics, proteomics, phosphoproteomics and immune infiltration data acquired from the same tumour samples.

In summary, the work presented in this thesis highlights strategies incorporating mass spectrometry-based proteomics and automation, allowing a greater number of samples to be analysed, more data to be extracted from samples and making better use of this data, which would hopefully improve biomarkers discovery, contributing to the field of personalised medicine.

Popular Science Summary

Since the second half of the 20th century, we have experienced a transition period marked by the increasing health impact of noncommunicable diseases (NCDs), such as cardiovascular disease, diabetes, and cancer. According to estimates from the World Health Organization (WHO), these diseases account for the vast majority of premature deaths worldwide.

Cancer, for instance, ranks amongst the main culprits. The WHO projects that over 28 million new cases will arise in 2040, representing an overall increase of 47% compared to current estimates. These estimates are the result of a population increase, aging, and increased prevalence of risk factors, but it also accounts for earlier and better detection of the disease. In connection with incidence, cancer mortality rates are also projected to increase, which highlights the need for new and better treatment options, and biomarkers are essential for this process.

A biomarker can be defined as a characteristic which, once measured, can serve as an indicator of a process, normal, pathogenic or in response to treatment. Different measurements can work as biomarkers, including, for instance, molecules or images such as features in X-ray radiographs.

In the recent past, the field of molecular medicine has advanced dramatically to a point where we have unprecedented amounts of data collected. Historically, the classification of diseases was done based on signs and symptoms, often resulting in a therapeutic approach commonly known as “one size fits all”.

The abundance of data, together with recent developments, allows for a different approach, one that addresses the individual needs of patients, also known as personalised medicine.

In the context of molecular medicine, a biomarker can be, for example, a gene or a protein. Proteins make interesting biomarkers because most functions are mediated by proteins. They are also amply used in the medical field for various purposes. The simplest example is a routine blood test, which results in the measure of several proteins found in blood that give an overall indication of an individual’s health status.

The characterisation of the entire protein fraction of a sample at a given time point, including proteins, their isoforms, modifications, and interactions defines the proteome, and the field responsible for its study, proteomics.

Amongst the different methods available in the proteomics toolkit, the use of mass spectrometry has become the method of choice in the analysis of complex samples. Essentially, mass spectrometers are capable of measuring the presence and abundance of molecules based on two main physical properties, mass and charge.

In this thesis, I describe how mass spectrometry can be used with different sample materials in a reproducible way in order to discover cancer biomarkers which can be used in personalised medicine.

We used two sample materials commonly found in the clinical setting, tissue biopsies and blood. The challenges associated with each are different, requiring different strategies.

Starting with blood, it is a very attractive source of potential biomarkers due to the minimally invasive and highly standardised procedure used to collect it. It contains a variety of proteins originated in blood and in other tissues, making it the most comprehensive proteome of human nature. However, the protein abundance also represents the main challenge associated with proteomic preparation of blood samples.

A few proteins, for example albumin, make the bulk of the blood proteome. In biomarker discovery efforts via mass spectrometry-based proteomics, this typically translates into most of the signal originating from these high-abundance protein species, ultimately limiting the utility of this sample material.

Despite this scenario, different tools are available to tackle this problem. In paper II, we use an approach based on the enrichment of low-abundance proteins. By using small antibody fragments, the concentration of low-abundance proteins in relation to the high abundance ones is increased, resulting in higher chances of detecting and quantifying these proteins. These antibody fragments had been previously developed and demonstrated to work in tissue samples. However, given the potential benefit of using blood for these analyses, an automated protocol was created to test a panel of 29 of these antibodies, both independently and in combination.

As a result, we were able to demonstrate not only the benefits of automation, which allows for a more robust sample preparation strategy, but also the use of a small number of antibody fragments for enrichment of low-abundance proteins, a small step which could be easily introduced in a workflow for improved proteome coverage.

The other material we worked with was based on tissue biopsies from breast cancer patients.

Breast cancer (BC) is the leading cause of incidence and mortality among women. Molecular medicine and the mammography screening program have contributed to considerable developments in the field. Nevertheless, a significant number of patients end up being undertreated or overtreated, either receiving insufficient treatment, ultimately leading to relapse, or undergoing procedures which could have been avoided, resulting in unnecessary pain, side effects and associated costs.

Since most of the developments in BC molecular medicine have been reported at the genome or the transcriptome levels, we were interested in investigating whether the proteome could provide additional information. For that, we used samples belonging to a clinical cohort. Since these samples had plenty of clinical and molecular data (transcriptomics) linked to them, they were ideal for exploring the potential benefits of including proteome data.

In paper I, we developed a workflow for the preparation of BC samples for proteomic analysis. Samples representing different subtypes of BC were used. One of the first objectives in developing the workflow was investigating the impact of different mass spectrometry data acquisition strategies.

Data dependent acquisition (DDA) is the most common approach. However, the way it works makes it inherently biased towards higher abundance molecules. To address this bias, a different approach, namely data independent acquisition (DIA), has become increasingly popular in the past few years. By comparing these two strategies in the BC samples, we were able to demonstrate that DIA is better suited for this sample type, allowing for overall increase in the number of detected and quantified proteins.

Given the abundance of clinical and molecular data available from these samples, we were also able to demonstrate that our proteome analysis could contribute to additional information not captured by other data. However, given the small number of samples per subtype and the fact that the samples were selected without a clinical question in mind, we were limited in what clinical conclusions we could draw from this study.

For that reason, paper IV was devised. Consulting with experts in the field, a relevant clinical question was proposed on the effects of lymph node metastasis and distant metastasis in BC, more specifically oestrogen receptor-positive BC. We built on the previously developed workflow and included protein phosphorylation data (phosphoproteomics) in a dataset that is the most comprehensive of its kind. The addition of phosphorylation data may provide additional information to the analysis, as protein phosphorylation is considered an important feature, responsible for activation of proteins and cellular communication. By integrating the already available transcriptome data with proteomics and phosphoproteomics, we were able to find, both at the pathway and single marker level, changes associated with metastasis to lymph nodes or distant sites.

The quality of the data and the clinical question behind it also permitted us to explore the immune component in these samples. The presence of immune cells in a tumour and their importance are not new. In fact, it has important prognostic implications. In the era where an abundance of omics data is available, it has been demonstrated that immune cell infiltration can be estimated based on various transcriptomics data. Until recently, proteomics data has not been used for this purpose.

Building on the knowledge that proteins make ideal biomarkers because they mediate most of an organism's functions, the estimation of immune cells based on proteomics data could also benefit from a similar rationale, where the actual effector molecules and markers are being used for improved accuracy.

In paper III, we explore various approaches to perform this estimation based on proteomics data and demonstrate, in one of the first efforts of its kind, that the choice of method and reference play a fundamental role in the accuracy of the estimations. Although transcriptomics has been typically used for such purposes, our results show promising applications of proteomics data.

In the context of paper IV, the immune infiltration data also contributed towards exploring potential subtypes associated with the presence of inflammation and immune cells, enabling a discussion of alternative therapeutic strategies in such cases.

In summary, in this thesis, the need for better and improved biomarkers is evidenced by shedding a light on the increasing burden of cancer in the coming decades. The discussion then moves towards the use of proteins as biomarkers, given they are more closely linked to functions compared to other molecular entities such as genes and transcripts as well as more amply used in the clinical setting. We then proceed to develop and optimise workflows for the analysis of different clinical samples, culminating in a project where data from multiple sources are combined to aid in answering clinically relevant questions.

Overall, the workflows presented here can contribute to the more ample use and implementation of proteomics as a biomarker discovery tool. Finally, by linking these workflows with a clinical need, we are able to suggest targets worth further exploring, hopefully leading to better patient stratification, and contributing to the field of precision medicine.

List of Papers

Paper I

Mosquim Junior, S., Siino, V., Ryden, L., Vallon-Christersson, J. & Levander, F. Choice of High-Throughput Proteomics Method Affects Data Integration with Transcriptomics and the Potential Use in Biomarker Discovery. Cancers (Basel) 14 (2022). https://doi.org/10.3390/cancers14235761

Paper II

Mosquim Junior, S., Levander, F. "Automated multiplexed affinity-based enrichment of peptides for LC-MS/MS plasma proteomics". Proteomics (Submitted manuscript).

Paper III

Zamore, M., S. Mosquim Junior. S.L. Andree, C. Altunbulakli, M. Lindstedt and F. Levander. "Considerations for immune cell deconvolution using proteomics data". (Manuscript).

Paper IV

Mosquim Junior, S, M. Zamore, J. Vallon-Christersson, L. Rydén and F. Levander. "Multiomic profiling of metastatic potential in oestrogen-receptor positive breast cancer". (Manuscript)

Author's contribution to the papers

Paper I

I was responsible for all practical laboratory work and mass spectrometry data collection. I performed most computational analyses and produced all visual elements. I wrote and edited the manuscript.

Paper II

I performed practical laboratory work and mass spectrometry data collection. I performed all computational analyses and produced all visual elements. I wrote, reviewed, and edited all versions of the manuscript.

Paper III

I supervised and took part in sample preparation, acquired the mass spectrometry data, discussed the computational workflow, and edited the manuscript.

Paper IV

I participated in planning of the study and selected the study samples, performed all practical laboratory work and mass spectrometry data collection. I performed most computational analyses and produced all visual elements. I wrote and edited all versions of the manuscript.

Other Publications

Siino, V., Ali, A., Accardi, G., Aiello, A., Ligotti, M. E., Junior, S. M., Candore, G., Caruso, C., Levander, F., & Vasto, S. (2022). Plasma proteome profiling of healthy individuals across the life span in a Sicilian cohort with long-lived individuals. Aging Cell, 21, e13684. https://doi.org/10.1111/acel.13684

Abbreviations

2D-GE	Two-dimensional Gel Electrophoresis
ACN	Acetonitrile
AFFIRM	Affinity Selected Reaction Monitoring
AIF	All-Ion Fragmentation
BC	Breast Cancer
BEST	Biomarkers, EndpointS and other Tools
BSE	Breast Self-Examination
CCA	Canonical Correlation Analysis
CCD	Central Composite Design
CIMS	Context Independent Motif Specific
CNB	Core Needle Biopsy
COA	Clinical Outcome Assessment
CPTAC	Clinical Proteomic Tumor Analysis Consortium
DDA	Data-Dependent Acquisition
DESI	Desorption Electrospray Ionisation
DIA	Data-Independent Acquisition
DOE	Design Of Experiments
DTT	Dithiothreitol
ELISA	Enzyme-Linked Immunosorbent Assay
ER	Oestrogen Receptor
ESI	Electrospray Ionisation
FASP	Filter-Aided Sample Preparation
FDA	Food and Drug Administration
FDR	False Discovery Rate
FNA	Fine Needle Aspiration
FT	Fourier Transform
FT-ARM	Fourier Transform All Reaction Monitoring

FTMS	Fourier Transform ion cyclotron Mass Spectrometer
GPS	Global Proteome Survey
GSEA	Gene Set Enrichment Analysis
HAP	High Abundance Protein
HDI	Human Development Index
HER2	Human Epidermal Growth Factor Receptor 2
HILIC	Hydrophilic Interaction Chromatography
IAA	Iodoacetamide
IBC	Invasive Breast Cancer
IBC-NST	Invasive Breast Carcinoma of No Special Type
ICD-O	International Classification of Diseases for Oncology
IHC	Immunohistochemistry
ILC	Invasive Lobular Carcinoma
IMAC	Immobilised Metal Ion Affinity Chromatography
ISH	In-Situ Hybridisation
JIVE	Joint and Individual Variation Explained
LAP	Low Abundance Protein
LC	Liquid Chromatography
LTQ	Linear ion Trap Quadrupole
MALDI	Matrix-Assisted Laser Desorption/Ionisation
MOAC	Metal Oxide Affinity Chromatography
MOFA	Multi-Omics Factor Analysis
MRI	Magnetic Resonance Imaging
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
MSI	Mass Spectrometric Imaging
MSIA	Mass Spectrometry ImmunoAssay
MWCO	Molecular Weight Cutoff

NCD	Noncommunicable Disease
NCI	National Cancer Institute
NIH	National Institute of Health
NMR	Nuclear Magnetic Resonance
ORA	Overrepresentation Analysis
P2CID	Parallel Collision-Induced Dissociation
PAC	Protein Aggregation Capture
PAcIFIC	Precursor Acquisition Independent From Ion Count
PAM	Prediction Analysis for Microarrays
PCA	Principal Component Analysis
PEA	Proximity Extension Assay
PLS	Projection to Latent Structures
PR	Progesterone Receptor
PRM	Parallel Reaction Monitoring
PSM	Peptide Spectrum Match
PTM	Post-Translational Modification
qRT-PCR	quantitative Real-Time Polymerase Chain Reaction
ROR	Risk Of Recurrence
RPPA	Reverse-Phase Protein Arrays
RSM	Response Surface Methodology
scFv	Single Chaing Variable Fragment
SDS	Sodium Dodecyl Sulphate
SDS-PAGE	Sodium Dodecyl Sulphate-Polyacrilamide Gel Electrophoresis
SISCAPA	Stable Isotope Standards and Capture by Anti-Peptide Antibodies
SPE	Solid-Phase Extraction
SRM	Selected Reaction Monitoring
SWATH	Sequential Windowed Acquisition of All Theoretical Mass Spectra
TCGA	The Cancer Genome Atlas

TFA	Trifluoroacetic Acid
TILs	Tumour-Infiltrating Lymphocytes
TNBC	Triple Negative Breast Cancer
TOF	Time-Of-Flight
U.S.	United States
UN	United Nations
WHO	World Health Organization
XDIA	Extended Data-Independent Acquisition

Introduction

The 21st century is marked by the ever-increasing socioeconomical impact of noncommunicable diseases such as cardiovascular disease and cancer. It is also marked by advances in molecular medicine, which has contributed to unprecedented amounts of omics data available for a multitude of different conditions, including cancer.

Such developments have also paved the way to a different approach when it comes to medical care, one that is not based on shared signs and symptoms, but rather addresses individual needs of patients, known as personalised medicine.

However, the capacity to translate these data and findings into clinical applications has been limited. For instance, very few biomarkers end up being approved for clinical use compared to the number of candidates published. This highlights a need for bridging the gap between laboratory research and clinical research, hopefully leading to the discovery and translation of better biomarkers for use in personalised medicine applications.

In the present thesis, different approaches are utilised and discussed with the aim of enabling better biomarker discovery in precision oncology.

First, although proteins are by far the most utilised class of molecules in clinical use, proteomics has not been amply utilised for biomarker discovery. This is partially due to most developments being focused on genomic and transcriptomic approaches, historically. In this context, a first focus is put on optimising and automating workflows for proteomic (and multiomics) data acquisition, allowing larger cohorts to be analysed, thereby increasing the chances of discovering more robust biomarkers. These aspects are discussed in Papers I and II, in the context of a large breast cancer clinical cohort and blood plasma, respectively.

A second focus, enabled by the developments achieved in the first part, is the integration of omics data for improved biomarker discovery. The rationale behind it is that multiple omics can potentially allow for a more complete molecular profile to be drawn, which would be beneficial for understanding and addressing complex issues such as cancer. Having established that proteomics can add a layer of information complementary to transcriptomics (Paper I), in Paper IV, data integration is implemented with the goal of profiling oestrogen receptor-positive breast cancer patient samples utilising transcriptomics, proteomics, phosphoproteomics and immune infiltration data.

The immune component has been shown to be important for patient classification and prognosis. Among different techniques available, deconvolution of immune cell types using bulk transcriptomics data has been shown to be successful. However, considering that these cells are usually characterised by the presence of cell surface proteins, the use of proteomics data for deconvolution could be beneficial. This approach is discussed and implemented in Paper III, and the resulting immune infiltration estimates are further used in Paper IV for the identification of subtypes.

In summary, this thesis is divided into four parts: (i) bringing attention to the need for better biomarkers for personalised medicine and the potential burden of cancer; (ii) reviewing and discussing implementations of mass spectrometry-based proteomics for biomarker discovery; (iii) reviewing and discussing the use of multiomics data; (iv) and discussing the applications adopted across the different papers in the context of breast cancer.

The Biomarker Problem

The term "biomarker" is very commonly used today. However, the concept of a biochemical or biological marker is much older (*1*). Over the decades, several definitions were proposed. For instance, in 2001, the Biomarkers Definitions Working Group defined a biomarker as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" (*2*). This definition is somewhat problematic in that it does not encompass non-pharmacological interventions nor observations which are not completely objective, such as imaging (*1, 2*).

In 2016, in a joint effort between the U.S. Food and Drug Administration (FDA) and the National Institute of Health (NIH), the term was further refined in the Biomarkers, EndpointS and other Tools (BEST) resource, a publicly available and continuously updated document aimed at standardising common definitions (*3, 4*).

Despite the broad definition, the BEST resource explicitly states that a biomarker is different from a measure of how a patient feels, functions, or survives (*3, 4*). Such measures are known as Clinical Outcome Assessment (COA) (*4*). It is important to distinguish between these two terms given that COAs are measures that are directly important to the patient, while biomarkers can have different purposes. An added layer of complexity is introduced in the context of defining endpoints. Both biomarkers and COAs may be used to for such purpose, and the distinction lies in the level of precision and scientific rigor required. This is necessary to guarantee reliability and reproducibility in a clinical study (*4*).

Compared to clinical endpoints, biomarkers have the advantage of being more easily and quickly measured, ultimately allowing clinical trials to be performed for shorter periods of time and with fewer patients. For example, a hypothetical trial designed to assess the effect of a certain intervention would require fewer patients and less time if using measures such as a patient's blood pressure and echocardiography compared to waiting for a clinical endpoint such as deaths from strokes (*1*).

They are routinely used in acute care settings and have been successfully implemented in other clinical applications such as risk stratification, population screening, disease subtyping and monitoring response to treatment. By functioning as surrogate markers, they also allow for the development of new drugs (*5*).

With the advent of molecular medicine and the rapid technological advances, the medical field is drastically changing given the number of measurements, computations and analyses that are being produced (*4, 6*). The increase in the amount of molecular data routinely collected has directly contributed to a better understanding of complex systems such as disease states. As a consequence of such development, it is unusual for a single biomarker to be able to recapitulate all required information for monitoring an intervention (*1, 4, 6*). Although that is perceived as a negative factor, given that, historically, the field has been built on single biomarker measures, e.g., high systolic blood pressure associated with increased risk of stroke, the use of complex biomarkers, or biomarker panels may indeed enable better predictions, as each biomarker would contribute to the outcome of interest (*4*).

The classification of diseases such as cancer is still heavily based on signs and symptoms. This results in dissonance with molecular medicine, as diseases with different molecular subtypes may be classified as a single entity (*6*). This more traditional approach is often referred to as "one size fits all", and its consequences lead to a paradoxical scenario of both undertreatment and overtreatment (*7*).

The emergence of a field that is accelerating rapidly emphasises the need for a different approach, one in which treatments target the needs of individual patients on the basis of genetic, biomarker, phenotypic, or psychosocial characteristics that distinguish such patient from others with similar clinical manifestations. This approach is known as precision medicine (*7, 8*). Compared to the more traditional approach, precision medicine could contribute not only to more efficient treatments with fewer side effect, but also reduce the associated costs (*7*).

In an era defined by the abundance of omics, informatics and imaging data, the necessity of precision medicine for better disease classification, prognostic and treatment implications becomes evident (*8*). The complexity associated with this system is, however, its biggest challenge, and for that, better biomarkers are needed to further aid in clinical decisions (*8*).

The number of biomarker candidates published has never been higher, with hundreds-to-thousands of them being routinely identified (*1, 5*). However, the problem is that there is a tremendous gap between the number of candidates and the number of approved biomarkers, with most of the successes in precision medicine being in the field of oncology (*5, 9, 10*). The challenges found in the progression of precision medicine are both scientific and non-scientific. Notably, the lack of clearly defined and relevant clinical questions, failure to understand the relationship between the pathophysiology of the disease and the mechanism of action behind the intervention, and failure to meet the high bar of clinical validation all contribute to this scenario (*1,* *5, 9, 10*). In this thesis, some of the scientific aspects are addressed in terms of automation, study design and data analysis for the discovery new biomarkers for applications in precision medicine.

The Burden of Cancer

Based on recent estimates from the World Health Organization (WHO), noncommunicable diseases (NCDs) are responsible for more than three quarters of premature deaths, i.e., occurring between the ages of 30 and 70 years (*11, 12*). The two main culprits are cardiovascular disease and cancer. Out of the 183 surveyed countries, cancer was the first or second leading cause of death in 127 (*12*). This takes place during a period of epidemiologic transition which initiated in the second half of the 20th century, progressively asserting the dominance of NCDs over infectious diseases (*12*). The burden of NCDs varies according to the Human Development Index (HDI), where countries with high HDI are characterised by having cancer as the main NCD, while low-HDI countries have a double burden from NCDs and infectious diseases (*12*).

According to the GLOBOCAN 2020, there were estimated 19.1 million cases of cancer worldwide and 10 million cancer deaths (*11*). In comparison, the GLOBOCAN 2022 estimated 19.9 million new cancer cases and 9.7 million cancer deaths (*13*). Based on estimates of incidence and mortality, the top 10 cancer types, in both sexes combined, account for over 60% of all newly diagnosed cancer cases and more than 70% of the associated deaths. In terms of incidence, female breast cancer (11.7% of total cases), lung (11.4%), colorectal (10.0%), prostate (7.3%) and stomach (5.6%) are the most commonly diagnosed types (*11*). Ranking the cases based on mortality, however, results in lung being the leading cause of cancer-related deaths (18.0% of all cancer deaths), followed by colorectal (9.4%), liver (8.3%), stomach (7.7%) and female breast (6.9%) (*11*).

When further stratified based on sex, the 2020 estimates show around 10.1 million cancer cases and 5.5 million cancer deaths in men; in women, those numbers are 9.2 million and 4.4 million, respectively. In terms of incidence in men, lung (14.3% of all cases) is closely followed by prostate (14.1%), and although lung remains the leading cause of cancer death in men, prostate ranks 5th (*11*). For women, breast is the leading cause of incidence and mortality, corresponding to 24.5% of all diagnosed cases and 15.5% of cancer deaths, followed by colorectal and lung for incidence, and the opposite for mortality (*11*).

For 2022, the estimates changed slightly. Lung cancer ranked first in terms of incidence for both sexes (12.4%), followed by female breast cancer (11.6%), colorectal, prostate and stomach. In number of deaths, lung continues to be the leading cause of cancer-related deaths (18.7%), followed by colorectal, liver, female breast and stomach (*13*). For numbers in incidence and mortality by sex, the trends remained the same for men, while for women, breast continues to rank first, although lung and colorectal now rank second and third both in the number of new cases and deaths (*13*).

Cancer incidence and death rates also have different profiles globally. The report demonstrated that if incidence and mortality rates are split based on a 4-tier Human Development Index (HDI) level, an increase in HDI leads to an increase in both metrics. More specifically, the difference in incidence was 2- to 3-fold higher in very high HID countries compared to low HDI ones. In terms of mortality, the difference is 2-fold in the same direction (*11*). When further subdividing based on sex, the overall cancer incidence in men was 19% higher compared to women. Among men, however, this number varied drastically, with incidence rates ranging almost 5-fold when comparing Australia/New Zealand (494.2 per 100,000) to Western Africa (100.6 per 100,000). Among women, the differences were up to 4-fold comparing Australia/New Zealand to South Central Asia (*11*).

These numbers reflect not only differences in exposure to risk factors, but also increased detection rates for the disease (*11*). For instance, in female breast cancer, incidence rates are 88% higher in transitioned countries compared to transitioning ones. However, the mortality rate in transitioning countries is 17% higher compared to that of transitioned countries (*11*).

Projecting these numbers to 2040, the GLOBOCAN 2020 estimates that 28.4 million new cases will arise in 2040, representing an increase of 47% compared to the current estimates. This is a result of population growth and aging, as well as increased prevalence of risk factors, and the projections are even more striking for low and medium HDI countries, 95% and 64%, respectively. The period of epidemiologic transition mentioned previously results in countries with emerging HDI levels to be the most affected, as the prevalence of risk factors often associated with high-income western countries increases together with a paradigm shift from infection-related and poverty-related cancers to cancers that are more common in developed countries. Ultimately, this requires a change in priorities in national cancer control strategies (*11*).

The increase in incidence will also lead to an increase in mortality rates. For that reason, resources must be allocated for both treatment and management of the disease. Even if prevention is a very effective way of controlling cancer, the implementation of effective interventions into health plans as well as the development of new interventions are necessary (*11*).

Proteomics

By now, it is clear that the burden of cancer will continue to increase in the coming years with an aging population and change in exposure patterns to risk factors. It is also clear that despite the considerable number of candidate biomarkers published, those that get implemented into clinical practice are very few (*5, 9, 14*).

In the realm of molecular medicine, a biomarker can take many forms. For example, they can be DNA, mRNA, proteins, metabolites, or pathways. Proteins are much more closely related to the phenotype than DNA and mRNA, as they represent the endpoint of gene expression and are responsible for most catalytic and structural functions in living organisms, as well as signalling and even gene expression (*15, 16*). They are also the most affected domain during the processes of disease, response, and recovery, and are well established in different clinical applications, including diagnosis, prediction of risk and detection of disease recurrence (*14, 16*). Therefore, the global analysis of proteins should be ideal for the discovery of new and better biomarkers.

The term proteome was first used in 1995 to describe a set of proteins originating from a genome (*17, 18*). Proteomics, on the other hand, corresponds to the study of the proteome. It englobes not only the proteins, but also their isoforms, modifications, interactions and almost everything "post-genomic" (*17*), and is aimed at characterising the entire protein fraction of a given sample at a given time point (*19*).

In broad terms, proteomics can be divided in three distinct types, namely expression proteomics, functional proteomics and structural proteomics (*20*). Expression proteomics is aimed at the study of the expression of proteins, both quantitatively and qualitatively. It includes techniques such as two-dimensional gel electrophoresis (2D-GE) and mass spectrometry (MS)-based technologies. Structural proteomics has the objective of understanding protein interaction and function through a structural basis; techniques such as X-ray crystallography and nuclear magnetic resonance (NMR) are utilised at the protein level together with electron microscopy and electron tomography for visualisation of complexes and cellular context (*17, 20*). Finally, the goal of functional proteomics is the study of protein function and molecular mechanisms (*20*).

The work presented in this thesis focuses on applications of MS-based proteomics. For the sake of simplicity, unless otherwise specified, the term proteomics will be used as synonym for MS-based proteomics.

Mass Spectrometry-based Proteomics

Mass spectrometers originated in 1912 and have undergone continuous development ever since (*21*). Their application in the field of proteomics has become the method of choice in the analysis of complex proteins samples, and it is due to the developments achieved through other omics, i.e., availability of sequence databases for genes and genomes, as well as technological advances in other areas, especially the development of protein ionisation methods, which awarded John B. Fenn and Koichi Tanaka the Nobel Prize in chemistry in 2002 (*17, 21, 22*).

Essentially, all mass spectrometers work by measuring the presence and abundance of molecules based on mass and net charge, more specifically the mass-to-charge (m/z) ratio, two fundamental properties of molecules (*21, 22*).

A mass spectrometer essentially consists of a source, responsible for the ionisation, a mass analyser that measures the m/z ratios of the resulting ions, and a detector, responsible for registering the number of ions at each m/z value (*22*).

Ionisation Methods

Starting at the source, because measurements take place in gas phase, ionisation is an essential step in converting proteins or peptides into ions (*21, 22*).

Two techniques are most commonly used for ionising proteins and peptides for mass spectrometry analysis, namely Electrospray Ionisation (ESI) and Matrix-Assisted Laser Desorption/Ionisation (MALDI) (*22*). They are referred to as soft ionisation techniques in that analyte fragmentation is prevented or minimised (*23*).

In ESI, the ionisation takes place directly from a liquid sample, allowing the system to be readily coupled to liquid-based separation tools such as liquid chromatography (*22, 23*). In principle, the method works by having a liquid sample containing the molecule of interest pumped at low flow rates through a hypodermic needle with high voltage applied to it. The high voltage results in the dispersion of the sample into small droplets, an electrospray, which occurs at atmospheric pressure. The droplets quickly evaporate, imparting charge onto the molecules present (*24*).

In MALDI, the sample is sublimated and ionised out of a matrix via laser pulses (*22*). More specifically, the analyte is coprecipitated with an excess of the matrix and allowed to dry onto a metal substrate. Nanosecond laser pulses, often nitrogen lasers at wavelength 337nm, irradiate the resulting solid (*24*). It also tends to produce singly charged ions, making the resulting spectra easy to interpret. However, the ability to couple ESI-MS with on-line liquid chromatography, thereby enabling simultaneous sample cleanup, concentration, and separation, has made this choice popular in the analysis of complex protein samples (*23, 24*).

Mass analysers and instrument configurations

The second component in the system is the mass analyser, and it is central to the technology. In the field of proteomics-based MS, sensitivity, resolution, mass accuracy and the ability to generate tandem mass spectra or MS/MS spectra, necessary for peptide sequencing, are key parameters (*22, 24*). As mentioned previously, all mass spectrometers measure m/z ratios of analytes. Based on the principle, three different approaches can be used to achieve mass separation: separation on the basis of time-of-flight (TOF MS); separation by quadrupole electric fields generated by metal rods (quadrupole MS); separation by selective ejection of ions from a three-dimensional trapping field, e.g., ion trap MS or Fourier transform ion cyclotron MS (FTMS) (*22, 24*). It is important to note that each has their own set of advantages and disadvantages, but they can also be used in combination in order to harness the strengths of each type (*22*). A representation of different ion sources and instrument configurations is given in **Figure 1**.

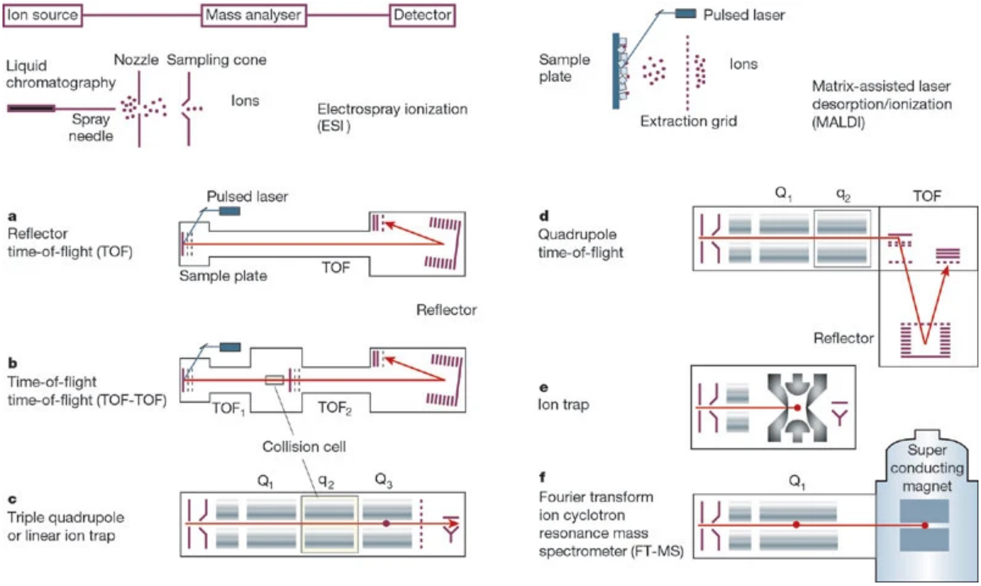

Ion source
Mass analyser
Detector
Liquid chromatography
Nozzle
Sampling cone
Spray needle
Ions
Electrospray ionization (ESI)
Pulsed laser
Sample plate
Ions
Extraction grid
Matrix-assisted laser desorption/ionization (MALDI)
a
Reflector time-of-flight (TOF)
Pulsed laser
Sample plate
TOF
Reflector
b
Time-of-flight time-of-flight (TOF-TOF)
TOF1
TOF2
Collision cell
c
Triple quadrupole or linear ion trap
Q1
q2
Q3
d
Quadrupole time-of-flight
Q1
q2
TOF
Reflector
e
Ion trap
f
Fourier transform ion cyclotron resonance mass spectrometer (FT-MS)
Q1
Super conducting magnet

Figure 1: The upper panels showcase sample ionisation and introduction in ESI and MALDI. Different instrument configurations are illustrated in panels (a-f). From (*22*) with permission from the publisher.

The principle behind TOF analysers is quite simple. It measures the m/z ratio of ions by determining the time required for them to traverse a flight tube, which then reflects the ions back through the tube and into a detector. Small ions have higher velocity and as a result are detected earlier than heavier ions, generating a TOF spectrum (*24, 25*). TOF analysers are normally coupled to MALDI for the analysis of intact peptides, while ESI is often used in conjunction with quadrupoles and ion traps in the analysis of fragment ions. However, instruments such as the TOF-TOF in which a collision cell is introduced between two TOF analysers, and the qTOF in which the quadrupole acts as a filter to select ions for fragmentation and then TOF analysis are also available (*22*), see **Figure 1**.

Quadrupole mass analysers are one of the most common mass analysers. Basically, the quadrupole, an array of four parallel metal rods, creates an electric field which can either be used to transmit all ions or act as a filter to only allow the transmission of ions of a certain m/z ratio (*24, 25*). Most commonly, quadrupoles are coupled to ESI instruments, for instance in a triple quadrupole mass spectrometer, where two quadrupoles act as filters, while a central one contains ions during fragmentation. A different configuration also involves replacing the third quadrupole by a TOF analyser (qTOF), allowing for high accuracy and resolution typically found in TOF instruments (*24*).

Ion trap analysers trap analytes in a three-dimensional field. The ions are first captured and then subjected to single or tandem MS analysis. The number of ions that can be trapped is determined by the ion trap space charge, and it corresponds to the maximum number of ions that can be introduced without a distortion in the applied field (*22, 24, 25*). Ion traps are compact, robust, sensitive, and relatively inexpensive, and so a considerable amount of the reported proteomics data in the literature comes from this type of instrument (*22, 24, 25*).

Fourier transform ion cyclotron MS (FTMS) is a variation of ion trapping, in which the ions are captured in a combination of electric fields and a strong magnetic field. Sensitivity, mass accuracy, resolution and dynamic range can be outstanding, but the complexity of the instrument together with low peptide fragmentation efficiency has limited their use (*22, 24, 26*).

The high complexity and running costs of FT instruments, together with the relatively low space charge capacity made evident the need for other ion trapping approaches of comparable performance, but more compatible with a typical laboratory environment (*26, 27*). With that in mind, the orbitrap mass analyser made its debut as a proof-of-concept device in the year 2000 (*27*). This analyser is based on the principle of orbital trapping, first implemented in 1923, and consists of two electrodes shaped to create a quadro-logarithmic electrostatic potential. The outer electrode is barrel shaped, while the inner one has a spindle-like shape. Injected ions rotate around the inner electrode and oscillate harmonically along its axis with a frequency characteristic of their m/z ratios. These oscillations are then converted into frequency spectra using a Fourier transform similar to the one used in FTMS instruments (*26, 27*). Compared to other electrostatic traps such as linear and segmented ring, the orbitrap offers the best performance with the smallest dimensions, making it the most suitable alternative to traditional FTMS instruments (*26, 27*).

The orbitrap was demonstrated to have high resolving power, internal mass accuracy and high space charge capacity when using pulsed ion sources. This, in turn, created a problem with sources which produced continuous ion beams such as ESI, which was later resolved by introducing a linear rf-only quadrupole that acted as ion storage (*26*). Initially, the injection of ions into the orbitrap was performed axially, which ultimately limited the mass range, transmission, dynamic range and mass accuracy over a wide mass range (*26*). The orthogonal injection of the ions using short pulses alleviated these issues and paved the way to an instrument configuration which combined the tandem MS capabilities of a linear ion trap (LTQ) instrument with the high resolution and mass accuracy of the orbitrap (*26*).

Despite the improved sensitivity of Orbitrap mass analysers, such instruments are often limited by the speed at which MS/MS spectra can be acquired, limiting ultra-high-throughput applications. TOF analysers, on the other hand, routinely generate spectra at much higher rates, but they have historically suffered from poor sensitivity (*28*). During the writing of this thesis, the new Astral mass analyser was introduced, addressing the limitations of both other approaches and combining features from quadrupole, ion trap and orbitrap into a hybrid instrumentation, the Thermo Scientific™ Orbitrap Astral Mass Spectrometer (**Figure 2**). It could potentially offer improved resolution and acquisition rate, thereby allowing higher throughput and less variation caused by instrument stability over time. These factors would, in theory, benefit biomarker discovery efforts.

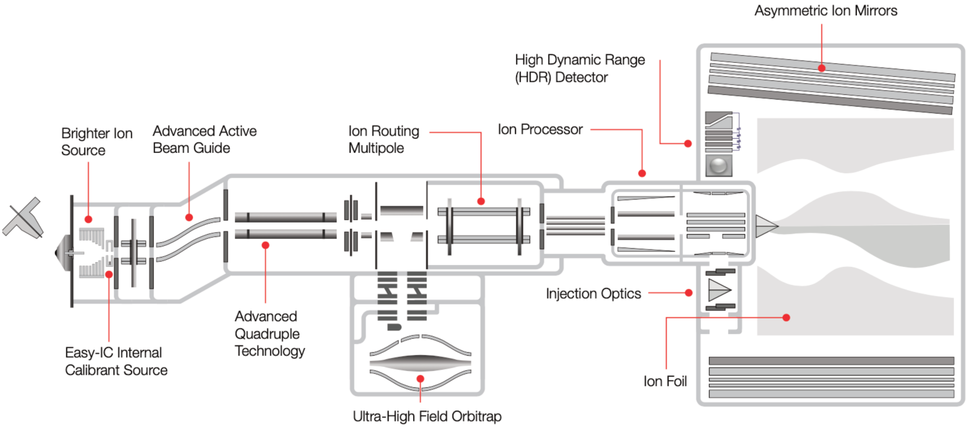

Asymmetric Ion Mirrors
High Dynamic Range (HDR) Detector
Brighter Ion Source
Advanced Active Beam Guide
Ion Routing Multipole
Ion Processor
Injection Optics
Easy-IC Internal Calibrant Source
Advanced Quadruple Technology
Ion Foil
Ultra-High Field Orbitrap

Figure 2 Schematic of ion path on Orbitrap Astral Mass Spectrometer. Image courtesy of Thermo Fisher Scientific. Reproduced from (*29*).

The data presented in Papers I-IV were acquired on a Thermo Scientific™ Q Exactive HF-X Hybrid Quadrupole-Orbitrap Mass Spectrometer instrument (**Figure 3**), a further development from the hybrid configuration initiated with the LTQ Orbitrap instrument and includes a quadrupole but no linear ion trap.

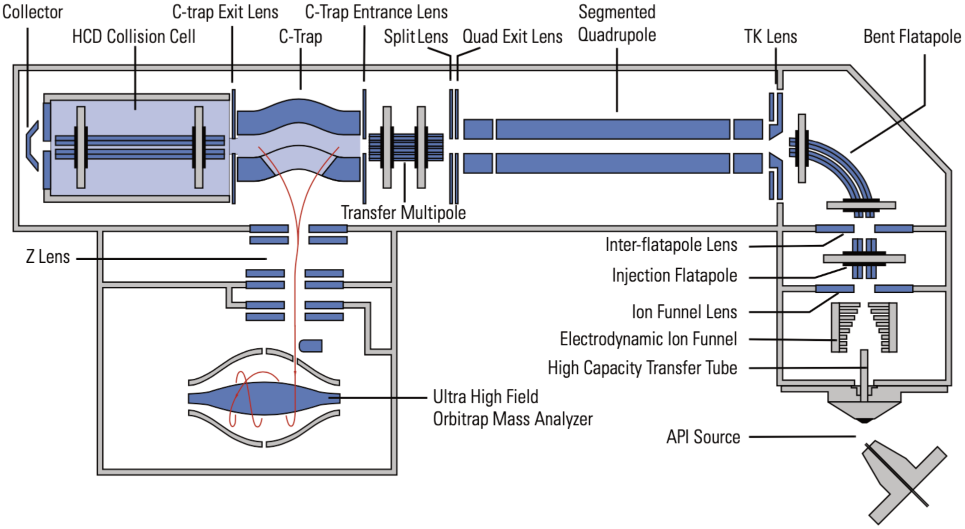


Figure 3 Schematic of the Q Exactive HF-X Mass Spectrometer. Image courtesy of Thermo Fisher Scientific. Reproduced from (*30*).

Top-down versus bottom-up MS/MS

As the field developed, mass-spectrometry-based proteomics became the method of choice for identifying and quantifying not only proteins, but also their interactions and post-translational modifications (PTMs) (*31, 32*). There are three approaches used in the analysis of MS-based proteomics data, namely top-down, middle-down and bottom-up (*31-33*), illustrated in **Figure 4**.

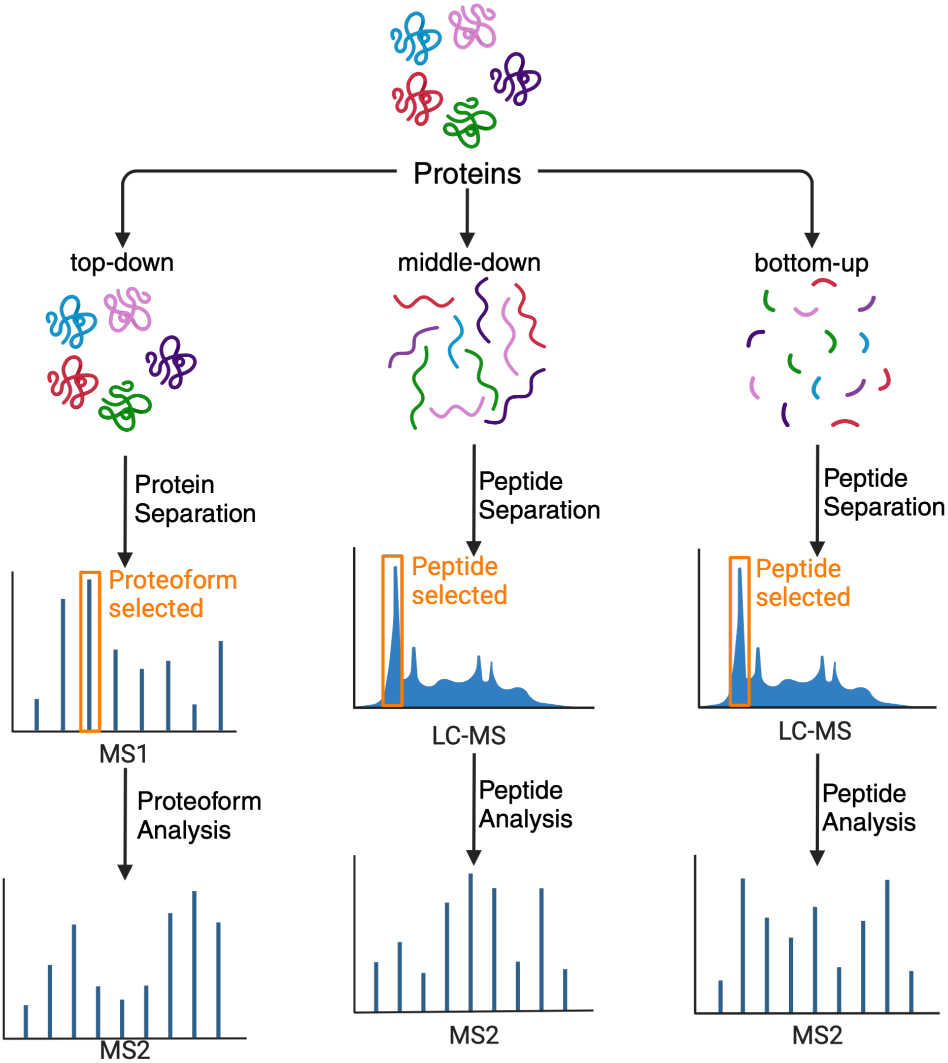


Figure 4 Schematic representation of the most common MS-based proteomics workflows. Created with BioRender.com.

In top-down proteomics, the analysis of intact proteins is performed. This approach combines ESI with high-performance MS, e.g., FTMS, given that ESI impairs more charge on the molecule compared to MALDI and FTMS instruments have high resolution (*31, 34-36*). This method is typically applied to samples of lower-complexity, in that a protein target or complex has already been predetermined and isolated, and it can be used for detecting modifications, investigating functional relationships between different PTMs on the same protein, identifying and quantifying different isoforms that would have been convoluted from endoproteinase digestion, and characterising drug-target interactions (*34, 37*). Moreover, in terms of sample preparation, it is also subjected to fewer artifacts compared to the bottom-up approach (*34*).

Middle-down refers to a workflow which makes use of a partial digestion step for the generation of large peptides (*33*). However, the most commonly used approach in protein analysis is the bottom-up approach, also known as shotgun proteomics (*32*). This method is based on a peptide-to-protein logic similar to that of the middle-down approach, i.e., proteins from a complex mixture first undergo proteolytic digestion prior to analysis by liquid chromatography (LC) tandem mass spectrometry (MS/MS) (*19, 32*). A process of protein inference is then performed in order to match the fragment peptide sequence with the protein from which it originated. Although the method is less direct than the top-down approach, it was largely adopted by the community (*19*).

Despite widespread adoption, the use of bottom-up proteomics is not met without challenges. In the biomarker discovery phase, it is not uncommon for the goal to be the identification of as many differentially abundant candidates as possible while only using a limited number of samples, which results is challenges relating to the variability of protein concentrations amongst the different samples and their complexity. A second challenge appears in the analysis of samples from very heterogenous populations, as mutations, polymorphisms, RNA processing and PTMs can give rise to different proteoforms (*19, 34*). The term "proteoform" was suggested in 2013 by Smith, Kelleher and The Consortium for Top Down Proteomics to refer to all possible molecular forms that proteins can possess, originating from genetic variations, alternatively spliced RNA transcripts and post-translational modifications (*38, 39*).

On a more technical side, the proteolytic digestion of proteins results in large amounts of data that need to be processed in order to match spectra to the original peptides. Moreover, the goal is often to retrieve the information from the original proteins and be able to quantify them. Because individual peptides might be present in different proteins, the process of assembling peptide sequences to infer the protein content of a sample becomes complex, and this is referred to as the protein inference problem. A similar logic applies when attempting to quantify a protein in a sample. As peptides might be present in different proteins, the correct assigning of a peptide to a protein directly influences the quantification estimates of said protein (*40, 41*). Top-down proteomic approaches could ameliorate the situation, as the analysis of whole proteins would improve PTM identification which could be lost in other approaches. However, protein separation, solubility and complexity are still challenges that resulted in this approach not being more popular (*20, 34, 39*).

Data acquisition strategies

As mentioned previously, bottom-up proteomics is the most widespread proteomic workflow. It is also the one used in all articles presented in this thesis, and for that reason it will be the main focus.

When moving into the realm of data acquisition strategies for bottom-up workflows, three distinct approaches are available, namely data-dependent acquisition (DDA), data-independent acquisition (DIA) and targeted acquisition (*19, 42*).

In targeted proteomics, as the name implies, the targets are predetermined and known. Selected Reaction Monitoring (SRM) is the most used targeted acquisition method. It requires prior knowledge about the peptide and the resulting fragment ions (transitions), which are included in a list and passed onto the acquisition method. Rather than discovering peptides and proteins in a sample, the goal of targeted approaches is to quantify and validate the presence of the target peptides (*19, 42*). Advances in instrumentation led to the development of a different targeted acquisition strategy, namely Parallel Reaction Monitoring (PRM). PRM takes advantage of the orbitrap (or TOF) higher mass accuracy and resolution to codetect all fragments from a peptide ion, thereby eliminating the need for extensive selection and optimisation of the method to the same extent as in SRM (*19, 42-44*).

The most widely used strategy to query the proteome in a discovery manner is DDA. In it, the instrument operates in cycles of MS1 and MS2 scans. All precursor ions at a given chromatographic time point are simultaneously scanned at the MS1 level. The instrument then selects the "Top N" most abundant precursor ions and sends them for fragmentation followed by MS2 scan (*19, 32, 42, 45*). MS1 and MS2 spectra are then used to query a database in order to identify the corresponding proteins, and quantification can be performed for instance by taking the area under the MS1 peak of precursor intensities by a process called label-free quantification (LFQ) (*45*). In the analysis of complex samples, many precursor ions will be injected at the same time, often surpassing the instrument sequencing capacity. For that reason, a selection criterion needs to be imposed. The most abundant precursors are selected as a way of increasing the likelihood of successful identifications (*19, 32*). Due to the imposed selection of precursors for fragmentation, variations in chromatographic performance result in a stochastic selection of precursors, ultimately resulting in poor run-to-run reproducibility even if the same sample is reinjected (*19, 45*).

Data-independent acquisition (DIA) was an alternative strategy created to address these limitations and combine the strengths of both DDA and targeted approaches. The instrument operates continuously acquiring MS2 spectra, similarly to what happens in targeted acquisition. Differently from these other two approaches, it does not assume the detection of a precursor (MS1 scan in DDA) for triggering MS2 acquisition. In that way, the acquisition is said to be independent of prior knowledge (*19, 45, 46*).

This can be accomplished by either fragmenting all ions injected at a given chromatographic timepoint, called broadband DIA, or by dividing the m/z range into smaller separate m/z isolation windows that are analysed consecutively (*46*). Within broadband DIA, several methods are available. A few examples include Shotgun Collision-Induced Dissociation (*47*), MS^E (*48*), parallel collision-induced dissociation (p2CID) (*49*), and All-Ion Fragmentation (AIF) (*50*). On the other hand, methods which utilise a select m/z window for fragmentation include, for instance, the original DIA (*51*), Precursor Acquisition Independent From Ion Count (PAcIFIC) (*52*), extended DIA (XDIA) (*53*), Sequential Windowed Acquisition of all Theoretical Mass Spectra (SWATH) (*54*) and FT-All Reaction Monitoring (FT-ARM) (*55*). Even though DIA strategies were developed in the early 2000s, the work by Gillet et. al. (*54, 56*) was crucial to its popularisation (*45*).

Analysis of Posttranslational Modifications

As it was mentioned in the introduction of this chapter, proteins are responsible for most catalytic and structural functions in any living organism. They are also the class of molecules most affected in the process of disease or other interventions. The close relation to the phenotype is the main reason behind the application of proteomics (*14-16*).

Behind the functionality of proteins lies the complexity of the proteome. Compared to the genome, the proteome is orders of magnitude more complex. This complexity is the result of two processes. The first process takes place at the transcriptional level and is the result of mRNA splicing and tissue-specific alternative splicing. The second, however, takes place after mRNA is translated into proteins (*57*).

Posttranslational modifications (PTMs) are covalent modifications that change the properties of a protein either by addition of a chemical group or cleavage of the peptide backbone (*57, 58*). Many were discovered serendipitously, and it was thanks to the developments in MS that it is now the method of choice for detection and identification of said modifications (*23, 58*).

Although the same principles behind MS also apply to the identification of PTMs, the task is increasingly more complex. Besides the low stoichiometry of some PTMs, the peptide containing the modification must remain stable during sample preparation and ionisation (*23*). The most common PTMs that occur by addition are phosphorylation, acylation, alkylation, glycosylation, and oxidation (*57*).

There are over 200 different types of protein modifications. Out of these, phosphorylation is probably the most well characterised and understood, both in terms of processes regulating it, but also its functional consequences (*23*). As the phosphoproteome is studied in Paper IV, it will be the focus of this section.

In humans, as with other mammals, the phosphoproteome consists of three residues, phospho-serine (pSer), threonine (pThr) and tyrosine (pTyr) at a ratio of roughly 99% for pSer and pThr in comparison to 1% for pTyr. Other residues, such as phospho-histidine (pHis) and phospho-aspartic acid (pAsp), also occur in bacterial and fungal phosphoproteomes (*24, 57*). In terms of stability, pTyr is the most stable, followed by pThr and pSer (*58*).

Phosphorylation is a process that involves the transport of a phosphate group from ATP to one of the residues by means of protein kinase activity. Once the target is phosphorylated, activation of signalling pathways occurs, and these can be involved in different human diseases (*59*). Phosphatases, on the other hand, are enzymes responsible for dephosphorylation and therefore deactivation of such pathways. Together with kinases, they modulate signal transduction (*23, 57*).

Considering that, in humans, phosphorylation primarily occurs in serine, threonine and tyrosine residues, and that, on average, the content of these amino acids in proteins is about 17%, this gives rise to nearly 700,000 potential phosphorylation sites, and the size of the human phosphoproteome is an active area of research (*60, 61*). Given the biological relevance of protein phosphorylation, that a signalling cascade can be initiated by a single phosphorylation event, and that sample preparation conditions might alter phosphorylation, the relevance of studying the phosphoproteome as well as potential challenges are highlighted, as well as the importance of resources such as PhosphoSitePlus for aggregation and annotation of relevant data (*23, 62*).

More traditional approaches for analysing the phosphoproteome involve two-dimensional gel electrophoresis (2D-GE), western blot, autoradiography, and protein sequencing via Edman degradation. These approaches, however, suffer from a few shortcomings, including poor reproducibility, low throughput, low dynamic range, and use of hazardous reagents such as radioisotopes (*59*). In this context, mass spectrometry is a very powerful technique for identification of phosphorylation sites, given its high efficiency, sensitivity, and selectivity (*23, 59*).

As mentioned previously, one of the challenges behind PTM analysis via MS is the low stoichiometry. In practice, this means that an enrichment or purification step is required to increase the likelihood of detection (*23, 24, 59*). This is more important in DDA analyses, as the precursors selected for fragmentation are the most abundant ones (*23*).

The first way of achieving detection levels is through enrichment. Here, there are several protocols available, including metal oxide affinity chromatography (MOAC), immobilised metal ion affinity chromatography (IMAC), immunoprecipitation-based enrichment and domain-based enrichment (*59*). In the context of the present work, in Paper IV, high performance paramagnetic zirconium (Zr) IMAC beads (MagReSyn® Zr-IMAC HP, ReSyn Biosciences, Edenvale, Gauteng, South Africa) were utilised.

IMAC is a widely used technique based on the electrostatic interaction between the phosphate groups of phosphopeptides and the positively charged metal ions on the beads. A common downside of MOAC and IMAC methods is that peptides containing acidic amino acid groups also show affinity for the metal ions and co-purify with the phosphopeptide (*23, 24, 59, 63*). However, it has been demonstrated that using acidic loading conditions in the presence of organic compounds – 0.1M glycolic acid, 80% acetonitrile (ACN) and 5% trifluoroacetic acid (TFA) – leads to more acidic peptides being neutralised while phosphopeptides maintain their negative charge and affinity for the metal ions, allowing for reduced nonspecific binding. After washing, alkaline buffers can be used to elute the peptides (*63*). From this step, the protocol is the same as the standard sample preparation.

Affinity Proteomics

The potential of MS-based proteomics has been demonstrated not only in the identification of proteins, but also in their quantification. The combination of modern instrumentation with data processing and analysis workflows enabled the use of such technology in global efforts (*64*). However, limitations associated with sensitivity, dynamic range, resolution, and reproducibility are always part of the discussion, and they are directly linked to available instrumentation, with newer equipment being capable of going deeper into the proteome in a more sensitive manner (*65, 66*).

As mentioned in the first chapter, despite research efforts and heavy investments in biomarker discovery, the field remains plagued by very few passing the scrutiny of clinical validation (*1, 5, 9, 10, 67*). The limitations previously mentioned aggravate this scenario even further. For comparison purposes, the typical concentration ranges detected by MS-based proteomics lies around 4 orders of magnitude greater than that typically detected by immunoassays in the clinical setting (*67*).

Two main approaches are available to tackle this problem. The first is based on increasing the analytical sensitivity, but often requires extensive optimisation and may result in a decrease in the signal-to-noise ratio in the context of MS, ultimately resulting in decreased resolution (*67*). These approaches will not be discussed in the context of the present work. The second, on the other hand, involves enrichment, and will be further developed (*67*).

The primary objective of enrichment strategies is to compress of the relative concentration between low and high abundance species (*68*). Two main classes of molecules can be used for such purpose.

The first concerns molecules without a specific affinity, where they bind general classes of proteins (*67*). That is the case of metal ions such as the ones used for IMAC-based enrichment of phosphopeptides described in the previous section. One of the main advantages of this approach is that, by using promiscuous affinity baits, the loss of potentially interesting low-abundance molecules is minimised. Additionally, this unspecificity also allows for new classes of molecules to be captured (*67*).

The second class of molecules is the opposite of the first in that the affinity agents have a specific target. Antibodies, aptamers, enzyme substrates and other proteins or nucleic acids compose this class. The use of such affinity agents for enrichment of specific probes defines the concept of affinity proteomics (*69*). The main advantage of their use is the efficiency of enrichment, i.e., the efficiency with which a specific target is enriched. On the other hand, reusable devices suffer from loss of efficiency over time and may result in carry over, imposing a severe bias (*67*). Of note, the use of different affinity agents has been successfully implemented and used for direct detection by technologies such as Olink proximity extension assay (PEA) (*70*) as well as the nucleic acid Slow Off-rate Modified Aptamers (SOMAmers) used as part of the SomaScan assay (*71*). However, mass spectrometric detection offers advantages in terms of specificity and versatility in discovery efforts (*72*).

An additional problem with the use of highly specific affinity probes is directly linked to the specificity (*67*). Take antibody microarrays for instance, which offer a platform for direct detection of proteins using immobilised antibodies on a solid surface. In terms of throughput cost and amount of sample used, they represent an advance compared to performing analysis with more traditional immunoassays such as via enzyme-linked immunosorbent assays (ELISA) (*73*). The technology has been successfully applied in the clinical applications for diagnosis, prognosis, and classification (*65, 74, 75*).

Scalability may be an issue due to the resolution of such implementation being directly correlated with the number of antibodies used and their specificity. Moreover, in most cases, only antibodies of known specificities are used, requiring a preselection step, resulting in a hypothesis-driven approach which excludes the possibility of discovering new targets (*65, 69, 76*). MS-based proteomics, on the other hand, allows for a hypothesis-generating approach, one which is not limited by existing knowledge (*69*).

To advance the field and harness the benefits of both approaches, affinity and MS-based proteomics can be combined (*65, 76*). The concept is not new, and it has already been described both at the protein and peptide level using different approaches such as ProteoMiner or Equalizer beads (*77, 78*), Stable Isotope Standards and Capture by Anti-peptide Antibodies (SISCAPA) (*79*), Affinity SRM (AFFIRM) (*80*) and Mass Spectrometry ImmunoAssays (MSIA) (*81, 82*).

A conceptually new method, called Global Proteome Survey (GPS), was described by Olsson et. al. (2011) (*65*), and it is the basis of Paper II in this thesis. The method addresses some of the limitations of previously described approaches by allowing the probing of the proteome in a hypothesis-generating manner using a limited number of affinity agents (*65*).

The affinity agent described consists of single chain variable fragment (scFv) antibodies termed Context Independent Motif Specific (CIMS). These CIMS antibodies target short peptide motifs, 4 to 6 amino acids long, and they were designed to be found in up to a hundred human proteins, taking into consideration the protease used for digestion, i.e., presence of lysine or arginine in the C-terminal in the case of tryptic digests (*65, 83*).

A human recombinant scFv library composed of 2 x 10^{10} members was used for selecting the binders, simplifying the generation of new probes and allowing for scalability of the platform (*83*). Of interest, as the name suggests, these antibodies can be used for probing any proteome, irrespective of species of origin. Moreover, the ample range of specificities theoretically allows for coverage of approximately 50% of the nonredundant proteome with one-hundred antibodies (*65, 76, 83*).

The workflow with the CIMS antibodies essentially consists of sample digestion followed by incubation with the antibodies at the peptide level, resulting in an enriched fraction which is then analysed by MS (*65*). The protocol has been successfully applied in breast cancer tissues, using a combination of affinity and MS-based proteomics to define proteins associated with histological grade (*84*).

Although the study reported over 800 peptides which had not been previously reported in the PeptideAtlas and demonstrated the coverage of both high and low-abundance proteins (*84*), since it was conducted, different generations of instruments became available, each allowing increased sensitivity and deeper coverage of the proteome.

Aware of the fact that the study had been conducted using tissue samples, on older instrument with data acquired in DDA mode, and that the affinity captures had been performed manually, in Paper II, we proposed the creation of an automated workflow for the multiplexed affinity enrichment of plasma samples.

After reduction, alkylation and tryptic digestion, a total of 29 CIMS antibodies were tested. A protocol describing the use of magnetic beads for such binders had already been described (*65, 76, 84*). In our case, instead of carboxylic acid beads, paramagnetic pre charged nickel particles (MagneHis™, Promega Corporation, Madison, Wisconsin, United States) were used. A protocol was developed and optimised on the KingFisher Flex robotic system (Thermo Fisher Scientific, Waltham, Massachusetts, United States) for peptide-level affinity enrichment. Out of the 29 antibodies used in the first experiment, a second experiment was performed, this time using an equivolumetric ratio of different binders in order to maximise enrichment events in a single reaction.

Instead of tissue samples or cell cultures, as was described in previous publications (*65, 76, 84*), in Paper II we utilised blood plasma. There are different challenges and opportunities associated with this sample material, which will be elaborated more in depth in the subsequent chapters.

Data Collection and Data Analysis

Irrespective of the specific omics used, a common scenario is the analysis of only a few samples, and subsequent collection of many features (e.g., genes or proteins) due to the associated financial and temporal costs. This results in what is referred to as the "curse of dimensionality", i.e., the low number of samples compared to the number of features results in data sparsity in high-dimensional space (*85*). Taking into consideration the number of biomarker candidates that make it into clinical implementation, this demonstrates that special considerations need to be taken in sample selection, preparation, and data analysis (*9, 85*).

In this chapter, different aspects concerning the sample selection, preparation and subsequent data analysis will be addressed.

Sample Selection

Clinical biomarker discovery is a highly translational field, requiring interdisciplinary collaborations. Consequently, clinical design and experimental design are highly intertwined (*85*). In clinical research, two important aspects must be considered, the anatomy and the physiology of research. The anatomy concerns tangible elements such as the research question, subjects, measurements, and analyses. The physiology, on the other hand, relates to the usability of the study, i.e., how generalisable the findings in the study are to the great majority of the population not analysed by it (*86*).

Before selecting samples, the first step in the anatomy of a study is defining the research question (*85, 86*). The research question defines the aim of the study. The process of scholarly analysis is fundamental to establish a good research question, as flexible studies – where hypotheses are generated after the data collection or where biomarker discovery is a secondary aim – would increase the probability of obtaining significant findings, although those would often be false (*85, 86*).

According to Chapter 2 of Designing Clinical Research, a good research question should follow the FINER criteria, i.e., it should be Feasible, Interesting, Novel, Ethical and Relevant. Essentially, it should not only be intriguing to those working on it, but also be performed in an ethical way and have a significant impact on the field of knowledge, clinical practice, or health policy, while also being manageable, affordable, and fundable (*86*).

A common goal for biomarker discovery efforts or omics efforts is the translation of the findings into the "real world". This either involves applying the findings from laboratory research to clinical studies or applying the findings from such clinical studies into clinical practice. However, the skillsets required in laboratory research and population research do not overlap. Clinical research acts as a bridge between these two, but in practice good collaborations are essential for successful translation opportunities (*86*).

After establishing a good research question based on the FINER characteristics, the next step is the study design. When it comes to clinical research, this can take different shapes and forms, based on whether an intervention is to be applied or not. These designs are known as clinical trial designs and observational designs, respectively (*86*).

Observational designs refer to passive studies where the goal is to make measurements. Depending on how these measurements are taken, different designs can arise. They can be divided into cohort studies, where a group of subjects is followed over time, cross sectional studies, where a single observation is made at a single defined time point, and case-control studies, characterised by the comparison of two groups, one with the outcome of interest, and one without (*86*).

Clinical trial designs, on the other hand, refers to study designs in which an intervention is applied, and the effects of such intervention are evaluated. In this category, randomised blinded clinical trials, where groups are selected at random and the intervention in blinded, is the most recommended design, although nonrandomised and unblinded clinical trials can also be used (*86*).

Translating these concepts to the context of this thesis, Papers II and III are purely methodological. They do not make use of clinical samples, and therefore fall into the category of laboratory research. Paper I also has a methodological aspect to its aim. However, a second aim was investigating the potential use of proteomics data for biomarker discovery. For that purpose, if the work presented here is put on a spectrum ranging from laboratory research to clinical research, Paper I would be further ahead compared to II and III. Finally, Paper IV builds on the results from Paper I and advances it further by having a clinical rationale behind the research question. Here, knowing that the method developed in Paper I could be used for biomarker discovery, the research question is then focused on discovering potential candidates, setting this paper apart from the rest and making it more similar to an observational study of case-control design.

For the selection of samples, or study subjects, two aspects must be defined, the inclusion and exclusion criteria (*86*). Ideally, sample selection would be performed completely at random. The number of samples selected is also important, as both aspects would contribute to a more representative cohort in comparison to the actual population, leading to more generalisable results (*85, 86*). However, when it comes to omics studies, that is not always possible (*85*). Therefore, well defined inclusion and exclusion criteria are critical.

The selection criteria are used to define a study population in the context of the research question. Inclusion criteria can be any set of characteristics used to specify subjects that would be relevant to answer the research question and make the study more efficient, such as demographic, clinical, geographic, or temporal (*86*).

Exclusion criteria, on the other hand, refers to criteria used to define a set of samples that would meet the inclusion criteria, but which could interfere with the study, for instance by lacking a clinical outcome of interest (*86*). Extreme values or samples which might have confounding information should also be excluded in order to avoid bias (*85*).

When translating research from the laboratory to the clinic and ultimately achieving clinical implementation, the goal is to draw the right inferences. The first level concerns internal validity, which refers to inferring the right information from the study, i.e., drawing the correct conclusions based on the research question. The next step is external validity, which concerns the degree with which the findings translate to the population outside the study (generalisability). The right research question and appropriate selection of samples increase the chances of achieving generalisability while also increasing the likelihood of performing the study with a high degree of internal validity (*86*).

To achieve these goals, one must be aware of causal inference. In biomarker discovery studies, the aim is to find biomarkers which have a causal relationship to the outcome of interest. However, in the presence of confounding factors, it becomes impossible to separate causal effect due to the marker or a secondary cause (*85, 86*).

Dealing with sources of variability also requires special attention, as it is likely to introduce bias, as approaches are often not a completely random process (*85*). Specification and Matching are two common approaches in which a specific level of a confounder is stipulated and other values are excluded, and where groups have the same distribution of confounders, respectively (*86*).

Besides confounders, it is also important to be aware of mediators and colliders. Mediators are defined as factors that are naturally occurring as consequence of the outcome, while colliders provide noncausal effects. Differently from confounders, one should not control mediators and colliders at the risk of introducing bias and making cases and controls unnecessarily homogenous, undermining the discovery process (*85, 86*).

In the framework of this thesis, clinical samples were used in Papers I and IV. As the goal in Paper I was more methodological and the work falls more into laboratory research, the only special consideration taken for sample selection was having representation for all different subtypes, without a clinical outcome in mind. In Paper IV, on the other hand, results from Paper I led to the use of RNA concentration as a selection criterium for methodological purposes associated with the feasibility of proteomics analysis in the sample material. Other selection criteria were associated with the research question at hand, for instance receptor statuses and availability of clinical data.

Sample Preparation

From an intervention perspective, the best sample preparation method would be minimal or non-existent. However, the complexity of the proteome, especially that of higher animals and plants, and instrument limitations result in only a small portion of the proteome being actually acquired on a routine high-throughput basis (*87*). The usual steps in the analytical method include sampling, specimen preservation, appropriate sample preparation and data analysis (*87*).

In principle, any sample material may be used for MS-based proteomic analysis. However, the sample preparation process can be challenging. Due to the different nature of the samples as well as the intended analysis, there is no universally agreed method for sample preparation in proteomics, requiring a case-by-case development (*21, 87*).

The first step in sample preparation is the extraction and solubilisation of proteins. The main goal of this step is to obtain the highest possible yield, minimising losses. There are different protocols available for different sample types, but normally a combination of physical and/or reagent-based methods is used (*33*).

In the context of the work presented in this thesis, different sample materials were used, namely blood plasma in Paper II, tissue biopsies in Papers I and IV, and sorted cells in Paper III.

The blood plasma samples corresponded to raw pooled plasma from different individuals. Sodium dodecyl sulphate (SDS) was used as an ionic detergent to dilute, solubilise, and denature the proteins. A dilution step is critical in raw plasma, given the viscosity of the material. Heat was applied to the samples after addition of a reducing agent – Dithiothreitol (DTT) – in order to reduce the disulfide bonds, thus promoting protein unfolding (*87*). Given the instability of the sulfhydryl after reduction, an alkylating agent is required to stabilise it. In this case, iodoacetamide (IAA) was used.

The tissue biopsies samples used in Papers I and IV underwent a different preparation. The detailed protocol is described in Saal et. al. (*88*). Briefly, after routine assessment of the sample by a pathologist, the sample is placed in a collection tube with RNAlater solution. Part of this material is used for simultaneous isolation of DNA, RNA and proteins using the AllPrep method automated in a QIAcube machine (Qiagen). The flowthrough from such isolation, which contains the protein fraction as well as short nucleic acids was then further reduced and alkylated (*88, 89*).

Sample clean-up and digestion

Once proteins are in solution, the next step is removal of contaminants such as the detergents used for extraction and solubilisation, followed by digestion into peptides. Historically, three main approaches have been used for this purpose, in-gel digestion, in-solution digestion, and membrane-assisted protocols (*33*).

Gel-based protocols, through the use of sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE), allow for efficient removal of different contaminants before LC-MS analysis. Depending on the duration of the gel run, it is also possible to fractionate the sample, which is cut into one or more gel slices. However, compared to other methods, there is an element of stochasticity with the digestion efficiency in gels. Another downside is scalability, given the fact that each sample needs to be excised from a gel lane (*33*).

In-solution digestion offers an alternative to the gel-based protocols. It requires fewer steps, and it is more compatible with automation compared to gel-based alternatives. A limitation with this method is that contaminants must be removed prior to digestion, and it can suffer from poor recovery if detergents are not used (*33, 90*).

Methods which rely on a filtering membrane were developed as a way of combining the other two methods, where detergents could be used for increased solubilisation and the clean-up step would be efficient, thus avoiding the gel format. With that in mind, the filter-aided sample preparation (FASP) method was established (*90*). The method consists in using a molecular weight cutoff (MWCO) filter as a reactor, i.e., a scaffold where all necessary steps take place. This method is compatible with a variety of sample materials, but it required several centrifugation steps, ultimately restricting the throughput (*33, 90*). Since its creation, other methods were developed with the idea of being faster, simpler, and more reproducible. One such example is the suspension trapping (S-Trap) method, which traps a protein particulate created from and SDS-solubilised protein solution in a bedding material consisted of quartz or borosilicate glass depth filter and reverse phase membrane compartments (*91*).

A fourth alternative was envisioned inspired by methods developed for high-throughput transcriptomics, given the rapid expansion of next-generation sequencing efforts. These methods utilised paramagnetic beads in both manual and roboticised workflows (*92*). The SP3 method, described by Hughes et. al., builds on principles of solid-phase reversible immobilisation and nanodiamond technologies to enhance and simplify the sample preparation workflow for proteomics (*92*). A different method, named Protein Aggregation Capture (PAC) was described by Batth et. al. in which proteins are non-specifically immobilised, precipitated and aggregated on any type of sub-micron particles, irrespective of their chemistry (*93*). This method was developed with the goal of creating a universal sample preparation method that could be easily scaled to different amounts of input material while maintaining compatibility with different reagents and buffer compositions, robustness, reproducibility, cost effectiveness and practicality (*93*).

To put the different methods in the context of this thesis, in Paper I and II, a solid phase extraction protocol based on paramagnetic beads with hydrophilic interaction chromatography chemistry (MagReSyn® HILIC, ReSyn Biosciences, Edenvale, Gauteng, South Africa) was used. In Paper IV, the PAC protocol was used with hydroxyl terminated beads (MagReSyn® Hydroxyl, ReSyn Biosciences, Edenvale, Gauteng, South Africa). Since sample preparation can be complicated and a source of variability, a major consideration in the choice of protocols described here is possibility for automation. In this context, it was the main motivation behind the use of paramagnetic beads, given that all protocols were automated on a Kingfisher Flex purification system (Thermo Fisher Scientific, Waltham, Massachusetts, United States).

Design of Experiments

When first testing the HILIC solid phase extraction (SPE) protocol, a few questions arose regarding its optimisation and potential interaction between different factors. For this optimisation, the concept of Design Of Experiments (DOE) was applied, which is the process of planning an experiment so that appropriate data is collected and analysed, leading to objective conclusions (*94*).

Basically, an experiment consists of an input and output, and the input is affected by both controllable and uncontrollable factors. The goal of an experimenter is to determine the influence of such factors through a process called strategy of experimentation (*94*).

There are several strategies which can be used. One such approach is the best-guess approach, in which arbitrary combinations of factor levels are used over the course of the experiment, making adjustments and informed guesses along the way. This approach is quite frequently used in practice by engineers and scientists, and it works reasonably well due to the experimenter's background knowledge and experience. The main disadvantages of this approach are the stochasticity of the factor levels, and the difficulty in finding the optimal point for the different factors (*94*).

A second common approach is called one-factor-at-a-time. In this approach, one factor varies while the others remain constant. Over time, this allows for assessing how the different factors affect the outcome. A major advantage of this method is the ease of interpretation, as factors are tested independently, and compared to the first approach, it allows for an optimal solution to be found. However, it disregards the potential interaction between factors, ultimately leading to poor results if interactions are present (*94*).

A more correct approach would be the use of a factorial experiment, in which the factors and levels are determined and simultaneously varied. This allows for assessing the interaction between different factors as well as their effect size. Because of this, factorial experiments make the most use of the data produced (*94*).

For the HILIC SPE protocol, potential factors of interest were the starting amount of material, the bead-to-protein ratio, the binding time, the trypsin-to-protein ratio, and the digestion time. If a factorial experiment were to be adopted for such factors, a total of 2^k experiments would be required, which in this case would result in 32 runs. In order to minimise the number of runs due to costs and time, a fractional factorial design was chosen. These designs are useful in screening situations, where the goal is to identify which factors would be responsible for large effects (*94*).

Based on the results of this experiment, the main factors were determined to be the starting amount of material, the bead-to-protein ratio, and the trypsin-to-protein ratio. These factors also appeared to interact, which then required a more in-depth analysis. To achieve this, a response surface method was chosen.

Response Surface Methodology (RSM) corresponds to a collection of methods used for modelling a response, especially one which is influenced by multiple factors. This methodology allows the experimenter to optimise such response (*94*).

For determining the response surface, the method chosen was a Central Composite Design (CCD). This is a second-order model very commonly used for determining a saddle point in the response surface. It consists of a 2^k factorial design with added centre points and star points. The star runs are included in the design in order to allow the calculation of a quadratic model (*94*).

The results from the CCD suggested that for the intended application, i.e., with the amount of starting material planned to be used, a lower ratio of beads and trypsin could theoretically be used to reduce costs while still maintaining maximum response (**Figure** 5).

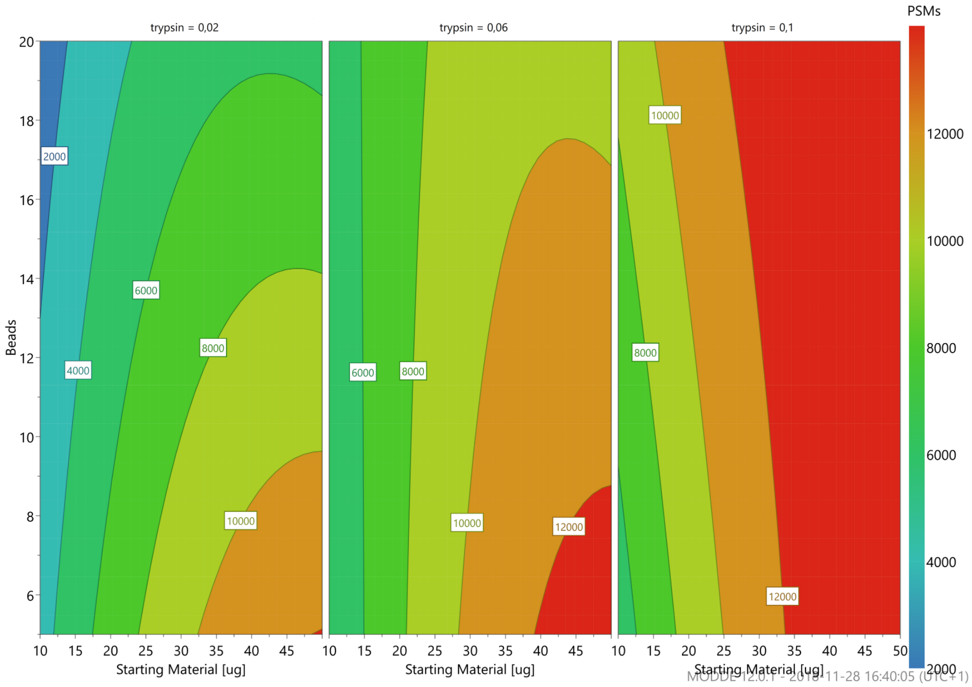


Figure 5 Contour plot mapping PSMs across different concentrations of starting material, bead-to-protein ratio, and protein-to-trypsin ratio.

Protein Depletion or Enrichment

Depending on the type of sample and the goal of the analysis, the proteomic dynamic range, i.e., the difference in concentration between the most abundant protein and the lowest, might prove to be problematic (*33*). The most obvious example is when working with blood serum or plasma, where the dynamic range has over 10 orders of magnitude, and a single protein, albumin, represents over 50% of the protein content (*33*).

Biofluids represent a simple resource for biomarker discovery, with ample use in the healthcare sector. In fact, many biomarkers in clinical practice are found in biological fluids (*87*). It constitutes a desirable resource, as it offers a non-invasive or minimally invasive alternative to biopsies or other surgical procedures (*95*). Blood plasma, for instance, is the most comprehensive proteome of human origin, containing proteins from other tissues and potentially elements of all proteins in the body, with several proteins found in blood already in clinical use (*16, 87, 96*). It is collected via a highly standardised procedure, involving the addition of anticoagulants to blood followed by a centrifugation step (*87, 97*).

Despite the great potential, there are caveats when working with blood-based proteomics. The presence of proteins from different tissues translates in a material of large complexity (*72, 96, 97*). Added to this complexity is the fact that, even though plasma proteins and peptides are required for several biological functions, those that could potentially act as markers of disease-related processes are likely to be low-abundance proteins (LAP), given that they would results from processes such as tissue leakage, cell death or destruction, or abnormal secretion (*98*).

In order to reach consistent detection levels for these proteins, separation strategies are usually required prior to LC-MS/MS analysis, including, for instance, enrichment, depletion of high-abundance proteins (HAP) or extensive fractionation, even with modern instrumentation (*66, 97*).

As the name suggests, depletion usually refers to the removal of HAPs from the sample, given that they might mask the signal from proteins of interest. Enrichment strategies, on the other hand, target dilute species, and the goal is to increase their signal for detection or quantification purposes. The same principles can be applied and performed both at the protein and peptide level, depending on the intended application (*87, 97, 98*).

There are different approaches available for both depletion and enrichment, but they can be broadly divided in physicochemical methods – which make use of properties such as molecular weight, charge, hydrophobicity, and isoelectric point – and affinity-based methods (*97*).

Within the different affinity-based methods, two main classes of molecules are routinely used. The first relies on specific binding to the target protein or peptide, and is represented by antibodies, aptamers, enzyme substrates or natural ligands. The second class has general affinity for classes of proteins, and is represented by dyes, metals, drugs, and other molecules which recognise an affinity tag (*67*).

From the perspective of the work presented in this thesis, the focus will be on affinity enrichment methods, and both classes of molecules are represented. In Paper II, single chain variable fragment (scFv) antibodies are used in blood for compression of the dynamic range. In Paper IV, on the other hand, ion metal affinity chromatography (IMAC) is used for the enrichment of phosphopeptides and subsequent phophoproteome data acquisition.

Data Acquisition

As mentioned in the previous chapter, different data acquisition strategies can be used in bottom-up proteomics workflows, including both targeted and untargeted approaches (*19, 42*). For biomarker discovery efforts, targeted approaches are not ideal, considering that an inclusion list of fragment ions must be provided. For that reason, these are not covered in the context of this thesis. Instead, data-dependent acquisition (DDA) and data-independent acquisition (DIA) are covered.

Data-dependent acquisition (DDA) has been the most common approach used in discovery efforts, where MS1 scans are performed, and the most abundant precursors are selected and sent for fragmentation and MS2 acquisition (*19, 32, 42, 45*). Due to the stochastic nature of this process, run-to-run reproducibility is a known concern in DDA (*19, 45*). Moreover, the same precursor selection process can also have a direct impact on proteome depth depending on the complexity of samples, potentially requiring processes such as fractionation to mitigate it.

Data-independent acquisition (DIA), on the other hand, is not based on precursor detection for fragmentation and MS2 acquisition, as MS2 spectra are continuously acquired. In that way, no prior knowledge is required, and it addresses limitations found in DDA (*19, 45, 46*). There are several methods available that allow for DIA, either via broadband DIA or by splitting the scan range into smaller segments that are consecutively analysed (*46*). The DIA methods used in this thesis belong in the second category, being derived from the SWATH method (*54*).

In comparison to DDA, DIA offers a more reproducible and complete proteomic map, achieving higher proteomic depth (*45, 99-101*). In a clinical context, a lot of research relies of differential expression analysis, making consistent and reproducible measures of analytes crucial (*100*). When comparing different samples, a common approach has been the introduction of labels, either metabolic or chemical, which offers excellent quantification accuracy (*19*). However, considering that such studies often use a large number of samples, this typically exceeds the number of labels which can be added and measured in a single run (*19*), thus making label-free approaches for identification and quantification the preferred choice (*19, 100*). Although more versatile, label-free approaches are more challenging compared to using labelled references, requiring more sophisticated normalisation and chromatographic alignment procedures to avoid erroneous identification. In turn, this makes the choice of strategy for data acquisition and querying essential in order to achieve consistent identifications across several samples (*19*).

A representation of different workflows commonly used in MS-based proteomics for the identification and quantification of peptides is given in **Figure 6**.

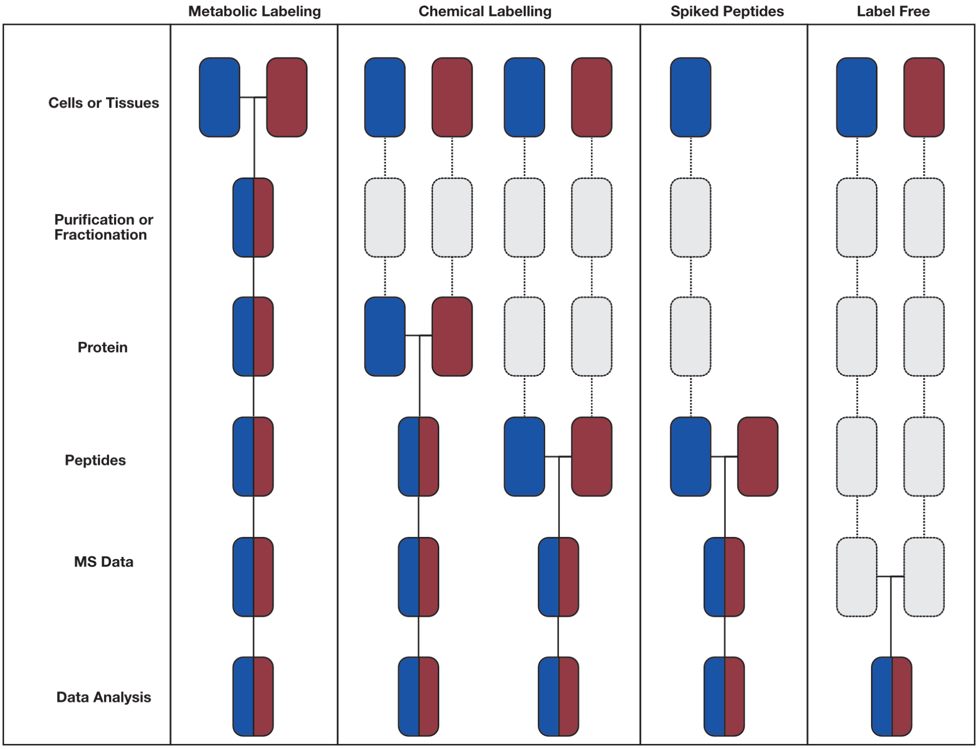


Figure 6 Representation of different quantitative workflows in mass spectrometry. Coloured boxes indicate different samples, while dashed lines and empty boxes indicate sources of experimental variation. Based on (*102*).

In terms of peptide identification, bottom-up proteomics approaches all use MS2 spectra originated from fragmented precursor ions. The different acquisition methods acquire such data at different intervals, being continuous for DIA and discontinuous in DDA (*19*). This makes DIA data amenable to more data querying strategies (*19*). However, the more complex data structure requires significantly more computational efforts compared to DDA (*19, 45, 99-101, 103*).

When it comes to quantification, traditional approaches involved the use of elements such as dyes, fluorophores or radioactivity (*102*). These methods have good sensitivity and dynamic range, but they often require gel-based protein separation, which limits the analysis in throughput and to soluble and high abundant proteins, and the identity of the quantified proteins is not revealed (*102*). MS overcomes these limitations but imposes a different challenge. Peptides have different physicochemical characteristics, resulting in different mass spectrometric responses. For that reason, MS is inherently not quantitative (*102*), although it shows quantitative response for individual peptides. That being said, a number of different techniques have been employed in acquiring quantitative information from MS data.

Typically, quantitative strategies are divided in two groups, those employing isotopic labels and those that do not (label-free quantification) (*64, 66, 102*). The use of isotopic labels has been considered the gold standard in quantification, but their use adds additional preparation steps, costs and offers an inherent limitation in the type or number of samples that can be utilised (*64, 66*). Label-free quantification, on the other hand, besides not being limited by the type or number of samples used in a study, typically offers improved proteome coverage and dynamic range compared to labelling strategies (*64, 66*).

Taking into consideration the different approaches mentioned, label-free acquisition was used throughout the work presented in this thesis, and in Papers II, III and IV it was performed exclusively via library-free DIA-NN (*101*).

Data Preprocessing

Starting from the data collected in the mass spectrometer, several analytical considerations and preprocessing steps are involved until translational questions can be answered. These steps involve peptide identification and quantification, data normalisation and downstream data analysis.

In the previous section, DDA and DIA approaches were highlighted as the most common techniques used in biomarker discovery efforts given that they do not require a list of targets to be provided, as is the case with targeted approaches. Additionally, since a large number of samples is typically used in clinical studies, label-free approaches for identification and quantification of proteins are the preferred choice (*19, 100*).

A well-established goal of proteomics is the identification and quantification of all proteins present in a sample at a given time point (*40*). From a workflow perspective, after data acquisition, bottom-up proteomics data must undergo essential steps, including, for instance, peptide identification, peptide quantification, protein assembly and protein quantification.

Starting with peptide identification, two different strategies have been devised, namely spectrum-centric and peptide-centric (*19, 45, 104*). Essentially, the difference between the two lies in the querying unit, i.e., whether spectra or peptides are queried for scoring and identification (*19, 104*).

There are different tools available for spectrum-centric analysis. As mentioned previously, in this approach, the MS2 spectra are considered the querying unit, and peptide identification can be performed via database searching, where experimental spectra are compared and scored against in silico generated theoretical spectra from a sequence database; via spectral library matching, where the obtained spectra are compared against a previously generated spectral library; or via de novo sequencing, where the identification of experimental fragments is performed without the use of a library or database (*19, 104*).

Most DDA studies make use of database searching as a spectrum-centric approach to peptide identification (*19, 104*). There are different software available to achieve this, including, for instance SEQUEST, Mascot, X!Tandem, MaxQuant, Comet, MS-GF+ and OMSSA (*104-111*).

Because the querying unit in spectrum-centric analysis is the MS2 spectra, a spectrum must first be matched to at least one peptide sequence to yield a peptide-spectrum match (PSM). These are then scored and an estimation of the degree of confidence is generated. A confidence estimate at the peptide level can then be generated based on aggregation of the best suited PSMs. However, because only the matched peptides are scored, those that were not assigned a PSM are automatically considered missing (*104*). Added to this complexity is the stochastic nature of DDA acquisition, ultimately making it impossible to confidently state that a certain peptide is missing from the sample, considering it may have never been selected for fragmentation by the survey scans (*104*).

In turn, the missing peptides are assigned very low confidence estimates. Because of this imputation of confidence estimate, this approach can lead to biases when inferring protein identity (*104*). The "protein inference problem" is a common challenge in bottom-up proteomics (*40*). It refers to the loss of information which happens due to the digestion of proteins into peptides, making it difficult to trace back the protein that originated a certain peptide, as the same peptide may be present across different proteins (*40*).

Despite the potential bias of the confidence estimate imputation, an advantage of spectrum-centric analysis is the direct estimation of the false discovery rate (FDR) (*19*). Because the search tools deployed are agnostic, decoy sequences can be introduced, and the number of decoy sequences identified can be directly used to estimate the FDR of a dataset (*19*).

For DIA data, despite the added complexity, most of the initial studies using this acquisition mode relied on using the same spectrum-centric tools originally designed for DDA data (*19, 54*). DIA solves the random sampling problem found in DDA, but it also introduces new challenges. Most notably, in order to cover the same m/z range as in DDA, wider precursor isolation windows must be employed, reducing precursor selectivity (*104*). Because traditional spectrum-centric analysis assumes that fragments originate from a single isolated precursor, these tools perform poorly when directly used for DIA data analysis (*104*).

Since DIA MS2 spectra are essentially all mixed, applying spectrum-centric analysis in such data would require allowances for contribution of multiple precursors, allocating intensities to contributing peptides and performing adjustments in confidence estimation. Some solutions involve deconvolving mixed spectra or matching mixed spectra to combinations of candidate peptides. A downside of these approaches is suppression of the dynamic range, in which signal from low-abundance peptides is overwhelmed by that of high abundance ones (*104*).

A new approach was introduced in 2012 by Gillet et. al. (*54*) that uses a chromatogram-based approach for analysis of DIA data (*19, 54, 104*). In it, ion chromatograms are used to identify and quantify query peptides (*19, 104*). This approach allows for a change in paradigm in which instead of relying on querying MS2 spectra, a peptide can be assessed as present or absent, generating a list of hypotheses ("Is this peptide present in the data?") (*19, 104*). This approach is known as "peptide-centric". The presence of a peptide is evaluated in a method akin to what is done in SRM, allowing a more biologically oriented way of querying LC-MS/MS data, thereby making it suitable to answer a variety of biological questions (*19, 104*).

Peptide-centric analysis is more suitable to answer biological questions because peptides can be directly queried and scored based on their presence or absence, differently from only identified peptides with a PSM (*104*). This also contributes to alleviating the protein inference problem, as no imputation of confidence estimate is performed for "missing" peptides, therefore making protein inference more transparent (*104*).

By not using individual spectra as querying units, peptide-centric analysis tolerates the mixed spectra generated in DIA, making it more suitable for this data type. However, that also imposes a complication, where the mixed spectra allow for the same spectrum to be assigned to multiple peptides (*104*). For estimation of the FDR, a similar use of decoys can be adopted. However, because the number of decoys increases with the number of peptides queried, the FDR cannot be estimated by simply counting decoys as it can be done in spectrum-centric analysis (*19, 104*).

There are different software solutions available for DIA data analysis. A popular commercial option for DIA data processing in the past few years has been Spectronaut (*99, 112*). Other alternative software, including open-access options include Skyline, OpenSWATH (*113*), DIA-Umpire (*114*), EncyclopeDIA (*103*), MaxDIA (*115*) and DIA-NN (*45, 99, 101*).

Traditionally, DIA data analysis has relied on project-specific libraries built either by fractionation or repeated injections (*99*). However, the use of such libraries has been questioned recently (*45, 116, 117*). Other approaches, such as publicly available libraries, gas-phase fractionation, and especially in silico generated libraries have been gaining popularity (*45, 99, 118, 119*).

Different studies have compared the different approaches to DIA data analysis (*45, 99, 100*). Gotti and colleagues (*45*) demonstrated that the choice of software for DIA data processing is fundamental for the quality of downstream analyses, as well as the size and number of isolation windows, and the use or not of a library. Lou and colleagues (*99*) further extended other benchmarks by evaluating application of different software both in global proteomics and phosphoproteomics applications, with mixed results. Of note, they highlight the use of DIA-NN in library-free mode as having overall best performance for global proteomics efforts, considering it had better false discovery rate (FDR) control, quantification accuracy and precision, as well as sensitivity and specificity in the detection of differentially expressed proteins (*99*). In phosphoproteomics efforts, the FDR control performed by DIA-NN was also the best, although Spectronaut analysis yielded higher sensitivity, suggesting that a combination of different software could be desirable for such applications (*99*).

In the work presented in this thesis, a benchmark of different tools was performed as part of the research question for Paper I. In it, DDA was compared to two different DIA approaches, one using chromatographic libraries (EncyclopeDIA), and one with in silico generated library (DIA-NN in library-free mode). According to our results, DIA-NN stood out both in terms of the overall number of features and in the number of features present across all samples analysed, even with match between runs enabled in all three approaches. Of note, even though the use of a library via EncyclopeDIA resulted in greater data completeness compared to DDA, as is expected, the overall number of features was lower, highlighting the importance of library generation when such approach is used, since only analytes present in the library will possibly be detected (*103*).

DDA data was analysed in MaxQuant, while DIA data was analysed using both EncyclopeDIA and DIA-NN (Paper I). While MaxQuant analysis is considered spectrum-centric, the other two methods incorporate aspects of both spectrum- and peptide-centric analysis.

For EncyclopeDIA, in addition to collecting DIA data from all samples analysed, a chromatogram library is built. In Paper I, a UniProt human FASTA file and the corresponding Prosit (*120*) spectral library from ProteomicsDB were used to create a spectrum library. A pool of different samples was used to collect narrow-window DIA data via gas-phase fractionation, which was then utilised to search the spectrum library and correct it, keeping only peptides detected in the pool and removing low-scoring matches to limit the search space, followed by inclusion of chromatographic and fragmentation data to create a corrected chromatogram library which is then used for searching the wide-window DIA data from individual samples (*89, 103, 121*).

DIA-NN, on the other hand, was used in library-free mode in Papers I through IV. In library-free mode, the DIA-NN workflow first involves in silico generation of a collection of precursor ions annotated with different fragment ions. Decoys are then generated, followed by extraction of chromatographic data, which includes retention time information and elution profiles of both precursor and fragments. Along the workflow, 73 peak scores are calculated, and a linear classifier is trained to select the best peak per precursor (*101*).

To avoid the complication of having a peak contributing to the identification of multiple precursors, a single, best, precursor is chosen per spectrum, thereby improving identification, and enforcing strict FDR control (*101*).

In label-free quantification, two types of information can be used to compare different samples, namely MS (or MS/MS) signal intensity or the number of MS/MS spectra (known as spectrum counting) (*64, 66, 102*).

Spectral counting approaches are based on the observation that the higher the amount of a protein in a sample, the more spectra will be collected. Because it uses fragment spectra, it benefits from extensive MS2 acquisition (*102*). There is some scepticism regarding spectral counting approaches in that no physicochemical property of peptides is actually measured, and it assumes that the linearity of response is the same for every protein at the same time that the number and length of peptides differ for different proteins, making smaller proteins suffer from more variability in quantification compared to larger ones (*66, 102, 122*).

Label-free quantification using MS signal intensity, called intensity-based quantification or precursor-based quantification is typically achieved by associating ion intensities to the elution profile (*122*). Compared to spectrum counting, it offers better accuracy and deeper coverage (*66, 122*). This process usually involves feature detection and chromatographic alignment (*66*). The alignment process allows for corrections in retention time across samples, ultimately allowing for more features to be detected by propagation of identifications across runs (*64, 66*). Important to note, quantification benefits from more scans across the chromatographic peak (MS1), while confidence in detection benefits from additional MS2 scans, ultimately introducing a conundrum especially when the instrument is operated in DDA mode where confidence in identification comes at the sacrifice of precision of quantification (*102, 122*). Operating the instrument in DIA mode, in turn, enables identification and quantification at the same time, evidenced by the robust quantification measures from DIA studies (*19, 45, 100, 122*).

One crucial requirement in label-free quantification experiments is reproducibility, and this comes in instrument stability over time, simplification of sample preparation steps and introduction of automation (*64, 66*). In this context, data normalisation is essential (*66, 123, 124*). Steps in sample handling and data processing can all introduce biases that, if left unchecked, could lead to misleading conclusions. Since the sources of bias can be several and unknown, this makes it challenging to determine a best normalisation method (*123, 124*). For that purpose, when preprocessing MS data in the different projects in this thesis, NormalyzerDE (*123*) was used for comparing different normalisation methods prior to selecting the best performing one.

Data Analysis

The resulting data, after these preprocessing steps are often presented in tabular format representing peptide or protein intensities across samples. The goal in biomarker discovery is often exploratory, aiming at identifying features and patterns which hopefully assist in answering the research question of interest. The first step is usually to find which proteins are differentially abundant when comparing groups of samples, as these could be potential biomarkers of interest. Because NormalyzerDE has that functionality built in, this step was also included across all projects presented in this thesis using LIMMA empirical Bayes statistics, since it was proven to perform well for this use case.

Differential analysis can be conducted using statistical tests and may lead to long lists of analytes to explore further. There are different tools available to aid in these efforts, such as enrichment analysis and different clustering methods. More recently, the ability to routinely collect data from different omics sources have made integrative analysis more common. Some of these methods are discussed in the next few sections.

Enrichment Analysis

With the increase in popularity of high-throughput omics technologies, the traditional approach of investigating biological questions with one or a limited number of features at a time has been challenged (*125*). Typically, such technologies produce long lists of features, and although differential expression analysis is a common and easy step to take, interpretation of results may prove challenging (*125-127*).

Starting with the assumption that if a given biological process is disrupted in a given group, features associated with that process will covary, the information given by individual features can be summarised in higher level biological elements such as pathways or processes, making the basis of enrichment analysis (*125, 127*).

The popularity of these methods promoted the development of several tools (*125*). Huang et. al. (*125*) classified different tools in three distinct categories, namely singular enrichment analysis, gene set enrichment analysis (GSEA) (*128*) and modular enrichment analysis (*125*).

Singular enrichment analysis is the most traditional approach in which preselected features are used to test biological terms on a one-by-one basis. The significance of the enrichment is then calculated by means of methods such as Chi-square, Fisher's exact test, binomial probability and hypergeometric distribution (*125*). Overrepresentation analysis (ORA) is an example of such approach (*129*).

Gene set enrichment analysis, on the other hand, differs from ORA in that all features are used for calculation of enrichment, instead of a preselected list, thereby eliminating the need of determining a threshold for feature selection. A maximum enrichment score is calculated based on a ranked list of all features and significance is assessed by means of Kolmogorov-Smirnov-like statistics (*125, 128*).

Since these enrichment tools depend on annotation, their usefulness is directly linked to availability of curated databases (*126, 127, 130*). In Papers I and IV, GSEA is performed utilising the "Hallmark" gene set, part of the Molecular Signature Database (MSigDB), a curated set of genes that portray well-defined biological processes while also reducing variation and redundancy (*126*). The analysis in both occasions is performed using an implementation available via the ClusterProfiler library (*131*).

Of note, the databases mentioned previously are gene centric. Although it is relatively trivial to map proteins to genes and perform analyses such as GSEA, in the event of PTM analysis, this information would be lost (*132*). Considering the importance of PTMs in the regulation, localisation and interaction of proteins, a PTM-centric database with data for similar enrichment analysis could provide an additional layer of information not seen at the gene or protein levels (*132*). In Paper IV, one such database, PTMsigDB (*132*) is used for the enrichment analysis of phosphosite data.

Consensus Clustering

In addition to the enrichment analysis tools mentioned previously, the widespread use and availability of omics data also incentivised the development of tools aimed at the discovery of new taxonomies, or classes based on intrinsic information from the analysed features (*133, 134*).

Cluster analysis is one way to achieve this, in which two main questions are addressed, namely the determination of the right number of clusters, and how to assign confidence measurements to such groups (*133, 134*).

As mentioned in the first chapter, the classification of diseases is still heavily based on signs and symptoms (*6*). The abundance of omics data highlights the potential for improved disease and patient classification, with prognostic and diagnostic implications that can be harnessed by precision medicine (*7, 8*). Among the different tools available, consensus clustering is commonly used in molecular cancer research (*135-137*), offering the benefit of class discovery based directly on intrinsic molecular features.

For this purpose, in Paper IV, the ConsensusClusterPlus (*133*) package is used to determine potential new subtypes utilising transcriptomics, proteomics and phosphoproteomics data. The package builds on the original consensus clustering method by Monti et. al. (*134*) which utilises a resampling-based method for class discovery and extends it further providing additional data visualisation and possibility of utilising custom clustering functions (*133*).

Integrative Omics Analysis

The advances of molecular medicine and decreasing costs of high-throughput technologies have led to an ever-increasing amount of collected biological data which, in turn, challenges traditional approaches of single biomarker measures and "one size fits all" interventions in favour of biomarker panels and precision medicine for better prediction and understanding of complex systems such as disease (*1, 4, 6-8, 138-141*).

However, our ability to translate these data for clinical implementation has been limited, as evidenced by the large gap between the number of biomarker candidate published and those that get approved (*1, 5, 9, 10, 138*). Integration of such data (multiomics analysis) could potentially address this issue by enabling the generation of a more complete molecular profile of diseases or patients compared to what individual omics allows (*141, 142*).

Traditionally, precision medicine would normally refer to the use of genomics data given early technological advances and availability of data. Early efforts in profiling mutations and/or chromosomal rearrangements in cancer have contributed to the identification of features associated with increased cancer risk, such as BRCA1 and BRCA2 mutations in breast and ovarian cancers (*138, 143*).

The advance of sequencing technologies shifted the analysis towards the transcriptome, and, since then, transcriptomics has directly contributed to the understanding of different processes (*138, 144*). Moreover, transcriptomics data has been used for predicting patient out come and treatment response, and several gene expression-based tests are on the market aimed at predicting prognosis and risk of recurrence (*138, 144-148*).

More recently, more and more studies have investigated the proteome, as proteins are not only the main mediators of cellular function, but also well established in terms of clinical application (*14-16, 138*). Added to this complexity is the fact that PTMs can directly affect function, localisation and turn-over rate of most proteins, highlighting their potential (*15, 39*).

With these examples and given the fact that many different omics are available besides genomics, transcriptomics and proteomics (e.g., metabolomics, epigenomics, interactomics and lipidomics), it is now possible to see that the analysis of single omics can be overly simplistic (*15, 138, 140, 141*). Therefore, multiomics efforts are required in order to integrate these data and acquire a more complete understanding of these complex biological processes (*140, 142, 149*).

Argelaguet et. al. (*149*) defines the first step in data integration as the definition of anchors. The authors define three integration approaches (horizontal, vertical and diagonal) based on the use of either features or samples as anchors, or the absence thereof (*149*).

Horizontal integration is defined by study designs aimed at using features as anchors and integrating data across different samples (*139, 140, 149*). These approaches are commonly seen as batch correction problems, with the goal of addressing external sources of variation. One example of such approach is single cell RNA sequencing, where genes are the anchors, and the data is integrated across different samples (*149*).

Vertical integration, on the other hand, uses samples as anchors, i.e., multiple data types are collected for the same set of samples (*139, 140, 149*). Depending on the research question, vertical integration approaches can be further divided into global, when the aim is to identify overall patterns of covariation across omics, or local, when the aim is to investigate a specific question, for instance the mapping of a pathway, using data from different omics (*149*).

Finally, diagonal integration, as the name suggests, concerns those study designs where no anchors are used, i.e., data is integrated across different modalities and samples (*149*).

Based on when in the analysis process the integration takes place, these integrative approaches can be further divided into early, intermediate and late integration (*139, 141*).

In early integration methods, all data is combined as part of the first step, creating a single matrix or graph, which is then used as input for modelling. It represents the simplest form of integration, and it has the advantage of allowing for the model to consider any type of association, although this comes at the cost of potentially working with very large data structures containing highly correlated features, outliers and noise (*139, 141, 142*).

In late integration, each data modality is first modelled independently, and these results are then used as features in a second-order model, which is used for prediction or majority voting. It is a good approach when the different data modalities are known to have distinct predictive power, and ensemble machine learning models are a good example of late integration. However, as a downside, there is the possibility to miss potential associations across different omics, since no direct integration step is involved (*139, 141, 142*).

Intermediate integration sits between the other two approaches, i.e., the integration does not rely on combining input data, nor does it involve modelling the data separately. Instead, it aims at keeping the structure of the original data while introducing preprocessing steps to address issues such as redundancy. This approach can have better performance compared to the other two, but it is often considered more demanding from a development standpoint (*139, 141, 142*).

By collecting publications between 2018 and 2021 with the term “multiomics” in the title or abstract, as well as “integration” and “disease” as general terms in the text, Athieniti et. al. (*141*) were able to get an overview of the types of data that are most commonly integrated, and which integration approach and method are commonly employed (*141*).

The authors identified transcriptomics as the most commonly used data modality, both in cancer and non-cancer studies. Epigenomics and genomics are next in the context of cancer-related studies, while proteomics and metabolomics are more common in diseases other than cancer. In addition, they also investigate which omics are more often combined, and show that, in cancer, transcriptomics and epigenomics are often combined, while in other diseases transcriptomics is usually used alongside proteomics (*141*).

As far as integration approaches go, intermediate integration models are more commonly adopted, aimed at detecting molecular patterns, identifying subtypes or understand regulatory processes. Within these categories, the authors highlight different types of models which can be used, including joint dimensionality reduction, kernel-based methods, network-based methods and deep learning (*141*).

In the scope of the present thesis, although transcriptomics and proteomics data were analysed as part of Paper I, a more integrative approach was adopted particularly in Paper IV, where the aim was to combine transcriptomics, proteomics, phosphoproteomics and immune infiltration data generated based on Paper III to gain insight into metastatic processes in oestrogen-receptor positive breast cancer, evaluate the occurrence of different unknown subtypes and identify biomarker candidates associated with such processes.

A common approach when analysing single omics is to perform dimensionality reduction. Principal Component Analysis (PCA) is very commonly applied in this context, where high-dimensional data is projected into a low-dimensional space by means of orthogonal components that maximise variance. Given the relative simplicity and interpretability of the method, different generalisations of PCA have been developed considering the integration of multiple data, including methods based on matrix factorisation such as Canonical Correlation Analysis (CCA) (*150*), Joint and Individual Variation Explained (JIVE) (*151*), Multi-Omics Factor Analysis (MOFA) (*152, 153*), Projection to Latent Structures (PLS) (*154*), among others (*149*).

Specifically, in Paper IV, MOFA was adopted. Despite the various models mentioned previously, a common downside is interpretability (*153*). The model uses a probabilistic Bayesian framework to decompose the data into factor and weight matrices (*141*). Sparsity is also adopted in order to remove features and factors, keeping the most important information, and the model can handle missing values (*141, 153*).

Although the model is linear, which may result in it missing non-linear associations in the data, it provides a framework that allows for interpretation of the factors in terms of features and contribution of different data modalities, as well as correlation with clinical covariates (*153*).

Applications in Breast Cancer

Breast Cancer Epidemiology and Risk Factors

As mentioned in the third chapter, according to recent estimates from the World Health Organization (WHO), NCDs are responsible for over 75% of premature deaths (*11, 12*), with cardiovascular disease and cancer being the two mainly responsible for these numbers. To put in perspective, cancer is responsible for one in six deaths (16.8%) and one in four deaths (22.8%) from NCDs, being among the two leading causes of death in 127 out of 183 surveyed countries (*12, 155*).

According to the latest GLOBOCAN estimates from 2022, 20 million new cases and 9.7 million deaths occurred across the world due to cancer (*155*). Among different cancer types, female breast cancer was responsible for over 2.3 million new cases and 666,000 deaths worldwide, representing 11.6% of all cancer cases and 6.9% of all cancer deaths (*13, 155*). Although lung ranks higher both in terms of incidence and mortality, breast cancer ranks first for women in both aspects (*11, 13, 155*). These numbers represent an important aspect to consider in terms of life expectancy, but it is also important to highlight the disproportional cancer mortality in women, with over one million children becoming orphans in 2020 due to their mothers dying of cancer, half of which were attributable to female breast and cervical cancers (*155, 156*).

Assuming that cancer rates remain the same, the GLOBOCAN estimates that 28.4 million new cases will occur in 2040, and 35 million new cases in 2050. This is the result of changes in population growth and aging, with the global population estimated to reach 9.7 billion by 2050, but also increased prevalence of risk factors (*11, 155*). These aspects highlight the need for better treatment and management options (*11*).

As with other cancers, breast cancer is also highly correlated to HDI, with higher HDI being associated with an increase in both incidence and mortality (*11, 155, 157*). This is illustrated by the highest incidence rates being found in France, Australia/New Zealand, North America and Northern Europe, with a 4-fold increase in incidence when compared to South-Central Asia and Middle Africa (*11, 155*). This same trend, however, is not observed in terms mortality, where transitioning countries have significantly higher mortality rates compared to transitioned countries (*11, 155, 157, 158*). Of note, despite the higher incidence in high-HDI regions, when considering population density, most of the global population is found in less developed regions, translating into over 50% of breast cancer cases occurring in these regions, resulting in a significant burden of the disease (*157*).

In order to understand these different trends in incidence and mortality, one must consider the associated risk factors for developing breast cancer. In general, risk factors can be divided between modifiable and non-modifiable (*159*). The two most important risk factors are being female and increased age (*157, 159*). Other non-modifiable factors which have a high relative risk include increased breast density, presence of precancerous breast lesions, previous chest wall irradiation and genetic predisposition (*159*).

Modifiable risk factors are both reproductive and non-reproductive but are often associated with a so-called Western lifestyle (*157, 160*). They include early age at menarche, later age at menopause, older age at first childbirth, nulliparity (or lower parity), decreased duration of lactation, exogenous hormone administration (oral contraceptive, postmenopausal hormone replacement therapy), as well as alcohol consumption, smoking, obesity and physical inactivity (*155, 157, 159, 160*).

When considering the risk factors in conjunction with the HDI, it becomes easier to understand the trends observed in transitioned and transitioning countries. In transitioned countries, an initial increase in incidence between the 1980s and 2000s was then met by recommendations against hormone replacement therapy. In terms of mortality, however, the trend is not the same, and that is a reflection of early detection, e.g., via mammographic screening, as well as better access to effective treatment options (*155, 157*). In less developed regions, however, where over two thirds of the breast cancer deaths in 2020 were recorded, delayed presentation is more common, ultimately contributing to increased mortality due to the disease being detected in advanced (stage III) or metastatic (stage IV) stages (*155, 157, 161*).

Due to the increasing burden of breast cancer, in 2021 the WHO launched the Global Breast Cancer Initiative (*161*). High-quality and accessible cancer programmes are lacking in low- and middle-income countries, contributing to the cancer burden and representing a threat to public health, economic growth and the achievement of the United Nations (UN) Sustainable Development Goals (*161*). Considering that the risk factors with highest relative risk are non-modifiable, risk factor reduction alone is insufficient for breast cancer control, requiring systematic changes to be implemented (*159, 161*).

With the goal of reducing breast cancer mortality at an yearly rate of 2.5%, the program is based on the implementation of three pillars: (i) health promotion and early diagnosis through education about risk-reduction strategies and basic breast health, and education of healthcare providers on signs and symptoms of early presentation of breast cancer to improve early detection; (ii) timely diagnosis, focused on establishing accessible and rapid-diagnosis systems; and (iii) comprehensive breast cancer management, through improving access to high-quality treatment options and implementing a personalised treatment and rehabilitation plan, minimising financial toxicity (*155, 161*).

Diagnosis and Clinical Management

Breast cancer is an incredibly heterogeneous disease, and this heterogeneity is seen etiologically, histopathologically and molecularly (*148, 160, 162, 163*). For instance, in the context of epithelial tumours of the breast, a total of 43 different morphology codes are defined by the International Classification of Diseases for Oncology (ICD-O-3.2), spread across different types of adenosis and benign sclerosis lesions, adenomas, epithelial-mesenchymal tumours, papillary neoplasms, non-invasive lobular neoplasia, ductal carcinoma in situ, invasive breast carcinoma, rare and salivary gland-type tumours and neuroendocrine neoplasms (*160*).

In terms of diagnosis, most procedures fall into three categories, namely screening tests, diagnostic tests and monitoring tests (*162*). Screening tests include self-performed manual palpation of breast (BSE) and are performed routinely in individuals without suspected breast cancer. A palpable mass is the most common clinical sign of invasive breast cancer (IBC), though other signs may include skin retraction, nipple inversion, nipple discharge and changes in the size or shape of the breast, or changes in skin texture or colour (*160, 162*). Upon presence of one or more of these signs, a diagnostic test is required to establish a definitive diagnosis, given that common symptoms of breast cancer can also occur in benign breast disease (*160*).

Diagnostic tests include different forms of imaging techniques, such as ultrasonography, mammography, Magnetic Resonance Imaging (MRI), different forms of biopsies such as Core Needle Biopsy (CNB), Fine Needle Aspiration (FNA) or surgical diagnostic biopsy, and determination of a histopathological phenotype, which has been the main diagnostic method used (*160, 162*).

In terms of imaging techniques, although population-wide mammographic screening has been the standard, it is not necessary for sustained reduction in breast cancer mortality (*161*). While ultrasonography has a higher false-positive rate, it is the recommended imaging procedure for women under the age of 40 years, and it can also be used to improve sensitivity for mammographically dense breasts (*159, 160*). Of note, combined mammography and ultrasonography results in very low false negative rates, in the range of 0% to 3% (*160*). MRI is considered the most sensitive method, but it suffers from lack of specificity, and when associated costs are factored in, it results in MRI being recommended mostly in very high-risk cases (*160, 164*).

In the histomorphological classification of breast cancer, a variety of features are taken into consideration, including the histological subtype, Nottingham grade, tumour spread in angiolymphatic spaces, and associated in situ component. Additionally, tumour size, distance to margins, stromal changes and the presence of Tumour-Infiltrating Lymphocytes (TILs) are also considered important features for tumour classification (*160*). Based on different histological features, histological subtypes are defined. Essentially, a special histological type is assigned if ≥ 90% is considered to be of that special type, e.g., invasive lobular carcinoma (ILC); otherwise, the invasive breast cancer of no special type (IBC-NST/ NST) is used (*160*).

Other classifications, such as tumour staging, offer crucial prognostic information. For instance, the 10-year survival rate of patients diagnosed with early-stage breast cancer is over 90%. On the other hand, advanced stages such as metastatic breast cancer have an associated 5-year relative survival rate of approximately 25% (*165*). However, the histopathological classification performed by pathologists constitutes the basis for all other classification systems (*165*).

Clinically, all IBCs are classified in terms of biomarker-defined subgroups based on oestrogen receptor (ER) and human epidermal growth factor receptor 2 (HER2) status. Based on this analysis, four subtypes are defined, namely ER+/HER2-, ER+/HER2+, ER-/HER2- and ER-/HER2+. These groups do overlap in terms of morphological features, but they provide diagnostic and prognostic information (*160, 165, 166*).

The main role of the ER status lies in the predictive utility in identifying a group of patients that would potentially benefit from endocrine therapy (*160, 165, 166*). The ER is predominantly nuclear and exists in two forms, ER-α and ER-β (*167*). They are both associated with proliferation, albeit in opposite ways, i.e., ER-α is said to enhance proliferation, while ER-β inhibits it (*167*). That being said, current immunohistochemistry (IHC) guidelines only stain for ER-α (*167*). ER is also considered prognostic, given that, in general, ER+ tumours have better prognosis over short term compared to negative cases (*160*).

When it comes to IHC staining of ER, there's an ongoing discussion regarding the cutoff due to the potential benefit in endocrine treatment. More specifically, ER staining between 1% and 10% constitutes a heterogeneous group which may have more similarities to ER-negative cases instead of ER-positive, requiring special attention (*160, 167*). Current recommendations consider ER-positive when ≥1% cells stain, and ER-negative when <1% or 0% (*160*). However, in the case of the percentage of stained cells being 1-10%, ER positivity should be reported as "low positive", and for negative cases, it should be noted whether results were <1% or 0% (*160*).

In conjunction with ER staining, staining for the progesterone receptor (PR) is also typically done (*160, 167*). Specifically in ER+ cases, PR is considered a prognostic marker, with higher PR levels associated with a better outcome. Similarly to ER staining, an optimum cutoff needs to be determined for PR measurement via IHC (*160, 167*). Although there is evidence that ER+/PR- tumours have a worse response to endocrine therapy compared to ER+/PR+ tumours, it is not well validated for that purpose, and so patients still typically receive endocrine therapy (*160, 165*).

Besides ER and PR, HER2 status is also investigated. As the name suggests, HER2 is part of the family of growth factor receptors, and is involved in the regulation of cell proliferation, development and survival. The protein is located on the cell surface, and 10% to 20% of breast cancer cases are characterised by overexpression of HER2, which happens as a result of the amplification of its gene (*ERBB2*). This overexpression is linked to a more aggressive phenotype, characterised by increased proliferation, cell motility and angiogenesis (*160*).

Since HER2 is characterised both as protein overexpression and gene amplification, its status can be determined via IHC at the protein level or utilising In Situ Hybridisation (ISH) to detect the gene amplification, although the lower costs and availability of IHC makes it the preferred method of choice for determining HER2 status (*160, 165, 167*). HER2 is also considered both prognostic and predictive due to the more aggressive phenotype. However, the main application of HER2 is predictive, as it is used to identify a group of patients which would benefit from anti-HER2 targeted therapies such as trastuzumab (*160, 165, 167*).

Molecular Classification

Considering the large number of morphological codes available for breast cancer classification and the overlap found in the phenotypes determined by ER and HER2 status, it becomes clear that breast cancer is a very heterogeneous disease, and that heterogeneity extends to the molecular level (*148, 160, 162-165, 168-171*).

To give a historical overview of breast cancer classification, the first diagnoses would only consider visible signs and symptoms of the disease, and it was only in the 18th century that it was understood as a local disease that would eventually spread to become systemic. As a consequence, mastectomy was the standard of care until the second half of the 20th century (*165*).

The second half of the 20th century brought many advances to breast cancer classification and therapeutic approaches. The 1960s were marked by the approval of Tamoxifen as an anti-oestrogen drug, the 1980s by the introduction of mammographic screening programs, and the 1990s by the introduction of novel chemotherapeutic options, implementation of sentinel lymph node biopsy, identification of the role of BRCA1 and BRCA2 mutations in breast cancer pathophysiology, and the introduction of the first anti-HER2 targeted therapy drug, Trastuzumab (*165*). The changes implemented during this period were crucial for a reduction in the incidence and mortality trends seen in the early 2000s (*155*).

Based on the assumption that the phenotypic diversity seen in breast cancer would be met by a corresponding diversity at the molecular level, in the early 2000s, Perou et. al. (*172*) proposed the molecular classification of breast cancer using gene expression patterns of 1753 genes captured via cDNA microarrays in conjunction with a hierarchical clustering algorithm (*172*). This work defined five distinct molecular subtypes, known as the "intrinsic" subtypes, namely Luminal A, Luminal B, HER2-enriched, Basal-like and Normal-like (*171, 172*). These intrinsic subtypes show correlation with overall survival, prognosis and therapy response, and have dominated breast cancer research in the last two decades (*148, 162, 165, 168, 173, 174*).

A more modern definition of such subtypes was developed in 2009 by Parker et. al. (*171*) in the form of a quantitative Real-Time Polymerase Chain Reaction (qRT-PCR) test comprised of 50 genes plus a classification algorithm called Prediction Analysis for Microarrays, receiving the name of PAM50 (*171*). In addition to the subtype prediction, Parker et. al. (*171*) investigated the utility of this multigene signature in the prediction of risk of recurrence (ROR) (*171*). Both are now available and part of the Prosigna assay (*166, 170, 171*).

These subtypes were shown to correlate to different clinical subtypes (*165, 166, 170*), illustrated in **Figure** 7. Specifically, ER positivity is a proxy for luminal subtypes (luminal A and luminal B), HER2 overexpression/amplification is a proxy for the HER2-enriched subtype, and negative ER, PR and HER2 status, known as triple negative breast cancer (TNBC), is a proxy for the basal-like subtype (*165, 166, 170*). This resulted in IHC staining of ER, PR, HER2 and the proliferation marker KI67 being used for classification of tumours in these subtypes (*160, 166*).

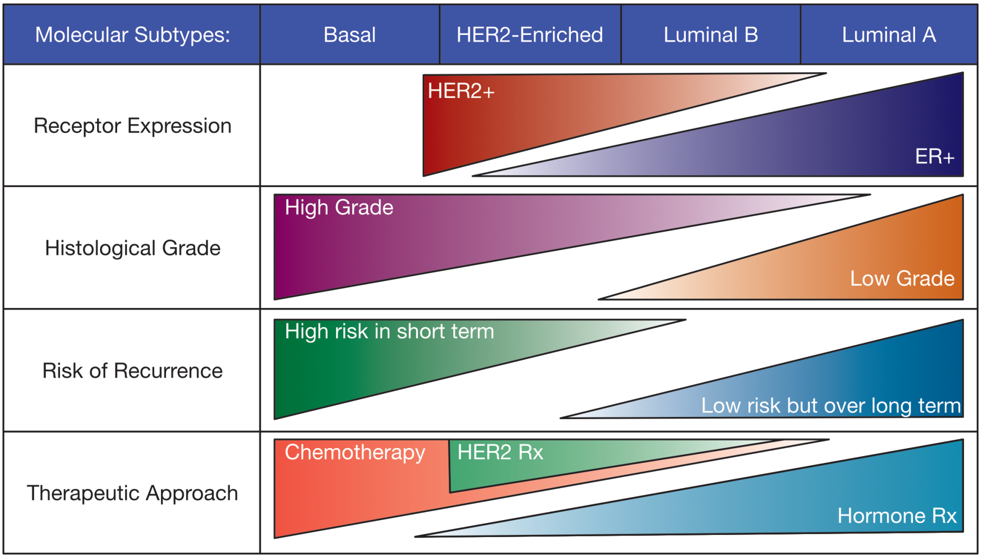


Figure 7 Correlation of clinicopathological features across the molecular subtypes. Based on (*160*).

In addition to the information provided by the Prosigna assay (ROR + PAM50 subtype), a number of different multigene signatures have been developed either for defining subtypes via gene expression profile, or for prognostic or predictive risk predictors (*148, 162, 166, 167, 170*). Rather than relying on single genes such as the IHC-based classification, these tools use a combination of different number of genes to reflect more complex tumour biology (*165*). A few examples include Oncotype DX (*175*), MammaPrint (*176*), the Breast Cancer Index (*177*), EndoPredict (*178, 179*), and the Genomic Grade Index (*180*).

Vallon-Christersson et. al. (*170*) evaluated 19 different multigene signatures using a consecutive observational cohort of 3520 resectable primary breast cancer samples from the south of Sweden (*170*). The main aims were to investigate the association of such signatures to overall survival and assess the classification consensus based on this cohort. The authors conclude that the use of multigene signatures can support clinical decisions but emphasise the need for further development to reach higher consensus (*170*).

Figure 8 gives an overview of different classifications typically used for BC.

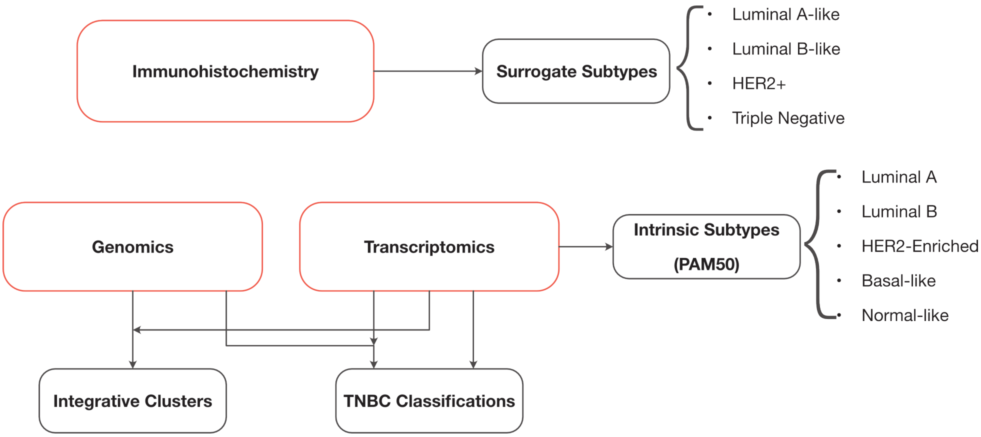


Figure 8 Overview of different classifications of breast cancer. Based on (*160*).

Of note, despite the number of assays available, there is a gap in the clinical validation of such signatures in prospective cohorts (*170*). From the examples given above, i.e., MammaPrint, Oncotype DX, Prosigna, EndoPredict, Breast Cancer Index and Genomic Grade Index, in the context of assisting in decision-making for chemotherapy use in ERpHER2n patients, only the MammaPrint and Oncotype DX assays have concluded prospective validation in the MINDACT and TAILORx trials, respectively (*160*). Another prospective trial, RxPONDER is ongoing, investigating the use of Oncotype DC, as well as the OPTIMA and ASTER70 trials for the use of the Prosigna assay and Genomic Grade Index, respectively (*160*).

Proteomics and Multiomics in Breast Cancer

As discussed in the previous sections, breast cancer is an incredibly heterogeneous disease, and the heterogeneity observed at the histopathological level is also seen at the molecular level (*162*). Despite the progress made, there are still unresolved questions concerning drug resistance and relapse, and the overall burden of the disease is still a considerable challenge (*162, 164*).

Historically, most studies performing a molecular investigation in breast cancer have utilised of genomics and transcriptomics (*163*). That is a consequence of the evolution of omics technologies, with an initial focus being set on genomics after the completion of the “Human Genome Project”, followed by transcriptomics and proteomics (*164*).

Initial efforts focused generating molecular profiles of somatic mutations of human tumours, such as the The Cancer Genome Atlas (TCGA) (*163, 181-183*). Such efforts have contributed immensely to the understanding of molecular mechanisms in cancer biology (*183*). However, the resulting list of significantly mutated genes is relatively short, and most mutations occur either in regulatory or non-coding space rather than in protein-coding sequences (*183*).

One of the issues arising from these strategies is that available databases tend to focus on signalling networks and interactomes derived from genomic data. As a result, biochemistry is often inferred rather than observed, which translates in a knowledge deficiency on how these genomic changes result in phenotypic phenomena, i.e., what drives proteins and phosphoproteins to execute said phenotype (*183, 184*). Therefore, a clear link between the genotype and phenotype must be established in order to improve the translational potential of these molecular information (*183*).

The technology used to acquire proteome data in the TCGA efforts was Reverse-Phase Protein Arrays (RPPA), but this approach is limited by the availability of antibodies (*184*). It also makes it a hypothesis-driven approach given that the specificity of the different antibodies is known, compared to MS-based proteomics, which allows for a hypothesis-generating approach (*65, 69, 76*). With this in mind, the National Cancer Institute (NCI) launched the Clinical Proteomic Tumor Analysis Consortium (CPTAC) to use MS-based proteomics to analyse the TCGA samples (*183*).

The field of proteomics has already made meaningful contributions to the use of biomarkers in breast cancer, with the clearest example being the utilisation of IHC in the investigation of ER, PR and HER2 status for subtyping, being an advantageous strategy considering the availability, reduced costs and familiarity (*160, 165, 168*).

It was discussed in previous chapters that one of the main advantages of performing proteomic studies lies in the close proximity of proteins and the observed phenotype (*14-16*), with proteomics being placed downstream from genomics and offering a platform that explains how changes in the genome can affect functions and phenotypes (*183*). However, considering the complexity and heterogeneity of diseases such as breast cancer, enormous potential is present in harnessing the synergies that the different platforms can provide, allowing for a more complete molecular profile of diseases and patients and facilitating biomarker discovery (*141, 142, 183*).

The potential benefits of LC-MS/MS and multiomics analysis have been discussed in previous chapters. In the context of breast cancer, this has been successfully demonstrated in different publications, e.g., (*135, 184-186*) and in Papers I and IV presented in this thesis.

Besides the use of LC-MS/MS for proteomic profiling of different diseases or patients, MS-based workflows can be used for a myriad of different applications, for instance in Mass Spectrometric Imaging (MSI) (*164, 187, 188*) – through MALDI-MSI, and Desorption Electrospray Ionisation Mass Spectrometric Imaging (DESI-MSI) – which could be utilised in place of techniques such as IHC, as well as in establishing surgical margins, and applications such as drug discovery and development and drug repurposing (*164, 189*).

In the context of this thesis, besides demonstrating that proteomics adds complementary information to transcriptomics data (Paper I), a multiomics approach incorporating transcriptomics, proteomics, phosphoproteomics and immune infiltration estimates is utilised for profiling metastatic processes in IBC-NST and ILC (Paper IV).

From a disease perspective, Paper IV represents, to the best of our knowledge, the most comprehensive multiomics dataset of metastatic processes in ER-positive breast cancer. The efforts presented in this paper highlight possible subtypes with distinct survival and markers of both lymph node involvement and distant metastasis. Although the findings need to be further investigated, they also underscore the potential to classify tumours for possible usage of immunotherapy and adjuvant therapy.

Concluding Remarks and Future Outlook

This thesis begins by bringing to the forefront two very crucial and complementary problems, namely “The Biomarker Problem” and “The Burden of Cancer”.

On one hand, technological advances have allowed insurmountable amounts of molecular and imaging data to be generated, but the lack of clearly defined and relevant clinical questions, and failure to meet rigorous clinical validation have resulted in limited approval of new biomarkers, posing an obstacle for the field of precision medicine.

On the other hand, an ever-growing and aging population, together with increased prevalence of risk factors have contributed to cancer ranking among the two leading causes of death worldwide. These factors also contribute to an increase in disease burden, and highlight the need for improved diagnosis and better, more efficient treatment options.

With these two aspects in mind, it becomes clear that progress in one front is necessary to also progress in the other. Therefore, the aim of this thesis focuses on MS-based proteomics and multiomics efforts for advancing biomarker discovery with the goal of contributing to precision medicine and oncology.

In Paper I, we start with 116 breast cancer samples, representing different molecular subtypes of the disease. We further demonstrate the use of this sample material, which corresponds to flowthroughs after DNA and RNA extraction, for acquiring matching proteomics data. Focusing on biomarker-discovery efforts, label-free proteomics was selected, and different data acquisition strategies were tested, concluding that DIA acquisition with library-free processing via DIA-NN resulted in better coverage and reproducibility. The resulting data was also evaluated in terms of contribution to already available transcriptome data. In this context, we demonstrate the added value in terms of GSEA and a decision tree model for distinguishing the different molecular subtypes. Despite not having a clear clinical question as part of the study, the markers highlighted recapitulated important aspects of the different subtypes, showcasing the importance of proteomics data.

In Paper II, we focus on blood plasma as a source of protein biomarkers. The blood plasma contains a multitude of proteins and is collected via a highly standardised yet minimally invasive procedure, making it a very interesting sample material. These aspects also result in added complexity when analysing such material via standard proteomics workflows, especially due to the high dynamic range. Among the different strategies available for addressing this challenge, we focus on the GPS platform and propose a semi-automated protocol for multiplexed enrichment of plasma peptides, harnessing the complementarity of different affinity binders. We also highlight the potential use of the workflow with other affinity binders, enabling matching complementary enrichments, e.g., enrichment of different PTMs from the same sample material.

In Paper III, we bring forward the prognostic and predictive value of immune infiltration in cancer. Although direct methods for estimation of immune infiltration exist, e.g., IHC, the use of omics data pose an interesting alternative. Most commonly, transcriptomics is used for this purpose, although proteomics data could potentially be more accurate, given its closer association with the phenotype. With that in mind, we compared different deconvolution algorithms and propose preprocessing steps concerning biological ID handling, normalisation and handling of missing data/imputation, culminating in the development of the proteoDeconv R package with the goal of streamlining the use of proteomics data for immune deconvolution. Factors such as signature matrix used and protein content of different immune cells are also important to consider, and although these were not investigated in depth in this context, developments in single-cell proteomics would directly contribute to better tailoring immune deconvolution algorithms to be used with proteomics data.

In Paper IV, we adopt a multiomics approach for the profiling of metastatic processes in ER-positive breast cancer, combining LC-MS-based proteomics and phosphoproteomics with immune infiltration estimates and transcriptome data. Through consensus clustering, we identified six potential subtypes with different expression patterns, immune infiltration and survival. In terms of the immune component, we discuss a potential role of interferon signalling in promoting an exhausted phenotype, one which could potentially benefit from immunotherapy. MOFA was used in combination with differential expression analysis to identify markers of lymph node and distant metastasis, and we discuss the potential role of some markers as well as possible therapeutic interventions. Further investigation of the reported markers is important to establish a causal relationship to the metastatic processes and potential prognostic and predictive roles, but they show promising leads to further develop precision oncology in ER-positive breast cancer.

In many ways, the work presented in this thesis is the summation of all the parts developed in the other papers. It draws from the experience in developing robust automated protocols for reproducible sample processing, highlighted in Papers I and

II, with more advanced data analysis pipelines implemented in Papers III and IV. The findings in this thesis underscore the importance of robust sample processing and the potential of multiomics efforts for precision medicine.

Considering the advances achieved in high-throughput molecular technologies and decreasing associated costs, it has never been easier to implement an integrative approach to biomarker discovery and validation. Diseases such as cancer make evident the molecular complexity behind such pathological processes, rendering it not only more difficult, but, to certain extent, unlikely to be fully understood from a single-omics perspective.

Historically, omics technologies have made use of genomics and transcriptomics. Consequently, the vast majority of databases are annotated in terms of genes and transcripts, often inferring function, despite proteins and their proteoforms being responsible for most functions in a living organism. Considering different mechanisms of transcriptional regulation, post-transcriptional and post-translational modifications, mapping genes to proteins is a lot more complex than mapping proteins to genes. For that reason, and since proteoforms can serve as direct measures of function, it would be beneficial to adopt a proteoform-centric base for functional annotation.

Finally, in order to alleviate the issues posed by the biomarker problem and the burden of cancer, special consideration must be put into study designs. Since laboratory research and population research do not overlap, it is essential to have strong collaborations in order to be able to translate knowledge from laboratory to clinical research and from research to clinical implementation.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to everyone who has supported me throughout this journey. This would not have been possible without the encouragement, guidance, support and love from so many of you.

I started this journey in 2018 very excited to join a programme focused on accelerating innovation and clinical implementation in oncology. It has been an unforgettable ride, full of ups and downs, twists and turns, dark times and happy times, harmony and chaos. It has made me a better person, and for that I am very grateful.

To my supervisor, Fredrik. Thank you for believing in me and for giving me this opportunity back when I knew very little about bioinformatics and mass spectrometry. You have taught me many lessons over the years. Thank you for our discussions, for your invaluable comments and suggestions. Early on you taught me that I shouldn't get too paranoid about getting papers out as fast as possible, that things would fall into place in due time, and I believe that has kept me grounded during difficult times. Thank you for always having your door open for me, even if I came around an infinite number of times on the same day. You always met me with a good humour and an air of serenity, and that is truly inspirational.

To my co-supervisors and collaborators, your additional perspectives, support and guidance have been invaluable. Your expertise and collaborative spirit have greatly contributed to my personal development and to that of this project. A special thanks to you, Valentina. You were the first with whom I worked closely in the lab and who taught me about instrument usage and maintenance. You were always a text message away and we spent many hours laughing and helping each other.

To everyone in the Proteomics group and my colleagues at the Department of Immunotechnology, old and new, thank you for the happy and positive environment you created. Your camaraderie, support and friendship have created a wonderful environment, and I am very proud to be part of it. Your enthusiasm, kindness, and sense of humour have brought light to the darkest days. Whether it was sharing a laugh during fika and afterworks, offering a listening ear, or providing words of encouragement, you have all contributed to making this journey more enjoyable.

To everyone at the Onkologiavdelning 87, Onkologimottagning, Hematologiavdelning 4 and Intensivvårdsavdelning. Thank you for you care and support. A special thanks to Emmy, Olof, Marie, Maria, Pernilla and Nenad for the heartfelt conversations, for the company and for making me feel supported and at ease during the most difficult times I encountered.

To my friends. Some of you I brought from Brazil, some I met in Sweden, and some are scattered across the world. Thank you for your enormous support and for keeping me grounded and motivated. Thank you for accepting me as my nerdy paranoid overthinker self, for all the moments around the table, for all the laughs and for the countless hours on video call when I felt the most hopeless and lonely. Your presence in my life has been a source of immense joy and strength, and I am very grateful for each and every one of you. I look forward to sharing many more moments together.

To my family, thank you for your unwavering support and belief in me. Your constant encouragement and sacrifices have been the foundation of my success. Your endless patience and love have been my guiding light. Your support has been a constant source of motivation. Thank you from the bottom of my heart.

Last but not least, to you, Beb. You have been my rock and my confidant. Thank you for your endless support and for being my anchor in the storm. Your love has been a beacon of light, guiding me through the darkest moments. Your sense of humour has brought laughter and joy when I needed the most. Your presence has made the journey not only (bear)able but also beautiful. You have made every challenge worth facing and every milestone worth celebrating. Your support has made all the difference, and I am very glad to have you by my side. I look forward to sharing many dreams, achievements and adventures with you. Älskar dig.

References

1. J. K. Aronson, R. E. Ferner, Biomarkers-A General Review. *Curr Protoc Pharmacol* **76**, 9 23 21-29 23 17 (2017).
2. G. Biomarkers Definitions Working, Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical pharmacology and therapeutics* **69**, 89-95 (2001).
3. F.-N. B. W. Group. (Silver Spring (MD), Bethesda (MD), 2016).
4. R. M. Califf, Biomarker definitions and their applications. *Experimental biology and medicine* **243**, 213-221 (2018).
5. A. G. Paulovich, J. R. Whiteaker, A. N. Hoofnagle, P. Wang, The interface between biomarker discovery and clinical validation: The tar pit of the protein biomarker pipeline. *Proteomics. Clinical applications* **2**, 1386-1402 (2008).
6. V. B. Kraus, Biomarkers as drug development tools: discovery, validation, qualification and use. *Nature reviews. Rheumatology* **14**, 354-362 (2018).
7. M. J. Duffy, N. O'Donovan, E. McDermott, J. Crown, Validated biomarkers: The key to precision treatment in patients with breast cancer. *Breast* **29**, 192-201 (2016).
8. J. L. Jameson, D. L. Longo, Precision medicine--personalized, problematic, and promising. *The New England journal of medicine* **372**, 2229-2234 (2015).
9. C. A. Borrebaeck, Precision diagnostics: moving towards protein biomarker signatures of clinical utility in cancer. *Nature reviews. Cancer* **17**, 199-204 (2017).
10. L. Knowles, W. Luth, T. Bubela, Paving the road to personalized medicine: recommendations on regulatory, intellectual property and reimbursement challenges. *J Law Biosci* **4**, 453-506 (2017).
11. H. Sung *et al.*, Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a cancer journal for clinicians* **71**, 209-249 (2021).
12. F. Bray, M. Laversanne, E. Weiderpass, I. Soerjomataram, The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer* **127**, 3029-3030 (2021).
13. J. Ferlay *et al.* (International Agency for Research on Cancer, Lyon, France, 2024), vol. 2024.
14. N. Rifai, M. A. Gillette, S. A. Carr, Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nature biotechnology* **24**, 971-983 (2006).

15. M. Mann, C. Kumar, W. F. Zeng, M. T. Strauss, Artificial intelligence for proteomics and biomarker discovery. *Cell Syst* **12**, 759-770 (2021).

16. N. L. Anderson, The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. *Clinical chemistry* **56**, 177-185 (2010).

17. M. Tyers, M. Mann, From genomics to proteomics. *Nature* **422**, 193-197 (2003).

18. M. R. Wilkins *et al.*, From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Bio/technology* **14**, 61-65 (1996).

19. L. C. Gillet, A. Leitner, R. Aebersold, Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. *Annu Rev Anal Chem (Palo Alto Calif)* **9**, 449-472 (2016).

20. S. Al-Amrani, Z. Al-Jabri, A. Al-Zaabi, J. Alshekaili, M. Al-Khabori, Proteomics: Concepts and applications in human medicine. *World journal of biological chemistry* **12**, 57-69 (2021).

21. A. Sinha, M. Mann, A beginner's guide to mass spectrometry–based proteomics. *The Biochemist* **42**, 64-69 (2020).

22. R. Aebersold, M. Mann, Mass spectrometry-based proteomics. *Nature* **422**, 198-207 (2003).

23. R. Aebersold, D. R. Goodlett, Mass spectrometry in proteomics. *Chemical reviews* **101**, 269-295 (2001).

24. M. Mann, R. C. Hendrickson, A. Pandey, Analysis of proteins and proteomes by mass spectrometry. *Annu Rev Biochem* **70**, 437-473 (2001).

25. P. R. Graves, T. A. Haystead, Molecular biologist's guide to proteomics. *Microbiology and molecular biology reviews : MMBR* **66**, 39-63; table of contents (2002).

26. A. Makarov *et al.*, Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Analytical chemistry* **78**, 2113-2120 (2006).

27. A. Makarov, Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Analytical chemistry* **72**, 1156-1162 (2000).

28. H. I. Stewart *et al.*, Parallelized Acquisition of Orbitrap and Astral Analyzers Enables High-Throughput Quantitative Analysis. *Analytical chemistry* **95**, 15656-15664 (2023).

29. ThermoScientific, Product Specifications: Orbitrap Astral Mass Spectrometer. 2023.

30. ThermoScientific, Exactive Series Operating Manual. 2017.

31. B. T. Chait, Chemistry. Mass spectrometry: bottom-up or top-down? *Science* **314**, 65-66 (2006).

32. C. D. Kelstrup *et al.*, Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics. *Journal of proteome research* **17**, 727-738 (2018).

33. J. C. Rogers, R. D. Bomgarden, in *Modern Proteomics – Sample Preparation, Analysis and Practical Applications,* H. Mirzaei, M. Carrasco, Eds. (Springer International Publishing, Cham, 2016), pp. 43-62.

34. D. P. Donnelly *et al.*, Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. *Nature methods* **16**, 587-594 (2019).

35. X. Han, M. Jin, K. Breuker, F. W. McLafferty, Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons. *Science* **314**, 109-112 (2006).

36. N. L. Kelleher *et al.*, Top Down versus Bottom Up Protein Characterization by Tandem High-Resolution Mass Spectrometry. *Journal of the American Chemical Society* **121**, 806-812 (1999).

37. S. R. Shuken, An Introduction to Mass Spectrometry-Based Proteomics. *Journal of proteome research* **22**, 2151-2171 (2023).

38. L. M. Smith, N. L. Kelleher, P. Consortium for Top Down, Proteoform: a single term describing protein complexity. *Nature methods* **10**, 186-187 (2013).

39. L. M. Smith, N. L. Kelleher, Proteoforms as the next proteomics currency. *Science* **359**, 1106-1107 (2018).

40. A. I. Nesvizhskii, R. Aebersold, Interpretation of shotgun proteomic data: the protein inference problem. *Molecular & cellular proteomics : MCP* **4**, 1419-1440 (2005).

41. M. Claassen, Inference and validation of protein identifications. *Molecular & cellular proteomics : MCP* **11**, 1097-1104 (2012).

42. R. Aebersold, M. Mann, Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347-355 (2016).

43. A. C. Peterson, J. D. Russell, D. J. Bailey, M. S. Westphall, J. J. Coon, Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Molecular & cellular proteomics : MCP* **11**, 1475-1488 (2012).

44. P. Picotti, R. Aebersold, Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nature methods* **9**, 555-566 (2012).

45. C. Gotti *et al.*, Extensive and Accurate Benchmarking of DIA Acquisition Methods and Software Tools Using a Complex Proteomic Standard. *Journal of proteome research* **20**, 4801-4814 (2021).

46. J. D. Chapman, D. R. Goodlett, C. D. Masselon, Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. *Mass Spectrom Rev* **33**, 452-470 (2014).

47. S. Purvine, J. T. Eppel, E. C. Yi, D. R. Goodlett, Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *Proteomics* **3**, 847-850 (2003).

48. J. C. Silva *et al.*, Quantitative proteomic analysis by accurate mass retention time pairs. *Analytical chemistry* **77**, 2187-2200 (2005).

49. A. A. Ramos, H. Yang, L. E. Rosen, X. Yao, Tandem parallel fragmentation of peptides for mass spectrometry. *Analytical chemistry* **78**, 6391-6397 (2006).

50. T. Geiger, J. Cox, M. Mann, Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Molecular & cellular proteomics : MCP* **9**, 2252-2261 (2010).

51. J. D. Venable, M. Q. Dong, J. Wohlschlegel, A. Dillin, J. R. Yates, Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature methods* **1**, 39-45 (2004).

52. A. Panchaud *et al.*, Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. *Analytical chemistry* **81**, 6481-6488 (2009).

53. P. C. Carvalho *et al.*, XDIA: improving on the label-free data-independent analysis. *Bioinformatics* **26**, 847-848 (2010).

54. L. C. Gillet *et al.*, Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Molecular & cellular proteomics : MCP* **11**, O111 016717 (2012).

55. C. R. Weisbrod, J. K. Eng, M. R. Hoopmann, T. Baker, J. E. Bruce, Accurate peptide fragment mass analysis: multiplexed peptide identification and quantification. *Journal of proteome research* **11**, 1621-1632 (2012).

56. C. Ludwig *et al.*, Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Molecular systems biology* **14**, e8126 (2018).

57. C. T. Walsh, S. Garneau-Tsodikova, G. J. Gatto, Jr., Protein posttranslational modifications: the chemistry of proteome diversifications. *Angewandte Chemie* **44**, 7342-7372 (2005).

58. M. Mann, O. N. Jensen, Proteomic analysis of post-translational modifications. *Nature biotechnology* **21**, 255-261 (2003).

59. M. Ke *et al.*, Identification, Quantification, and Site Localization of Protein Posttranslational Modifications via Mass Spectrometry-Based Proteomics. *Advances in experimental medicine and biology* **919**, 345-382 (2016).

60. K. Sharma *et al.*, Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell reports* **8**, 1583-1594 (2014).

61. D. Ochoa *et al.*, The functional landscape of the human phosphoproteome. *Nature biotechnology* **38**, 365-373 (2020).

62. P. V. Hornbeck *et al.*, PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic acids research* **43**, D512-520 (2015).

63. T. E. Thingholm, M. R. Larsen, Phosphopeptide Enrichment by Immobilized Metal Affinity Chromatography. *Methods in molecular biology* **1355**, 123-133 (2016).

64. J. Cox *et al.*, Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & cellular proteomics : MCP* **13**, 2513-2526 (2014).

65. N. Olsson *et al.*, Proteomic analysis and discovery using affinity proteomics and mass spectrometry. *Molecular & cellular proteomics : MCP* **10**, M110.003962 (2011).

66. M. Sandin, A. Chawade, F. Levander, Is label-free LC-MS/MS ready for biomarker discovery? *Proteomics. Clinical applications* **9**, 289-294 (2015).

67. B. Kim *et al.*, Affinity enrichment for mass spectrometry: improving the yield of low abundance biomarkers. *Expert Rev Proteomics* **15**, 353-366 (2018).

68. E. Gianazza, I. Miller, L. Palazzolo, C. Parravicini, I. Eberini, With or without you - Proteomics with or without major plasma/serum proteins. *Journal of proteomics* **140**, 62-80 (2016).

69. B. Ayoglu *et al.*, Systematic antibody and antigen-based proteomic profiling with microarrays. *Expert review of molecular diagnostics* **11**, 219-234 (2011).

70. E. Assarsson *et al.*, Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PloS one* **9**, e95192 (2014).

71. J. C. Rohloff *et al.*, Nucleic Acid Ligands With Protein-like Side Chains: Modified Aptamers and Their Use as Diagnostic and Therapeutic Agents. *Mol Ther Nucleic Acids* **3**, e201 (2014).

72. E. W. Deutsch *et al.*, Advances and Utility of the Human Plasma Proteome. *Journal of proteome research* **20**, 5241-5263 (2021).

73. B. B. Haab, Applications of antibody array platforms. *Current opinion in biotechnology* **17**, 415-421 (2006).

74. A. Carlsson *et al.*, Molecular serum portraits in patients with primary breast cancer predict the development of distant metastases. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 14252-14257 (2011).

75. M. Sanchez-Carbayo, N. D. Socci, J. J. Lozano, B. B. Haab, C. Cordon-Cardo, Profiling bladder cancer using targeted antibody arrays. *The American journal of pathology* **168**, 93-103 (2006).

76. N. Olsson, P. James, C. A. Borrebaeck, C. Wingren, Quantitative proteomics targeting classes of motif-containing peptides using immunoaffinity-based mass spectrometry. *Molecular & cellular proteomics : MCP* **11**, 342-354 (2012).

77. L. Guerrier, P. G. Righetti, E. Boschetti, Reduction of dynamic protein concentration range of biological extracts for the discovery of low-abundance proteins by means of hexapeptide ligand library. *Nature protocols* **3**, 883-890 (2008).

78. P. G. Righetti, E. Boschetti, L. Lomas, A. Citterio, Protein Equalizer Technology : the quest for a "democratic proteome". *Proteomics* **6**, 3980-3992 (2006).

79. N. L. Anderson *et al.*, Mass spectrometric quantitation of peptides and proteins using Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA). *Journal of proteome research* **3**, 235-244 (2004).

80. A. Sall *et al.*, AFFIRM--a multiplexed immunoaffinity platform that combines recombinant antibody fragments and LC-SRM analysis. *Journal of proteome research* **13**, 5837-5847 (2014).

81. F. Weiss *et al.*, Catch and measure-mass spectrometry-based immunoassays in biomarker research. *Biochimica et biophysica acta* **1844**, 927-932 (2014).

82. O. Poetz, S. Hoeppe, M. F. Templin, D. Stoll, T. O. Joos, Proteome wide screening using peptide affinity capture. *Proteomics* **9**, 1518-1523 (2009).

83. C. Wingren, P. James, C. A. Borrebaeck, Strategy for surveying the proteome using affinity proteomics and mass spectrometry. *Proteomics* **9**, 1511-1517 (2009).

84. N. Olsson *et al.*, Grading breast cancer tissues using molecular portraits. *Molecular & cellular proteomics : MCP* **12**, 3612-3623 (2013).

85. J. Forshed, Experimental Design in Clinical 'Omics Biomarker Discovery. *Journal of proteome research* **16**, 3954-3960 (2017).

86. S. B. Hulley, *Designing clinical research*. (Wolters Kluwer/Lippincott Williams & Wilkins, ed. 4th, 2013).

87. F. E. Ahmed, Sample preparation and fractionation for proteome analysis and cancer biomarker discovery by mass spectrometry. *Journal of separation science* **32**, 771-798 (2009).

88. L. H. Saal *et al.*, The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. *Genome medicine* 7, 20 (2015).

89. S. Mosquim Junior, V. Siino, L. Ryden, J. Vallon-Christersson, F. Levander, Choice of High-Throughput Proteomics Method Affects Data Integration with Transcriptomics and the Potential Use in Biomarker Discovery. *Cancers (Basel)* **14**, (2022).

90. J. R. Wisniewski, A. Zougman, N. Nagaraj, M. Mann, Universal sample preparation method for proteome analysis. *Nature methods* **6**, 359-362 (2009).

91. A. Zougman, P. J. Selby, R. E. Banks, Suspension trapping (STrap) sample preparation method for bottom-up proteomics analysis. *Proteomics* **14**, 1006-1000 (2014).

92. C. S. Hughes *et al.*, Ultrasensitive proteome analysis using paramagnetic bead technology. *Molecular systems biology* **10**, 757 (2014).

93. T. S. Batth *et al.*, Protein Aggregation Capture on Microparticles Enables Multipurpose Proteomics Sample Preparation. *Molecular & cellular proteomics : MCP* **18**, 1027-1035 (2019).

94. D. C. Montgomery, *Design and Analysis of Experiments*. (John Wiley & Sons, Inc., ed. 9, 2017).

95. H. T. Tan, Y. H. Lee, M. C. Chung, Cancer proteomics. *Mass Spectrom Rev* **31**, 583-605 (2012).

96. N. L. Anderson, N. G. Anderson, The human plasma proteome: history, character, and diagnostic prospects. *Molecular & cellular proteomics : MCP* **1**, 845-867 (2002).

97. C. D. King, K. L. Kapp, A. B. Arul, M. J. Choi, R. A. S. Robinson, Advancements in automation for plasma proteomics sample preparation. *Mol Omics* **18**, 828-839 (2022).

98. B. He, Z. Huang, C. Huang, E. C. Nice, Clinical applications of plasma proteomics and peptidomics: Towards precision medicine. *Proteomics. Clinical applications*, e2100097 (2022).

99. R. Lou *et al.*, Benchmarking commonly used software suites and analysis workflows for DIA proteomics and phosphoproteomics. *Nature communications* **14**, 94 (2023).

100. K. Frohlich *et al.*, Benchmarking of analysis strategies for data-independent acquisition proteomics using a large-scale dataset comprising inter-patient heterogeneity. *Nature communications* **13**, 2622 (2022).

101. V. Demichev, C. B. Messner, S. I. Vernardis, K. S. Lilley, M. Ralser, DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature methods* **17**, 41-44 (2020).

102. M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, B. Kuster, Quantitative mass spectrometry in proteomics: a critical review. *Analytical and bioanalytical chemistry* **389**, 1017-1031 (2007).

103. B. C. Searle *et al.*, Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nature communications* **9**, 5128 (2018).

104. Y. S. Ting *et al.*, Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. *Molecular & cellular proteomics : MCP* **14**, 2301-2307 (2015).

105. J. K. Eng, B. Fischer, J. Grossmann, M. J. Maccoss, A fast SEQUEST cross correlation algorithm. *Journal of proteome research* 7, 4598-4602 (2008).

106. T. Koenig *et al.*, Robust prediction of the MASCOT score for an improved quality assessment in mass spectrometric proteomics. *Journal of proteome research* 7, 3708-3717 (2008).

107. R. Craig, R. C. Beavis, TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466-1467 (2004).

108. J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* **26**, 1367-1372 (2008).

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

110. S. Kim, P. A. Pevzner, MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature communications* **5**, 5277 (2014).

111. L. Y. Geer *et al.*, Open mass spectrometry search algorithm. *Journal of proteome research* **3**, 958-964 (2004).

112. R. Bruderer *et al.*, Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Molecular & cellular proteomics : MCP* **14**, 1400-1410 (2015).

113. H. L. Rost *et al.*, OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature biotechnology* **32**, 219-223 (2014).

114. C. C. Tsou *et al.*, DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature methods* **12**, 258-264, 257 p following 264 (2015).

115. P. Sinitcyn *et al.*, MaxDIA enables library-based and library-free data-independent acquisition proteomics. *Nature biotechnology* **39**, 1563-1573 (2021).

116. E. Govaert *et al.*, Comparison of fractionation proteomics for local SWATH library building. *Proteomics* **17**, (2017).

117. K. Barkovits *et al.*, Reproducibility, Specificity and Accuracy of Relative Quantification Using Spectral Library-based Data-independent Acquisition. *Molecular & cellular proteomics : MCP* **19**, 181-197 (2020).

118. L. K. Pino, S. C. Just, M. J. MacCoss, B. C. Searle, Acquiring and Analyzing Data Independent Acquisition Proteomics Experiments without Spectrum Libraries. *Molecular & cellular proteomics : MCP* **19**, 1088-1103 (2020).

119. Y. Yang *et al.*, In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nature communications* **11**, 146 (2020).

120. S. Gessulat *et al.*, Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature methods* **16**, 509-518 (2019).

121. B. C. Searle *et al.*, Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nature communications* **11**, 1548 (2020).

122. M. Bantscheff, S. Lemeer, M. M. Savitski, B. Kuster, Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Analytical and bioanalytical chemistry* **404**, 939-965 (2012).

123. J. Willforss, A. Chawade, F. Levander, NormalyzerDE: Online Tool for Improved Normalization of Omics Expression Data and High-Sensitivity Differential Expression Analysis. *Journal of proteome research* **18**, 732-740 (2019).

124. A. Chawade, E. Alexandersson, F. Levander, Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. *Journal of proteome research* **13**, 3114-3120 (2014).

125. W. Huang da, B. T. Sherman, R. A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* **37**, 1-13 (2009).

126. A. Liberzon *et al.*, The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417-425 (2015).

127. A. Liberzon *et al.*, Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739-1740 (2011).

128. A. Subramanian *et al.*, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545-15550 (2005).

129. E. I. Boyle *et al.*, GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**, 3710-3715 (2004).

130. M. Ashburner *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25-29 (2000).

131. T. Wu *et al.*, clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**, 100141 (2021).

132. K. Krug *et al.*, A Curated Resource for Phosphosite-specific Signature Analysis. *Molecular & cellular proteomics : MCP* **18**, 576-593 (2019).

133. M. D. Wilkerson, D. N. Hayes, ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572-1573 (2010).

134. S. Monti, P. Tamayo, J. Mesirov, T. Golub, Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning* **52**, 91-118 (2003).

135. H. J. Johansson *et al.*, Breast cancer quantitative proteome and proteogenomic landscape. *Nature communications* **10**, 1600 (2019).

136. J. Lehtiö *et al.*, Proteogenomics of non-small cell lung cancer reveals molecular subtypes associated with specific therapeutic targets and immune-evasion mechanisms. *Nature Cancer* **2**, 1224-1242 (2021).

137. K. Asleh *et al.*, Proteomic analysis of archival breast cancer clinical specimens identifies biological subtypes with distinct survival outcomes. *Nature communications* **13**, 896 (2022).

138. M. Olivier, R. Asmis, G. A. Hawkins, T. D. Howard, L. A. Cox, The Need for Multi-Omics Biomarker Signatures in Precision Medicine. *International journal of molecular sciences* **20**, (2019).

139. M. Zitnik *et al.*, Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. *Inf Fusion* **50**, 71-91 (2019).

140. M. Picard, M. P. Scott-Boyer, A. Bodein, O. Perin, A. Droit, Integration strategies of multi-omics data for machine learning analysis. *Comput Struct Biotechnol J* **19**, 3735-3746 (2021).

141. E. Athieniti, G. M. Spyrou, A guide to multi-omics data collection and integration for translational medicine. *Comput Struct Biotechnol J* **21**, 134-149 (2023).

142. D. Chicco, F. Cumbo, C. Angione, Ten quick tips for avoiding pitfalls in multi-omics data integration analyses. *PLoS computational biology* **19**, e1011224 (2023).

143. A. W. Kurian, BRCA1 and BRCA2 mutations across race and ethnicity: distribution and clinical implications. *Current opinion in obstetrics & gynecology* **22**, 72-78 (2010).

144. N. D'Agostino, W. Li, D. Wang, High-throughput transcriptomics. *Scientific reports* **12**, 20313 (2022).

145. J. Botling *et al.*, Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis, and tissue microarray validation. *Clinical cancer research : an official journal of the American Association for Cancer Research* **19**, 194-204 (2013).

146. W. Li, R. Wang, Z. Yan, L. Bai, Z. Sun, High accordance in prognosis prediction of colorectal cancer across independent datasets by multi-gene module expression profiles. *PloS one* **7**, e33653 (2012).

147. C. W. Duarte *et al.*, Expression signature of IFN/STAT1 signaling genes predicts poor survival outcome in glioblastoma multiforme in a subtype-specific manner. *PloS one* **7**, e29653 (2012).

148. A. Prat, M. J. Ellis, C. M. Perou, Practical implications of gene-expression-based assays for breast oncologists. *Nature reviews. Clinical oncology* **9**, 48-57 (2011).

149. R. Argelaguet, A. S. E. Cuomo, O. Stegle, J. C. Marioni, Computational principles and challenges in single-cell data integration. *Nature biotechnology* **39**, 1202-1215 (2021).

150. H. HOTELLING, RELATIONS BETWEEN TWO SETS OF VARIATES*. *Biometrika* **28**, 321-377 (1936).

151. E. F. Lock, K. A. Hoadley, J. S. Marron, A. B. Nobel, Joint and Individual Variation Explained (Jive) for Integrated Analysis of Multiple Data Types. *Ann Appl Stat* **7**, 523-542 (2013).

152. R. Argelaguet *et al.*, MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology* **21**, 111 (2020).

153. R. Argelaguet *et al.*, Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology* **14**, e8124 (2018).

154. A. Singh *et al.*, DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* **35**, 3055-3062 (2019).

155. F. Bray *et al.*, Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **74**, 229-263 (2024).

156. F. Guida *et al.*, Global and regional estimates of orphans attributed to maternal cancer mortality in 2020. *Nature medicine* **28**, 2563-2572 (2022).

157. L. Wilkinson, T. Gathani, Understanding breast cancer as a global health concern. *Br J Radiol* **95**, 20211033 (2022).

158. M. Arnold *et al.*, Current and future burden of breast cancer: Global statistics for 2020 and 2040. *Breast* **66**, 15-23 (2022).

159. A. C. Kataki, D. Barmon, *Fundamentals in Gynaecologic Malignancy.* (Springer Nature Singapore, 2023).

160. W. H. O. C. W. H. O. C. T. E. Board, I. A. f. R. o. Cancer, *WHO Classification of Breast Tumours: WHO Classification of Tumours, Volume 2.* (World Health Organization, 2019).

161. B. O. Anderson *et al.*, The Global Breast Cancer Initiative: a strategic collaboration to strengthen health care for non-communicable diseases. *The lancet oncology* **22**, 578-581 (2021).

162. R. Benacka, D. Szaboova, Z. Gulasova, Z. Hertelyova, J. Radonak, Classic and New Markers in Diagnostics and Classification of Breast Cancer. *Cancers (Basel)* **14**, (2022).

163. N. Cancer Genome Atlas, Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70 (2012).

164. R. N. Behera, V. S. Bisht, K. Giri, K. Ambatipudi, Realm of proteomics in breast cancer management and drug repurposing to alleviate intricacies of treatment. *Proteomics. Clinical applications* **17**, e2300016 (2023).

165. E. A. Rakha, G. M. Tse, C. M. Quinn, An update on the pathological classification of breast cancer. *Histopathology* **82**, 5-16 (2023).

166. S. Veerla, L. Hohmann, D. F. Nacer, J. Vallon-Christersson, J. Staaf, Perturbation and stability of PAM50 subtyping in population-based primary invasive breast cancer. *NPJ Breast Cancer* **9**, 83 (2023).

167. M. J. Duffy, S. Walsh, E. W. McDermott, J. Crown, Biomarkers in Breast Cancer: Where Are We and Where Are We Going? *Adv Clin Chem* **71**, 1-23 (2015).

168. C. Mueller, A. Haymond, J. B. Davis, A. Williams, V. Espina, Protein biomarkers for subtyping breast cancer and implications for future research. *Expert Rev Proteomics* **15**, 131-152 (2018).

169. A. N. Neagu *et al.*, Proteomics and its applications in breast cancer. *Am J Cancer Res* **11**, 4006-4049 (2021).

170. J. Vallon-Christersson *et al.*, Cross comparison and prognostic assessment of breast cancer multigene signatures in a large population-based contemporary clinical series. *Scientific reports* **9**, 12184 (2019).

171. J. S. Parker *et al.*, Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **27**, 1160-1167 (2009).

172. C. M. Perou *et al.*, Molecular portraits of human breast tumours. *Nature* **406**, 747-752 (2000).

173. C. Sotiriou *et al.*, Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 10393-10398 (2003).

174. J. H. Norum, K. Andersen, T. Sorlie, Lessons learned from the intrinsic subtypes of breast cancer in the quest for precision therapy. *The British journal of surgery* **101**, 925-938 (2014).

175. S. Paik *et al.*, A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *The New England journal of medicine* **351**, 2817-2826 (2004).

176. M. Kittaneh, A. J. Montero, S. Gluck, Molecular profiling for breast cancer: a comprehensive review. *Biomark Cancer* **5**, 61-70 (2013).

177. J. M. S. Bartlett *et al.*, Breast Cancer Index and prediction of benefit from extended endocrine therapy in breast cancer patients treated in the Adjuvant Tamoxifen-To Offer More? (aTTom) trial. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO* **30**, 1776-1783 (2019).

178. M. Filipits *et al.*, A new molecular predictor of distant recurrence in ER-positive, HER2-negative breast cancer adds independent information to conventional clinical risk factors. *Clinical cancer research : an official journal of the American Association for Cancer Research* **17**, 6012-6020 (2011).

179. I. Sestak *et al.*, Prognostic Value of EndoPredict in Women with Hormone Receptor-Positive, HER2-Negative Invasive Lobular Breast Cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **26**, 4682-4687 (2020).

180. C. Sotiriou *et al.*, Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute* **98**, 262-272 (2006).

181. N. Cancer Genome Atlas Research *et al.*, The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics* **45**, 1113-1120 (2013).

182. M. J. Ellis, C. M. Perou, The genomic landscape of breast cancer as a therapeutic roadmap. *Cancer Discov* **3**, 27-34 (2013).

183. M. J. Ellis *et al.*, Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discov* **3**, 1108-1112 (2013).

184. P. Mertins *et al.*, Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55-62 (2016).

185. S. Tyanova *et al.*, Proteomic maps of breast cancer subtypes. *Nature communications* **7**, 10259 (2016).

186. T. De Marchi *et al.*, Proteogenomic Workflow Reveals Molecular Phenotypes Related to Breast Cancer Mammographic Appearance. *Journal of proteome research* **20**, 2983-3001 (2021).

187. M. N. Stillger, M. J. Li, P. Honscheid, C. von Neubeck, M. C. Foll, Advancing rare cancer research by MALDI mass spectrometry imaging: Applications, challenges, and future perspectives in sarcoma. *Proteomics*, e2300001 (2024).

188. D. Pietkiewicz *et al.*, Mass spectrometry imaging in gynecological cancers: the best is yet to come. *Cancer cell international* **22**, 414 (2022).

189. F. Meissner, J. Geddes-McAlister, M. Mann, M. Bantscheff, The emerging role of mass spectrometry-based proteomics in drug discovery. *Nature reviews. Drug discovery* **21**, 637-654 (2022).

Towards Precision Oncology

With cancer incidence and mortality rising, the need for individualised treatment has never been greater. Recent advances in omics technologies have supported a paradigm shift from traditional classification of diseases to personalised medicine. To achieve this, however, novel biomarkers are needed.

Given the extensive clinical use of proteins and their close proximity to disease phenotypes, their holistic analysis could contribute to the development of better biomarkers. Moreover, as cancer is a highly complex disease, combining multiple omics could help create a more complete map of the disease.

This thesis explores different strategies incorporating mass spectrometry-based proteomics and automation for biomarker discovery in blood plasma and breast cancer biopsies. A multiomics approach to data analysis is adopted, yielding promising results for breast cancer by identifying potential subtypes with differential immune infiltration and markers associated with metastatic processes.

Printed by Media-Tryck, Lund 2024 NORDIC SWAN ECOLABEL 3041 0903

LUND
UNIVERSITY

Department of Immunotechnology
Faculty of Engineering

ISBN 978-91-8104-115-6

9 789181 041156