On the application and validation of multiplexed affinity assays

TEA DODIG-CRNKOVIĆ

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Technology in Biotechnology on https://kth-se.zoom.us/j/64701056914, Friday the 2nd October 2020, at 1:00 p.m
To my parents, Ivica and Gordana
Abstract

Proteins are essential macromolecules that carry out complex functions in human cells, tissues, and organs. They regulate a diverse set of biological processes and protect against pathogens. However, dysregulation or malformation of proteins can cause disease. By characterizing proteins in health and disease, we can gain insights into disease aetiology and identify druggable targets to treat disorders. By bringing protein discoveries from the research lab into clinical practice, protein assays have been and will continue to be important tools for enabling and improving medical decision-making.

The work presented in this thesis concerns both exploratory and targeted affinity-based assays applied for the study of proteins. High-throughput and multiplexed suspension bead arrays have been the primary technology for measuring proteins with antibodies in samples such as human blood. Identification and validation of protein-protein interactions that may provide novel insights into the druggable proteome have also been carried out. Throughout the projects, methods for validating the observations have been pursued and include replication in independent sample sets, as well as the assessment of antibody selectivity via other proteomics assays or orthogonal methods such as genetic associations.

In Paper I, we used multiplexed exploratory antibody arrays comprising almost 1,500 affinity binders to study proteins that circulate in plasma. Here, the focus was to determine the longitudinal variability of proteins. We analysed samples from 101 clinically healthy individuals, collected each third month for one year. The protein data provided insights into inter-individual diversity and the unique molecular fingerprint of each participant. We found that 49% of the studied proteins were stable across one year, as these had low variability in each individual. Eight modules, each containing 11-242 proteins, were found to co-vary across one year. We also found genetic variations to influence 15 of the detected protein profiles and confirmed selected indications in an independent set of 3,000 subjects. In summary, we observed the existence of individual-specific protein profiles and found that short-term and continuous changes occurred in almost every participant.

In Paper II, we investigated blood-derived serum and plasma to identify age-associated proteins. We started from a large set of exploratory antibody bead arrays to screen 156 individuals aged 50-92 years. We found protein profiles of the histidine-rich glycoprotein (HRG) to be significantly associated with age. This association was further corroborated by the analysis of >4,000 individuals from eight additional and independent sets of blood samples. We further validated
the HRG protein profiles by sandwich assays and protein microarrays developed in-house. Comparing genetic data and HRG profiles obtained by two independent antibodies, we observed strong but inverse associations to the genetic variants for two anti-HRG antibodies.

**In Paper III**, we applied multiplexed assays for the detection of autoantibodies against cancer-testis antigens (CTAs) in 133 non-small cell lung cancer (NSCLC) patients. We found reactivity against 29 unique CTAs exclusively in cases, compared to 57 matched controls with benign lung diseases. The presence of six CTAs was further confirmed in an independent set of 34 NSCLC cases. Analysis of longitudinal samples from seven patients demonstrated that the presence of CTA autoantibodies was stable over time for each of the individuals.

**In Paper IV**, we developed a novel multiplexed sandwich-immunoassay for the detection of interaction partners to G-protein coupled receptors (GPCRs). This pharmaceutically important family of membrane proteins is believed to be regulated by another group of receptor activity-modulating proteins (RAMPs) by the formation of protein complexes. We studied cell lysates expressing combinations of 23 GPCRs with three RAMPs. We confirmed most of the previously reported interaction pairs and additionally found evidence for 15 new GPCR-RAMP complexes. All interactions were validated using epitope tags that were engineered onto the proteins. Selected complexes were further validated by *in situ* proximity ligation assays performed in cell membranes.

In summary, the work included in this thesis describes the use of multiplexed affinity-based assays for research within plasma proteomics and the interrogation of protein complexes. The work highlights the method’s potential for the identification of circulating proteins that may aid and add to the current knowledge about human health and disease.

**Keywords**: Affinity Proteomics, Antibody, Autoantibody, Multiplexed Assays, Protein Microarray, Plasma Proteins, Suspension Bead Array
Sammanfattning

Proteiner är makromolekyler som utför essentiella funktioner i människans celler, vävnader, och organ. De deltar i många olika biologiska processer och kan exempelvis skydda mot patogen, så som bakterier och virus. Proteiner är en av kroppens viktigaste byggestenar och förändringar i deras aktivitet kan leda till sjukdom. Genom att studera proteiner i friska och sjuka individer kan vi få en bättre inblick om de bakomliggande molekylära processer som orsakar sjukdom, samt identifiera målprotein för läkemedelsutveckling. Proteinanalys har varit och kommer att fortsätta vara ett viktigt verktyg inom sjukvården.

Arbetet i denna avhandling berör affinitetsbaserade metoder för proteinanalys. Antikropparrayer med hög kapacitet att mäta många proteiner parallellt har tillämpats för att studera proteiner i blod. Dessutom har metoden använts för att identifiera och validera proteininteraktioner som kan vara relevanta för läkemedelsstrategier. Forskningsprojektet som presenteras här har ämnat att validera de undersökningarna som utförts genom att bland annat replikera resultaten i olika patientprov. Antikropparnas selektivitet har bekräftats genom jämförelser av olika proteinanalyser, antikroppsfria metoder, samt genetisk variation.


i blod som uppmättes med två olika antikroppar påvisade även stark association till en genetisk variant av HRG.

I **Paper III** studerades förekomsten av autoantikroppar mot cancer testis antigens (CTAs) i 133 patienter med icke-småcellig lungcancer (NSCLC). Vi identifierade reaktivitet mot 29 unika CTAs i patienter med NSCLC som ej påvisade reaktivitet i 57 prov från individer med godartade lungsjukdomar. Reaktiviteten för sex av dessa CTAs kunde bekräftas i ytterligare 34 NSCLC-patienter. Analys av longitudinella prover från sju patienter påvisade att uttrycket av CTA-autoantikroppar var stabilt under studieperioden för samtliga patienter.

I **Paper IV** utvecklades en ny antikroppsbaserad analysmetod för detektion av proteiner som bildar komplex med G-proteinkopplade receptorer (GPCRs). Denna familj av membranprotein är viktig för många läkemedel. Det finns underlag för att GPCRs funktioner kan regleras via receptor activity-modulating proteins (RAMPs), en annan grupp av proteiner som kan bilda komplex med GPCRs. Med den nya analysmetoden studerade vi 23 stycken GPCRs i kombination med tre stycken RAMPs i cellysat. Vi kunde bekräfta majoriteten av tidigare rapporterade komplex, och kunde vidare identifiera ytterligare 15 helt nya GPCR-RAMP-komplex. Ett urval av interaktionerna validerades med hjälp av epitoptaggar på proteinerna, samt med hjälp av *in situ* proximity ligation assays.

Sammanfattningsvis beskriver arbetet i denna avhandling användning av en affinitetsbaserad metod för proteinforskning i blodplasma, samt undersökning av proteininteraktioner. Studierna belyser metodens potential för identifikation av cirkulerande proteiner som kan komma att ad- dera kunskap till det vi idag känner till om hälsa och sjukdom.
Popular scientific summary

Proteins are small organic molecules that can be found in all cells of the human body. They carry out many important tasks, such as speeding up chemical reactions (enzymes), signalling between cells (hormones), and fighting against foreign bacteria and viruses (antibodies of the immune system). However, if a protein is damaged or mutated it can lead to serious diseases. This is the reason why most drugs act by affecting protein activity.

Proteins can provide insights into what is going on inside of the body, at a given time point. For example, cancer cells can leak proteins into the bloodstream, which can then be detected by medical tests. Similarly, by identifying antibodies that circulate in the blood, a test may uncover if a person has been infected with a particular pathogen, such as the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). For this reason, blood samples are attractive for disease diagnoses as they can give a snapshot of a person’s current state of health. Blood tests are furthermore easy to perform, non-invasive, and can be collected and stored in biobanks for research purposes.

A central part of this thesis concerns the tools that are today used for measuring proteins in human samples. Here, we have utilized reagents (antibodies and antigens) for the detection of proteins. Antibodies are a group of Y-shaped proteins that are part of the immune system, and they are produced to incapacitate harmful molecules. These molecules are known as antigens and often consist of proteins that originate from viruses and bacteria. For research purposes and in medical tests, antibodies can be used as a strategy for capturing particular proteins in a biological specimen. Tools like antibodies and antigens that bind molecules with a certain specificity are known as affinity reagents, and these are widely used for identifying proteins and their functions in cells, tissues, and organs. Since 2003, the Human Protein Atlas (HPA) project has produced a large collection of affinity reagents for the mapping of human proteins. Reagents provided by the HPA have been also been used in the work presented in this thesis.

Although researchers have conducted large-scale and systematic analyses of human proteins, there are still many things about these essential molecules that we do not know in detail. How different are proteins between individuals? Do they change with time, lifestyle, or other factors? Which proteins can tell us if we are about to develop a disease?
To answer some of these questions, we have utilized affinity reagents for assays to study; proteins (Paper I-II) and antibodies (Paper III) that circulate in blood plasma, as well as protein interactions in cells (Paper IV). In Paper I, we followed a group of 101 clinically healthy individuals during one year and measured their proteins in the blood. We observed that each individual in the study had a unique, personal protein pattern – a protein “fingerprint” – that was retained throughout a whole year. In Paper II, we studied proteins in plasma and serum from 156 elderly individuals to identify proteins that are related to ageing. We found one protein of particular interest that we further validated in >4,000 individuals. In Paper III, we focused on the detection of antibodies that may appear due to cancer cell antigens. We analysed blood plasma from 133 lung cancer patients and validated six antigens of interest in additional sets of independent samples. In Paper IV, we developed a novel antibody-based method for detecting protein complexes. We found evidence for 15 complexes that have not been previously described, and these new insights could be valuable for future drug development.

In conclusion, the work presented in this thesis describes different applications of a protein analysis method for researching proteins. The studies highlight the method’s potential for the detection of proteins and protein complexes that may advance our current knowledge about human health and disease.
This thesis will be defended **October 2nd 2020 at 13:00**, for the degree of Doctor of Technology in Biotechnology. With regard to COVID-19, the defense will be viewable online using the following Zoom link:

https://kth-se.zoom.us/j/64701056914

The event will be held for invited attendees in in Air & Fire, Science For Life Laboratory, Tomtebodavägen 23A, 171 65 Solna, Sweden.

**Respondent**

Tea Dodig-Crnković, M.Sc. in Biotechnology

Division of Affinity Proteomics, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH Royal Institute of Technology, Sweden

**Faculty opponent**

Dr. Norman Leigh Anderson

Co-Founder, Chairman and CEO of SISCAPA Assay Technologies, Inc (www.SISCAPA.com), United States

**Evaluation committee**

Assoc. Prof. Stefan Enroth

Faculty of Medicine, Department of Immunology, Genetics and Pathology, Uppsala University, Sweden

Prof. Tove Fall

Department of Medical Sciences, Molecular Epidemiology, Uppsala University, Sweden

Assoc. Prof. Paivi Östling

Science for Life Laboratory, Department of Oncology-Pathology, Karolinska Institutet, Sweden
Chair of the thesis defense
Prof. Aman Russom
Division of Nanobiotechnology, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH Royal Institute of Technology, Sweden

Main supervisor
Prof. Jochen M Schwenk
Division of Affinity Proteomics, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH Royal Institute of Technology, Sweden

Co-supervisors
Prof. Peter Nilsson
Division of Affinity Proteomics, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH Royal Institute of Technology, Sweden

Dr. Mun-Gwan Hong
Division of Affinity Proteomics, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH Royal Institute of Technology, Sweden
List of publications and manuscripts

The presented thesis is based on the following four articles. Full versions of the papers are in the Appendix.


Facets of individual-specific health signatures determined from longitudinal plasma proteome profiling.


Profiles of histidine-rich glycoprotein associate with age and risk of all-cause mortality.
*Life Sci Alliance.* 2020 Jul 31;3(10). DOI: 10.26508/lsa.202000817


Detection of autoantibodies against cancer-testis antigens in non-small cell lung cancer.

**Paper IV** Lorenzen E, **Dodig-Crnković T**, Kotilar IB, Pin E, Ceraudo E, Vaughan RD, Uhlén M, Huber T, Schwenk JM, and Sakmar TP.

Multiplexed analysis of the secretin-like GPCR-RAMP interactome.
*Sci Adv.* 2019 Sep 18;5(9). DOI: 10.1126/sciadv.aaw2778

§The authors contributed equally to the work

#Shared senior authorship
Respondent’s contribution to the appended papers

**Paper I**
Main responsible for planning and performing experimental work, co-responsible for data analysis and data visualization, main responsible during manuscript writing.

**Paper II**
Co-responsible for planning and performing single-binder and sandwich immunoassay experiments, manuscript writing as co-responsible author.

**Paper III**
Co-responsible for planning experimental work, assisted with experiments, data analysis, and manuscript writing.

**Paper IV**
Co-responsible for data analysis and data visualization, assisted with manuscript writing.
Related work not included in the thesis

Related work in chronological order.

Hellström C, Dodig-Crnković T, Hong MG, Schwenk JM, Nilsson P, and Sjöberg R.

**High-density serum/plasma reverse phase protein arrays.**

*Methods Mol Biol.* 2017;**1619**:229-238. DOI: 10.1007/978-1-4939-7057-5_18

Chen Z, Dodig-Crnković T, Schwenk JM, and Tao SC.

**Current applications of antibody microarrays.**


**Systematic development of sandwich immunoassays for the plasma secretome.**


**The human secretome.**


**Integration of molecular profiles in a longitudinal wellness profiling cohort.**

Accepted 2020-08-03 in *Nature Communications.*

A translational multiplex serology approach to profile the prevalence of anti-SARS-CoV-2 antibodies in home-sampled blood.

medRxiv. 2020 Jul 02. DOI: https://doi.org/10.1101/2020.07.01.20143966

Manuscript to be resubmitted to Science Translational Medicine.

§The authors contributed equally to the work.
<table>
<thead>
<tr>
<th>Table of Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT .........................................................</td>
</tr>
<tr>
<td>SAMMANFATTNING ............................................</td>
</tr>
<tr>
<td>POPULAR SCIENTIFIC SUMMARY ........</td>
</tr>
<tr>
<td>THESIS DEFENSE ...............................</td>
</tr>
<tr>
<td>LIST OF PUBLICATIONS AND MANUSCRIPTS ....</td>
</tr>
<tr>
<td>RESPONDENT'S CONTRIBUTION TO THE APPENDED PAPERS</td>
</tr>
<tr>
<td>I. AN INTRODUCTION TO PROTEOMICS ................................</td>
</tr>
<tr>
<td>ABOUT PROTEINS ...................................</td>
</tr>
<tr>
<td>THE HUMAN PROTEOME AND PROTEOFORMS ...........</td>
</tr>
<tr>
<td>THE PLASMA PROTEOME ....................................</td>
</tr>
<tr>
<td>THE HUMAN PROTEIN ATLAS ....................................</td>
</tr>
<tr>
<td>II. AFFINITY-BASED METHODS IN PROTEOMICS ................................</td>
</tr>
<tr>
<td>BACKGROUND ..................................................</td>
</tr>
<tr>
<td>MULTIPLEXED IMMUNOASSAYS .......................</td>
</tr>
<tr>
<td>PROTEOMICS METHODS ........................................</td>
</tr>
<tr>
<td>III. VALIDATION ..................................................</td>
</tr>
<tr>
<td>STUDY DESIGN ..............................................</td>
</tr>
<tr>
<td>EXPERIMENTAL PLANNING .........................</td>
</tr>
<tr>
<td>DATA PROCESSING AND DATA ANALYSIS ...............</td>
</tr>
<tr>
<td>VALIDATION STRATEGIES ..........................</td>
</tr>
<tr>
<td>IV. APPLICATIONS ..................................................</td>
</tr>
<tr>
<td>PERSONAL PROTEOMICS PROFILES ................</td>
</tr>
<tr>
<td>OMICS DATA STUDIES ............................</td>
</tr>
<tr>
<td>V. PRESENT INVESTIGATION .............................</td>
</tr>
<tr>
<td>PAPER I: FACETS OF INDIVIDUAL-SPECIFIC HEALTH SIGNATURES DETERMINED FROM LONGITUDINAL PLASMA PROTEOME PROFILING</td>
</tr>
<tr>
<td>PAPER II: PROFILES OF HISTIDINE-RICH GLYCOPROTEIN ASSOCIATE WITH AGE AND RISK OF ALL-CAUSE MORTALITY ........................................</td>
</tr>
<tr>
<td>PAPER III: DETECTION OF AUTOANTIBODIES AGAINST CANCER-TESTIS ANTIGENS IN NON-SMALL CELL LUNG CANCER ...............</td>
</tr>
<tr>
<td>PAPER IV: MULTIPLEXED ANALYSIS OF THE SECRETIN-LIKE GPCR-RAMP INTERACTOME .....................</td>
</tr>
<tr>
<td>CONCLUDING REMARKS AND FUTURE DIRECTIONS ..........</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS ...............................................</td>
</tr>
<tr>
<td>REFERENCES ..........................................................</td>
</tr>
</tbody>
</table>

XIII
I. An introduction to proteomics

The concept of proteomics involves the large-scale study of protein expression in biological processes, disease states, or defined conditions [1]. Proteins are molecules with diverse structures and biological functions, and they are vital to all cells of the human body. The following chapter aims to provide an introduction to the world of proteins, the molecules that have been studied in this thesis. The chapter will also cover how protein variants arise, the interest of studying these essential molecules in our body, and their important role in diagnosing diseases. An overview of how proteins can be studied systematically in cells and tissues concludes the chapter.

About proteins

The modern study of human proteins has been greatly enabled by the completion of the first human genome draft in 2001 [2, 3]. Since then, the field of proteomics has continued to advance and entails the comprehensive and large-scale analysis of proteins in cells, tissues, and other biological systems [4]. One of the main interests that drives protein science is that proteins can provide knowledge about the dynamic molecular interactions that constitute health and disease. Furthermore, proteins are still the most common drug target and the top-selling therapeutic drugs are protein-derivatives [5].

Protein function

Proteins are biomolecules that are assembled inside of cells and are integral for the biological machinery that makes up living organisms. They are often referred to as one of the building blocks of life, as they carry out vital functions in the human body. Proteins such as enzymes, hormones, antibodies, receptors, and transcription factors interact in stable or transient networks to maintain biological systems. They can catalyse chemical reactions, transport molecules, relay information between neighbouring and distal sites, regulate the immune response, as well as repair and maintain the integrity of genetic information.

A protein can be multifunctional, meaning that it may have more than one distinct function [6]. In general, the functional activity is governed by the combination of different factors such as the protein sequence, folding, location, abundance, or the controlled addition of functional groups known as post-translational modifications (PTMs) [4, 7]. Dysregulation to any of these factors can cause disease. For instance, alterations in the protein structure may lead to misfolding, mutations, or truncation. Cancer is one example where proteins often change activity as a result of
mutations [8] and proteins that misfold into aggregates have been linked to neurological diseases, such as Alzheimer’s disease [9]. Protein malfunction can thus lead to the manifestation of clinical symptoms or other observable characteristic traits, known as disease phenotypes. In biology and life science, there is a large interest of finding the link between phenotypes and the causative molecules in order gather knowledge about diverse molecular processes that may help in disease detection, monitoring, treatment, and prevention.

**Protein biosynthesis**

Information about proteins is stored in the deoxyribonucleic acid (DNA) in the form of nucleotide sequences. The DNA molecule contains genes that code for all proteins that are produced in the human body. For practical reasons, a gene-centric view is sometimes adopted, simplified to: one gene encodes one protein [10]. The production of a protein starts by DNA transcription into messenger RNA (mRNA), which in turn is translated into a protein. Different levels of conformation complexity exist, ranging from simple peptides to multimeric proteins consisting of several subunits. During translation, a linear sequence of amino acids (the primary structure) is assembled. The sequence can fold into two- and three-dimensional structures, for example forming an α helix or β sheet (the secondary structure). A protein can further be organized into a more complex arrangement of domains (the tertiary structure). Larger functional proteins, such as haemoglobin that transports oxygen in the human body, are assembled by joining protein subunits (the quaternary structure) [7].

**The human proteome and proteoforms**

The human proteome can be viewed as the collection of proteins expressed by different cells, tissues, and organs at a given time point. It is dynamic in its protein expression, influenced by both genetic and environmental factors. For example, the collection of proteins in one individual at a given time point can be determined by interplaying factors such as age, sex, and medication. Adding yet another layer of complexity to the human proteome is the protein distribution that varies between different cells and tissues within the same individual. Diversity in protein expression is even found at the single-cell level within populations of discrete cell types [11].

**Sources of protein variants**

In 2010, The Human Proteome Organization (HUPO) initiated the Human Proteome Project (HPP), with the intent to systematically characterize the complex human proteome [12]. There are 20,438 protein-coding genes in the human genome (Ensembl v 100.38), whereas the number
of unique protein variants is by far larger. Why does the number of proteins exceed the number of genes? The following section aims to describe how one single gene can be mapped to multiple proteins, expanding the proteome complexity.

Through events that occur pre- or post-translationally, the catalogue of proteins can be diversified into >70,000 structurally unique molecules, increasing the total number of protein variants that make up the human proteome [13]. Molecular changes at the DNA, RNA and protein level can create variants of related proteins, known as proteoforms [14].

**DNA level**

At the genomic level, alterations in the genetic code occur through mutations, polymorphism and recombination events. Changes to the DNA that take place in a germ cell can be passed to progeny, as opposed to somatic mutations that occur in non-germline cells [15]. Permanent mutations can appear as a result of DNA transcription errors that escape the proofreading system, or due to environmental factors such as exposure to UV radiation. Single-nucleotide polymorphisms (SNPs) involve the change of one base in the DNA sequence. Unlike stochastic mutations, SNPs are present in at least 1% of a population and are thus considered to be part of the normal genetic variation [16]. SNPs and point mutations that occur inside of a coding gene can affect the translated RNA sequence, which in turn can result in a new proteoform (Figure 1A).

Two examples of a nucleic acid change with impact on the protein sequence are missense and nonsense mutations. In the former, the DNA alteration results in the change from one single amino acid to another amino acid. Such an alteration to the protein composition can affect the protein’s conformation, activity, stability, function, and binding [17]. In cancer, mutation to proteins that regulate proliferation, vascularization, and other essential processes for cell life are frequently mutated, giving the cancer cells survival advantages [8]. In the case of a nonsense mutation, the change of one amino acid causes truncation of a protein, often rendering it non-functional. The consequence of structural variants in proteomics and their analysis is further discussed in Chapter IV and Paper I-II.

Interestingly, the majority of identified disease-associated SNPs occur outside of protein-coding genes [18]. For example, a SNP in a non-coding region that occurs at a regulatory site of a gene could influence translation efficiency, thus affecting protein abundance. Furthermore, protein expression can be regulated through the activity of the DNA sequence [19]. By mechanisms such as tightly packing of DNA around protein histones or by the addition of methyl groups, the
accessibility to the genetic code can be opened or closed [20]. The downstream effect on the protein product can be partially increased, decreased or completely silenced (on/off – effect), depending on how tightly the genetic sequence is regulated.

![DNA RNA Protein](image)

**Figure 1**: Protein variants (proteoforms) arise as a result of alterations that occur at the DNA, RNA, and protein level. (A) DNA contains genes that are transcribed into RNA. Variations in the DNA sequence, here exemplified by a SNP, can give rise to different transcripts. (B) Introns are removed during the pre-processing (splicing) of a mRNA transcript. Alternative splicing can result in unique mRNA sequence variants that are then translated into different proteoforms. (C) Proteins are further diversified into distinct proteoforms by PTMs, such as the addition of phosphates (left proteoform) or glycans (right proteoform). SNP, single nucleotide polymorphism; DNA, deoxyribonucleic acid; RNA, ribonucleic acid; mRNA, messenger RNA.

**RNA level**

At the transcriptomic level, the use of different promoter sequences, translation start sites, or alternative splicing of exon sequences can result in multiple mRNAs that originally stem from one gene (Figure 1B). Another large source of protein variants, particularly in ageing or stressed cells, is due to errors that occur during translation [13].

**PTM level**

At the protein level, the addition of PTMs can alter protein structure and activity, with the potential of directing downstream protein interaction partners, localization, conformation and protein recycling (Figure 1C). Although modification events are not observed for every protein, a fraction of the proteins can carry PTMs at one or multiple positions. Some amino acid modifications are reversible while others are permanent. Structurally, some PTMs are simple (phosphorylation, acetylation, methylation) while others are highly complex (glycosylation,
polyubiquitination). The Universal Protein Resource (UniProt) Knowledgebase provides a comprehensive database on proteins including their sequence, splice variants, and annotation of PTMs [21].

Considering the different types of PTMs that have been identified thus far, there are hundreds of thousands of additional molecules that add to the number of theoretically possible proteoforms [13, 22]. However, there is a large discrepancy between the calculated number of potential PTM combinations and the actual observed number of protein variants. The limits of proteoform diversity lie both in technological restrictions of sensitivity for protein detection and in the natural capacity of the human cell, such as how many protein copy numbers can be retained in a single cell at a given time. Mapping of PTMs continues to engage researchers although PTM detection is challenging. Methods for protein detection and measurement are further described in Chapter II.

The plasma proteome

Blood is a systemic fluid that acts as a highway in the body, transporting oxygen, nutrients, peptide hormones, immunoglobulins, waste products, and other molecules. Small amounts of protein can be released or leak from cells and tissues into the circulation, reflecting what is happening in different parts of the body. In the clinic, blood is routinely collected and tested as it can provide clues about a person’s health or disease state. Given that blood is in contact with all organs of the body, the blood proteome can carry subsets of other tissue proteomes. This is one factor that makes the blood proteome a highly complex yet attractive sample for research and diagnosis [23-25]. The following section aims to introduce the blood proteome and describe the opportunities and challenges it presents.

Blood, plasma, and serum

Blood is carried through the body by networks of arteries, veins, and capillaries. The liquid portion of blood (55%) known as plasma consists of water and dissolved molecules, there among proteins. The remaining portion (45%) is made up by blood cells – erythrocytes transport oxygen to cells in the body, leukocytes are part of the immune system responsible for protecting against diseases, and thrombocytes seal wounds by clotting [26]. For medical decision-making in the clinic and proteomics analysis, whole blood, plasma, and serum are commonly used sample preparations. Whole blood contains blood liquid and cellular components, while plasma and serum are free from blood cells. Plasma is obtained by adding anticoagulants such as EDTA, heparin or sodium citrate, followed by centrifugation to create a sample free from cells. A serum sample
is prepared by first allowing clot formation at room temperature, and then by centrifuging the sample, the serum is separated from the clot and blood cells. Considering that the serum matrix is slightly less complex in protein content while plasma contains an additional chemical, the choice of sample type depends on application and assay [25, 27, 28]. In this thesis, the liquid portion of the blood proteome has been studied, here collectively referred to as plasma proteomics or plasma profiling.

Blood-derived plasma and sera are routinely collected using standardized protocols [28]. Compared to tissue biopsies or lumbar puncture, blood sampling is minimally invasive and often requires little effort from the blood donor. Due to its accessibility, blood can be collected from the same individual multiple times for health monitoring and check-ups, for instance when taken from a patient before and after drug treatment. Large blood-based population research is feasible by biobanking blood collected across studies and donors [29]. As described above, blood can be prepared in different formats like serum or plasma, which inevitably impacts what is detectable in the sample matrix [30]. Studies that are performed during a long period of time may need to place blood samples in long-term storage, which is possible by freezing the samples or collecting dried blood spots [31].

Technical artefacts that arise from sample handling continues to be an important factor in plasma proteomics research [25]. Sample processing, age, collection protocol, and thawing and freezing cycles are some of the many pre-analytical factors that are known to impact sample quality, and therefore needs to be carefully considered when handling blood-derived samples [32, 33]. Small differences in sample collection and processing can render results that are not possible to reproduce. More on the importance of reproducibility and validation is reviewed in Chapter III.

**Proteins circulating in the blood**

The plasma proteome has an extremely wide and dynamic range of protein concentrations, making it a challenging fluid to study with the current proteomics technologies [34]. It is believed that the concentration range of proteins in plasma spans over at least 12 orders of magnitude [25]. Among the reported 5,000 detectable proteins in plasma [34], the most abundant plasma proteins such as the transporter molecule albumin or the immune-related immunoglobulins (antibodies) are measured in mg/ml in blood, while interleukins, cytokines or tissue leakage proteins are often detected at the pg/ml range. One predicament caused by the highly abundant plasma proteins is that they can mask low abundant molecules in a sample. Different strategies are employed to enable measurement of low abundant proteins, which can tell more about ongoing cellular processes in the body. Protein depletion, enrichment, and plasma fractionation are some of the methods that can increase the sensitivity of an assay in order to detect proteins of low concentration [24].
Although plasma is such a commonly used sample material in life science and the clinic, researchers still seek to collect a deep and complete database characterizing the consensus plasma proteome. In 2002, a comprehensive compendium on 289 plasma proteins was published [23], and building on that effort, HUPO initiated the Human Plasma Proteome Project (HPPP) to expanded on the list of proteins that are detectable in plasma [34]. A recent study by The Human Protein Atlas has taken on the task of cataloging the human secretome by annotating over 2,600 proteins that are secreted by cells and classifying them according to their intended location [35]. By reviewing published literature, it was established that 730 proteins (<4% of the human protein-coding genes) are actively directed to the bloodstream, where these proteins carry out their main activities.

In general, the plasma proteins are mainly produced by the liver, and secondary sources include intestines, blood cells, and other tissues [23, 35]. Proteins that are not part of the circulating plasma proteins can still enter or leak into the blood as temporary passengers or as a result of the natural process of cell damage and death. Disease-related proteins that are for instance secreted by tumours or originate from pathogens can also reside in plasma [36]. In pregnant women, proteins can even pass through the placenta into the blood, in this way exchanging proteins between mother and baby [37]. Given that the protein content in the blood varies over time and between individuals, the heterogeneity of proteins found in plasma further adds a wealth of information to the plasma proteome [38].

**Clinical significance of proteins**

One major application of blood analysis is the usage of serological tests for protein, antigen and autoantibody identification. In clinical laboratories, enzyme-linked immunosorbent assays (ELISAs) are still considered to be the gold standard for measuring specific protein biomarkers in plasma [39]. According to the Food and Drug Administration (FDA) and National Institutes of Health (NIH) joined council, a biomarker is “a defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or biological responses to an exposure or intervention, including therapeutic interventions” [40]. Biomarkers can be further classified by their clinical application, and the three major groups are diagnostic, prognostic, and predictive biomarkers. A **diagnostic** biomarker allows the detection of a disease or condition. A **prognostic** biomarker assesses the natural course of a condition, such as the likelihood of relapse, disease progression, or other clinical events in patients. A **predictive** marker allows identification of individuals that are more likely to respond to treatment or exposure, in comparison to similar individuals. On average, not more than 1.5 novel protein biomarkers are approved by the FDA per year [36].
In order to be implemented in a clinical test, a biomarker must demonstrate clinical validity entailing high sensitivity and specificity. Sensitivity refers to the ability to identify all positive cases (true positives) where a disease is present, while specificity is the ability to detect negative cases (true negatives) where the disease is absent. A test that detects many false positives or false negatives thus has low sensitivity and low specificity. In addition to these two factors, other variables, such as the biological diversity between individuals, can be challenging to overcome during a biomarker’s journey from discovery to validation.

One biomarker that was early translated into a clinical assay is cardiac troponin, a protein that is specifically expressed in heart muscle tissue and is detectable in blood in the case of myocardial infarction. The clinical test for troponin detection has been well established and is considered to have a reliable diagnostic performance [41]. A handful of biomarkers for cancer diagnosis are also FDA approved for clinical use [42]. However, these are not relied upon as stand-alone tests but rather applied in combination with traditional diagnostic methods, such as tissue biopsies [43]. Not all biomarkers that have been introduced in the clinic have proved to be successful in terms of accurate diagnosis. One biomarker that has received controversial reputation is the prostate-specific antigen (PSA) for the detection of prostate cancer. It has been reported to be an insufficiently precise biomarker as PSA can also be elevated in individuals free from cancer, causing overdiagnosis [44].

Nowadays, it is in general recognized that one protein alone may not be specific enough to capture the complex biological processes of one disease [45]. Therefore, the future of biomarker discovery may not necessarily rely on one indicator, but rather utilize panels containing several biomarkers to increase sensitivity, accuracy, and minimize false-positives and false-negatives [46]. Another approach is to go beyond proteomics and incorporate several layers of biological information, such as combining protein biomarkers with genetic information, transcriptomics, metabolomics, and other omics data [47]. Personalized health monitoring incorporates yet another strategy, where markers of disease are evaluated on the level of the individual, rather than solely on population-based cut-offs. Here, critical biological parameters (such as proteoforms) are continuously followed and a digression from a person’s reference level could indicate transition into a disease [48, 49]. More about the trend towards personalized health assessment is covered in Chapter IV.

The Human Protein Atlas

The Human Protein Atlas (HPA) is a Swedish project initiated in 2003, intending to systematically map all proteins that are expressed in the human body [50]. To achieve this, mRNA and protein distributions are measured in tissues [51], and cells and their subcellular compartments [11]. Alongside these, mRNA and protein expression in major cancer tissues are described in a
pathology atlas [52]. Recently, three additional sub-atlases have been published that measure proteins in blood [35, 53], brain [54], and metabolic pathways [55]. The collected results from the HPA project are continuously made available through the online and open-source portal proteinatlas.org. Today (version 19.3, updated 2020-03-06), six atlases make up the HPA portal. In total, 17,058 unique proteins have been catalogued using 26,371 antibodies – a class of proteins that are components of the immune system and can also be used as a tool for protein capture and detection.

In order to build this comprehensive gene-centric database on proteins, the HPA project started by creating a pipeline for producing and validating affinity reagents - protein fragments and antibodies - for protein measurement. The resource of reagents provided by the HPA has been essential for enabling the research projects described in this thesis.

The pipeline for each protein starts with the in silico design of a recombinant protein fragment, called protein epitope signature tag (PrEST) [56, 57]. Each fragment consists of approximately 25-150 amino acids, which allows efficient cloning and protein expression in Escherichia coli. The length of the PrEST sequence is long enough for conformational epitopes to form, ideally resembling the native protein structure. The selected sequence itself is designed to have a low (<60%) sequence homology to other human proteins. Further, the sequence never corresponds to a transmembrane region of a protein or a signal peptide. All PrESTs carry an N-terminal dual tag consisting of six histidines in tandem (His$_6$) and an albumin-binding domain (ABP). His$_6$ is used for purifying the fragments on nickel columns, in order to remove any bacterial contaminants. Next, the PrESTs are introduced into rabbits and the Streptococcal protein G-derived ABP acts as an immunostimulant. The immunized rabbits produce mono-specific polyclonal antibodies, which are affinity-purified from the rabbit sera using the dual tags and the source PrEST [58]. Antibody twins are generated by introducing identical PrEST constructs into multiple rabbits. Antibody siblings are produced by immunizing multiple rabbits with PrEST sequences that stem from the same target protein but cover different epitopes. Finally, the produced antibodies are validated by antigen assays [59]. Only antibodies with high specificity towards its target antigen and low cross-reactivity with other PrESTs are approved for further characterization using immunohistochemistry, immunofluorescence and Western blots. HPA antibody twins, siblings and binders from commercial sources are compared for further validation of the generated data.

The affinity reagents generated within the HPA project have versatile applications and until today they have been used in >500 different research projects [60]. In this thesis, antibodies and PrESTs have been applied for protein- and autoantibody-profiling in blood serum and plasma (Papers I-III), and for the study of protein-protein interactions in cell lysates (Paper IV). Alongside microarrays, HPA reagents have also been used for mass spectrometry (MS); PrESTs
have been adapted to isotope-labelled protein standards for absolute quantification (QPrESTs) [61] and HPA antibodies have been used in immunocapture MS [62].

Alongside HPA, there are other large-scale initiatives of mapping the human proteome. The publicly available databases ProteomicsDB and Human Proteome Map both provide mass-spectrometry based drafts of the human proteome [63, 64]. The Human Cell Atlas is an ongoing international effort to systematically describe human cell types and their proteins and gene expression, particularly at the single-cell level [65].

Identification of proteins with the purpose of mapping the proteome can further our understanding of proteins and their mode of action, as well as how they contribute to complex biological pathways. In order to gain knowledge about disease mechanisms and establish accurate diagnostic tools, the field of proteomics has developed several methods for protein analysis, some of which are described in Chapter II.
II. Affinity-based methods in proteomics

The following chapter introduces technologies for measuring proteins, predominantly in blood, with a focus on affinity-based methods. While the choice of proteomics technology often depends on the research question at hand, additional aspects to consider include the availability of reagents, personnel expertise, instrumentation, and costs. Furthermore, sample type, throughput, multiplexing capacity, and required sensitivity are other crucial points when selecting proteomics technology. Here, throughput refers to the number of samples per time unit that can be analysed by a method, while multiplexing denotes the total number of analytes that can be measured in parallel.

Background

The three most common methods in proteomics are chromatography, MS, and affinity proteomics, or a combination thereof [66]. Chromatography and gel-based protein separation were developed as the first technique for protein detection and size estimation. Chromatography is used for separating or purifying proteins from mixtures and is a powerful tool when coupled to sample preparation in MS. Despite limitations concerning sample reproducibility, throughput and sensitivity, a gel-based method is still a valuable tool for small-scale target analysis or when used as a protein enrichment step combined with MS or antibody microarrays [67, 68]. The most widely used proteomics method today is MS, where proteins are ionized and passed through a mass analyser. There, peptide sequences are deconvoluted by matching the resulting mass-over-charge ratio spectra to theoretical spectra derived from sequence databases. However, restrictions in MS when it comes to sample throughput, computational time, and expensive instrumentation has led to an increased interest in protein analysis by affinity-based methods [69]. Compared to MS, affinity-based technologies rely on reagents, often antibodies, that are designed to bind proteins with high affinity. This concept is particularly advantageous when searching for low abundant proteins that are otherwise difficult to detect. Below follows a more detail overview of affinity proteomics using multiplexed immunoassays.

Multiplexed immunoassays

The following section will cover the concept of immunoassays and describe the suspension bead array (SBA) that is applied in Paper I-VI. For the analysis of soluble proteins, affinity-proteomics assays utilize capture reagents (binders or analytes), that are designed to bind to proteins of interest with high specificity and sensitivity [70]. As described already in 1989, an ambient analyte immunoassay can be constructed as a miniaturized assay using extremely small amounts of
antibodies added to a solid support. Under these conditions, protein detection is possible independent of antibody concentration or sample volume [71]. With the advances in automation of liquid handling robots and analysis instruments, modern immunoassays can process hundreds of samples and proteins in parallel.

**Forward-phase and reverse-phase arrays**

In a forward-phase array (Figure 2), binders are immobilized in arrays onto a surface, such as a glass or plastic slide, or a microsphere. In the next step, a sample containing proteins is added to the array so that the affinity binders can capture their target proteins from the solution. Unbound proteins are removed by washing of the arrays. Following the introduction of reporter molecules, the interaction between a bound protein and affinity reagent can be detected optically. This is achieved if the bound proteins are fluorescently labelled, or through the addition of a secondary antibody that carries a fluorophore or enzyme that generates a signal molecule. The described approach has proven to be an attractive method when the aim is to measure a large number of analytes in a complex biological sample matrix [72-74]. Recently, an array comprising of >60,000 monoclonal antibodies was designed for the purpose of proteome screenings [75].

![Diagram](image)

**Figure 2:** Immunoassays on suspension bead arrays. (A)-(C) Forward-phase arrays. (A) In a single-binder assay, antibodies are coupled to microspheres and incubated with biotinylated proteins. Detection is enabled by the addition of streptavidin conjugated to a fluorophore. (B) In a sandwich assay, two independent antibodies that share the same target protein are incubated with a sample. Capture antibodies are conjugated to the beads and incubated with a sample, and labelled detection antibodies are added for readout. (C) When applying a protein-protein-interaction (PPI) assay, two antibodies are applied, each targeting different proteins that appear in complexes. (D) In a reverse-phase array, antigens are applied to the beads and incubated with a biological sample. In the presence of a binding antibody, the immunoglobulin can be detected by the addition of a secondary labelled antibody.
Besides the mentioned format, a *reverse-phase array* can be a suitable approach when investigating a large number of samples while focusing on a few specific analytes. In a reverse-phase protein array (RPPA), biological samples are printed onto a solid surface while the capture reagent is added in solution to the array (Figure 2). Recent efforts have enabled the design of planar RPPAs that contain >12,000 serum samples (a whole biobank) on a single slide [76].

**The suspension bead array**

In the work presented in this thesis, multiplexed and high-throughput assays for protein detection in human blood have been constructed. This method was developed by the Luminex corporation [77] and relies on the conjugation of binders (here antibodies or protein fragments) to carboxylated magnetic and colour-coded microspheres (beads) [78]. Each bead contains a unique ID composed of an internal mix of fluorescent dyes, enabling the identification of each microsphere within a pool of IDs in suspension. Thus, the suspension bead array (SBA) allows multiplexing, up to 500 different binders, and can be adapted to multiple assay formats. In addition to blood-derived sera and plasma, the SBA assays and similar commercially available bead-based kits have been adapted to enable the analysis of other human sample types, such as cerebrospinal fluid [79], bronchoalveolar lavage [80] and saliva [81].

In the multiplexed single-binder assay (Figure 2A, Figure 3), plasma sample are diluted, labelled with biotin, and heat-treated at 56°C in 96-well microtiter plates (Figure 3A). In parallel, the SBA is prepared by immobilising one antibody type per bead ID and then pooling the beads together, creating the antibody array (Figure 3B). Beads are incubated with the biological samples in either 96-well or 384-well plates. The relative antigen binding can then be detected by the addition of streptavidin coupled R-phycoerythrin (Figure 3C). Data acquisition is enabled using a Luminex instrument with two lasers: one laser detects and classifies each bead ID based on a bead’s internal fluorescent dye, while the other laser measures the relative amount of bound protein via the fluorescence emitted by the reporter molecule. The median fluorescence intensity (MFI) is reported per bead ID.

Due to the high sample throughput compared to traditional proteomic approaches and the flexibility of adding any antibody into the bead pool, the single-binder SBA technology is an attractive platform for larger exploratory studies for relative protein quantification, as further described in Paper I-II. However, using the single-binder assay entails the risk of off-target binding or measuring of protein complexes rather than quantifying single proteins.
Figure 3: Antibody SBA workflow. (A) Plasma samples are diluted, biotinylated and heat treated. (B) Capture antibodies are conjugated to beads with unique IDs, usually ranging from ID 1-384. Beads are pooled together and dispensed into a microtiter plate. (C) Plasma samples are incubated with the SBA in microtiter plates, and a streptavidin coupled fluorophore is added for protein detection. Antigens that are captured by the antibodies are detected using a flow cytometry-like system (Luminex FlexMap instrument) for the detection of bead ID and relative fluorescence intensity.

In order to increase the stringency and detect one protein at a time, a sandwich immunoassay (SIA) can be utilized (Figure 2B). A SIA relies on two binding events: a capture antibody binds a target protein in solution, and after washing, a fluorescently labelled detection antibody that targets the same protein of interest is added [82]. In order to register a signal, it requires that two (preferably independent) antibodies bind the same target protein, which decreases the likelihood of cross-reactivity and off-target binding [83]. SIAs and similar dual-capture approaches (such as ELISAs) are attractive for small-scale antibody validation and absolute protein quantification. However, the challenge with a SIA is that it does require the availability of independent binders and it is difficult to multiplex beyond a few analytes. Further, the performance of the SIA entails that the assay protocol needs to be tailored per antibody pair and to the target protein. However, a recent effort where 2,170 antibody pairs were analysed in dilution series showed that through careful study design and assay optimization, it is possible to develop a pipeline for parallel screening of SIA pairs [84]. Another format of the dual-binder assay is the study of protein-protein interactions (PPI). Here, each antibody within the antibody pair targets a different protein, in order to measure if the two proteins appear in complexes with each other (Figure 2C). An in-house developed PPI assay for the detection of GPCR-RAMP complexes is further described in Paper IV.
Assays for profiling antigens can also be built on the SBA format by coating beads with antigens, represented as peptides, PrESTs, or full-length proteins. These SBAs are used for identifying the presence of antibodies via the detection of immunoglobulins (Figure 2D). The method has been successfully adapted for detecting anti-human IgG in the blood of patients with atopic dermatitis [85], multiple sclerosis [86], systemic lupus erythematosus [87], and lung cancer as described in Paper III.

Commercial immunoassays

Due to the wide interest in antibody-based assays and their applicability in protein research, pre-made microarrays have become commercially available [70]. Commercial multiplexed planar and bead-based assays often cover protein targets with a certain area of interest, such as cytokines [88], cell signalling [89], autoantigens [90], or PTMs such as phospho-arrays [91]. The company Olink is among the recently popular multiplexed immunoassay provider that offers panels for the detection of 92 proteins in 96 samples. Using the proximity extension assay (PEA) technology, proteins are detected through a dual-capture system where antibody pairs carry complementary oligonucleotide labels. When two antibodies bind in close proximity onto a common protein, the oligonucleotides anneal and the hybridization event can be amplified by PCR [92]. Similarly to PEA, in situ proximity ligation assays (PLA) can be used as an immunohistochemical tool for detecting protein interactions in tissues and cells [93]. For validation purposes, PEA and PLA were performed in Paper I and Paper IV, respectively.

Proteomics methods

This section aims to briefly describe different types of reagents, besides antibodies, that are used on microarrays, as well as the well-established MS concept in combination with affinity reagents.

Affinity reagents

Microarrays can be assembled using small antibody constructs such as single-chain variable fragments (scFv) and nanobodies. Due to the small size of the fragments, screening and generation of binders can be performed in large quantities by phage display. Through an iterative selection of high-affinity binders, the technology enables testing for protein-antigen, protein-protein and protein-DNA interactions in a high-throughput manner [94], as well as screening for therapeautic fragments [95]. Binders used in microarrays can also consist of other antibody mimetics or synthetically engineered capture reagents, including affibodies based on Staphylococcus aureus Protein A [96], designed ankyrin repeat proteins (DARPins) [97], and aptamers constructed of nucleic acids [98]. For the latter, the company SomaLogic has developed a multiplex aptamer-based
microarray that can measure >5,000 human proteins in human blood serum and plasma [99, 100].

**Mass spectrometry and affinity-proteomics**

The most widely used technology in the proteomics field is still MS. Unlike microarrays that are based on a set of pre-selected binders, MS can use both a targeted selection or a hypothesis-free approach. With a bottom-up MS approach, proteins are digested using sequence-specific enzymes (e.g. trypsin) and the resulting peptides are separated by liquid chromatography, ionized, and identified by matching the observed fragment spectra with spectra annotated in databases [101]. Advances to speed up the plasma proteomic pipeline now allows more accurate and rapid proteome profiling [32]. With a top-down MS approach, protein quantification and PTMs on intact proteins can also be studied. However, the method is still computationally and experimentally challenging. Proteins are interacting molecules, and the mapping of proteoform networks is essential for understanding the underlying mechanisms of health and disease. Therefore, MS methods for measuring protein interaction partners, stoichiometry and abundance have been developed [22].

MS methods for plasma proteomics have undergone a revival in recent years, aiming to increase translation of biomarkers into the clinic [25]. Despite its wide usage in research and even its quantitative properties, clinical MS is still less implemented in routine hospitals compared to ELISA or other immune-based assays [102]. However, immuno-MS that combines MS and affinity proteomics can facilitate translational proteomics, as demonstrated by the assay for thyroglobulin detection in blood [103]. The sensitivity of assays has been increased by utilizing antibodies for the enrichment of particular proteins or peptides in a sample, while MS is coupled for target detection [39]. Immuno-MS approaches such as the stable isotope standards and capture by anti-peptide antibody (SISCAPA) method have demonstrated how the detection of low abundant proteins can be significantly increased by utilizing antipeptide binders with MS [104]. Further, systematic antibody validation by immunoprecipitation with MS has also been developed by joining affinity proteomics and MS-based analysis [62]. The combination of MS and affinity proteomics can thus be a powerful approach both for proteomic studies and for antibody validation.

In summary, immunoassays allow the detection and study of proteins and combining multiple proteomics approaches can be a powerful validation tool, as further discussed in Chapter III.
III. Validation

Reproducibility is a fundamental element in science as it provides research validity and credibility. Yet today the field is facing a reproducibility crisis [105, 106]. Therefore, the following chapter will highlight selected aspects that are important to consider in proteomics studies when pursuing reproducible research.

Here, the term reproducibility will be used as an umbrella term to encompass: “repeatability (same team, same experimental setup), replicability (different team, same experimental setup) and reproducibility (different team, different experimental setup)” [107]. Whereas reproducibility of an immunoassay largely relates to technical aspects, validation is another commonly used term that can entail reproducibility, as well as incorporating additional technical and biological aspects. For instance, validation can refer to assay development that deals with accuracy, precision, reproducibility, limit of detection, and similar optimization parameters [108]. In comparison, the concept of antibody validation may further involve the establishment of what antigen or epitope a specific analyte binds [109]. From a biomarker discovery point of view, validation may signify the ability to observe similar associations between protein and disease outcome when studying independent sample sets [45]. Certainly, the concept of validation encompasses many aspects that are relevant to immunoassays and protein profiling. Thus, this chapter will discuss selected validation aspects that were considered during the work presented in this thesis.

Study design

The design of a study and its experimental procedures is critical in research projects. These two aspects should be carefully determined before an investigation is undertaken, as each type of design comes with set limitations and potential sources of bias. The choice of design will limit which hypothesis can be appropriately formulated, and which statistical measurements can be applied.

Research is often conducted in order to test for hypotheses concerning the causality between specific exposures and outcomes, such as “Does smoking (exposure) cause lung cancer (outcome)?”. Clinical studies can be divided into two broad categories, depending on the chosen study design (Box 1) [110]. In a clinical trial, individuals are assigned (often through randomization) to an intervention or control group for disease prevention or treatment. As opposed to a clinical trial, observational studies do not include any type of intervention. Instead, individuals are
allocated into groups determined by factors such as exposure or disease outcome. Here, the cohort study and the case-control study are described in more detail.

**Box 1:** A study design can be divided into two categories, clinical trials and observational studies. Individuals that are enrolled in a clinical trial can be assigned or randomized into groups, where one group is subject to an intervention. Observational studies include cohort studies, case-control studies, and cross-sectional studies. Both the cohort and case-control study have a temporal component, while a cross-sectional study analyses data at one specific point in time.

<table>
<thead>
<tr>
<th>STUDY DESIGN</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clinical trial</strong></td>
</tr>
<tr>
<td>- Non-randomized</td>
</tr>
<tr>
<td>- Randomized</td>
</tr>
<tr>
<td><strong>Observational study</strong></td>
</tr>
<tr>
<td>- Cohort study</td>
</tr>
<tr>
<td>- Prospective</td>
</tr>
<tr>
<td>- Retrospective</td>
</tr>
<tr>
<td>- Case-control study</td>
</tr>
<tr>
<td>- Cross-sectional study</td>
</tr>
</tbody>
</table>

A cohort study, a type of longitudinal study, is carried out in chronological order and can be a suitable approach when the aim is to investigate the natural history of a disorder or when determining incidence rates. A study that aims to assess the risk of developing lung cancer could use a prospective design by following healthy individuals longitudinally until the onset of disease. Retrospective cohort studies are instead initiated after an outcome is established, and with the help of medical records that can aid in identifying the effect of an exposure [111]. Since cohort studies contain a temporal component, they may help to untangle cause from outcome, such as “does A trigger B, or vice versa?” [10]. This longitudinal study design also allows for investigating if a single exposure can give rise to multiple outcomes. Another advantage with a longitudinal study design is that it allows each individual to be used as its own reference, for instance by comparing an individual before and after a disease outcome. However, studies that rely on repeated sampling often suffer from patient drop-outs or loss to follow-up. The investigation of rare exposures requires a large number of subjects and can end up being costly and time-consuming. Other forms of longitudinal studies exist where data is collected from individuals that do not share a pre-defined exposure (Paper I-II) [112]. The growing interest in individual-based longitudinal sampling is further discussed in Chapter IV.

Another frequent choice of study design is the case-control study (Paper III). Here, groups are not defined by exposure but rather by the outcome. If the outcome of interest is observed in a subject, then the individual will belong to the case group, and if not observed, then the individual belongs to the control group. A case-control study can be favourable to undertake if the incidence rate is low in a population, or when investigating diseases that take a long time to develop.
A reoccurring challenge is the selection of appropriate control groups that are representative of the background of the case group [113]. Choice of control group can be vulnerable to bias, and a strategy to circumvent this issue is to match cases and controls to critical and known variables that may influence the outcome, such as age, sex, lifestyle, medication, presence of disease, and genetic heterogeneity [114].

In the work presented in this thesis, sample collection and study design were performed by collaborators. Therefore, the main focus during project planning in our laboratory has instead been on experimental design, data analysis, and antibody validation.

**Experimental planning**

As earlier discussed in Chapter I, pre-analytical variables such as sample handling, processing, storage, and choice of sample type are important variables to factor in before committing to a study design. However, assay preparation and processing are also sources of variation that can introduce technical noise and changes in measured data. If these threats to validity are not addressed, they may lead to inaccurate interpretation of the gathered data, as well as non-reproducible results. True associations can be diluted, and artefacts may influence the outcome of statistical analysis. Thus, careful consideration of the experimental design is critical in order to minimize technical variability and bias.

**Batch effects**

High-throughput immunoassays entail the study of large sets of samples. Commonly, these samples are processed in microtiter plates that can accommodate 96 or 384 samples per analysis unit. Therefore, when investigating a larger number of samples or analytes that exceed the size of one microtiter plate, the analysis is split up into smaller batches [115]. Batch-to-batch variation is a reoccurring feature seen in molecular biology data, and it arises due to the accumulation of small processing differences between batches. Sample preparation, reagent lots, instrument, and time of the experiment are a few examples of technical factors that can contribute to batch effects [116]. In the SBA workflow, samples are biotinylated, heat-treated and diluted in 96-well plates and later transferred into the 384-well format. Hence, batch effects can arise from different steps both in 96-well and 384-well batches. If experiments are further divided across multiple days or operators, that additionally needs to be factored in prior to data interpretation. Systematic noise that appears as shifts of signal intensity across or within microtiter plates is a phenomenon that can occasionally be observed in immunoassays that utilize fluorophores for an optical read-out [117]. Notably, batch effects are not unique to protein data but can occur in all molecular biology studies. With the increase of high-throughput and multiplexed assays, methods for tackling
technical variability have been developed including practical guidelines on experimental design [118] and various data normalization methods [116].

Considering the experimental design, sample randomization [118] is a valuable strategy that can be employed to mitigate severe confounding factors, such as when batch effects are correlated with an outcome variable of interest [119]. By allocating samples across assays, potential pre-analytical variables, confounding factors, and technical noise are distributed across batches, thereby minimizing unintentional bias [120, 121]. However, hidden variables such as health history may be unknowingly present in a collection of samples and may impact the results if these, by chance, remain unbalanced across batches.

**Replicates, positive and negative controls**

Samples designated as internal quality control (QC) are another essential tool for evaluating the performance and reproducibility of an assay and for identifying batch effects (Figure 4A). In the SBA workflow, replicates, positive and negative controls are included in every assay.

*Technical replicates* can be allocated within and between sample plates for estimating intra- and inter-assay variability. Replicates can be created from a single sample or a pool of samples. The choice often depends on restrictions in sample volume or the number of available positions in a plate. Optimally, a complete replicate of an assay should be performed in order to estimate a workflow’s reproducibility. When planning for larger studies that will likely include several batches, perhaps combining samples from different sources or cohorts, it is important to create an experimental design that can accommodate new study set. If possible, the same technical replicates, or bridging samples, should be included in all batches to be later utilized for measuring variance and performing data normalization [122]. *Negative controls* can consist of a buffer that does not contain the proteins of interest [121]. These can thus indicate the signal background and is sometimes included for determining the assay’s performance range. Further, biological samples that display a similar data distribution as the negative controls can be identified as technically failed and excluded from downstream analysis. Furthermore, a blank position in the plate can be utilized as a reference for plate orientation [121]. *Positive controls*, such as standards containing proteins spiked-in at known concentrations, contain the proteins of interest. These are added to an assay making it possible during QC to evaluate the performance of the experiment. Sometimes, positive controls are used for quantifying the amount of protein in the analysed samples (Figure 4B).

Importantly, experimental planning is not only required from the sample point of view but also for analytes. Internal positive and negative control bead can also be included as a first step in
analyte QC. The positive controls used in plasma profiling are often anti-albumin and anti-human IgG antibodies [72, 123], which confirm the presence of labelled proteins in samples. As these proteins are present in high abundance in plasma, the signal intensities are expected to reach saturation (Figure 4C). Hence, data from other antibodies that correlate with positive controls may also be outside of the assay’s performance range. Negative controls include bare beads that resemble beads without any antibody, while beads carrying normal rabbit IgGs provide an idea on antibodies lacking a defined specificity. Therefore, data from antibodies that correlate with the negative controls may also demonstrate non-specific binding. Similarly, assays for the detection of autoantibodies present in plasma can utilize different types of antigens conjugated to beads. There, beads carrying anti-human IgG and Epstein–Barr virus nuclear antigen 1 (EBNA1) (an antigen commonly found reactivity towards) can act as positive controls. In comparison, bare beads and beads carrying common tags that can be engineered on the antigens may provide the expected background signal [124].

Figure 4: Immunoassay quality control and assay performance before data normalization. (A) Batch effects may appear due to small differences in assay processing. Here, the Y-axis is the detected signal intensity of one protein and the X-axis corresponds to different batches. As the box plots illustrate, here a differential signal intensity appears between batch A and B. (B) A protein standard can be used for quantifying the amount of protein in an analysed sample. The assay’s performance range, such as the upper and lower limits of detection (dashed lines), can be estimated from a protein that is spiked-in at known concentrations. Biological samples and technical replicates should preferably fall within the linear range of the assay, which is between the upper and lower limit of detection. (C) Quality control can include the evaluation of positive and negative analyte controls. The positive and negative control (dashed lines) may indicate the upper and lower signal limit of an assay, respectively.
Data processing and data analysis

The introduction of technical artefacts during assays is unfortunately inescapable in the laboratory. For experiments where batches are processed equally and the assumption is that the sample and analyte distributions are balanced across batches, data normalization approaches can be applied to reduce systematic variations in the data. Ideally, appropriate technical controls should be included early in the assay preparation, so that they may report the processing steps that each sample undergoes. By identifying effects that impact the generated data, quality control approaches can be applied to correct for observed technical variations.

In the SBA data workflow, the first QC step is to assess data and exclude or flag technical outliers (samples or analytes). Here, replicates, positive and negative controls are evaluated, as well as the total number of registered beads per sample and analyte. In the next step, systematic sources of technical variation are addressed, including batch effects and drifts of signal intensity across plates.

For SBA plasma profiling work presented in this thesis, two normalization strategies have been employed, as described by earlier studies [125, 126]. These normalization approaches entail probabilistic quotient normalization (PQN) that can account for differences in sample dilution [127] and Multidimensional MA for batch correction [115]. Linear regression on normalized data can further be performed on covariates such as sequential reading order, age, and sex, in order to adjust the data for technical or biological factors.

Transformation or standardization may also be a required step in data processing prior to statistical analysis. Parametric tests, such as the paired t-test or the one way analysis of variance (ANOVA), assume a particular data distribution [128]. The most common assumption in parametric tests is that the data is normally distributed, however, biological data often entails working with skewed data sets. Log transformation is one widely applied method for addressing data with a non-normal distribution. Another strategy for data standardization is the use of Z-scores, which describe the distance of a data point from a population mean in terms of standard deviations. Thus, a Z-score that is equal to 0 indicates that the measurement is identical to the population mean. When working with bimodal distributions, continuous measurements can be converted into binary data before initializing statistical analysis [129]. Proper selection of transformation or standardization method is necessary as it may impact on the outcome of the downstream data analysis [128].

Missing data is another important aspect in data analysis which will be briefly covered here, although it is not a common feature in SBA as compared to MS data. Besides the choice of technology, missingness can appear due to other factors such as the incompleteness of reported
clinical data or due to the choice of study design. In a longitudinal study, missingness can occur if a sample donor does not provide a sample at a particular time point or drops out of the study. Hence, a strategy needs to be formulated in the data analysis pipeline on how to deal with missing values. There are three general groups of missing data: missing completely at random, missing at random, and missing not at random [130]. As their names indicate, each group has a different type of missing data, here named according to increasing order of systematic missingness. Hence, missing values can be random or non-random in nature, which needs to be taken into account if applying imputation [131]. In order to determine the robustness of the results generated from imputed data, a sensitivity analysis may be performed [132].

Before initiating data modelling or biological interpretation, it is necessary to evaluate if the data processing methods are appropriate and help to improve the data quality. Studies have shown that data imputation can lead to bias, while normalization may risk introducing artificial structures in the data, leading to incorrect conclusions [133]. Furthermore, batch-correction and other data normalization methods can only do so much as to reduce certain aspects of the observed, systematic or predictable technical variability. Bias inherited from the experimental design remains challenging to overcome, and proper experimental planning remains crucial [116].

Data visualization and choice of statistical model for hypothesis testing are instrumental for data interpretation. Especially with large datasets, it is often not feasible for a researcher to quality control one point at the time. Therefore, methods that can provide a broad overview of the acquired data are desired. In order to visualize structures within a dataset, dimensionality reduction methods can be utilized, such as principal component analysis (PCA) [134], t-distributed stochastic neighbour embedding (t-SNE) [135], and uniform manifold approximation and projection (UMAP) [136]. These algorithms can help in identifying certain patterns in the data caused by either technical or biological variation. They may for example reveal if batch effects or donor-specific factors have an impact on the data or detect unexpected sub-groupings of individuals that can be of interest to follow up.

**Validation strategies**

As highlighted in Chapter II, antibodies have a broad utility and can for example be used for protein purification, target enrichment, and the measurement of proteins in biological samples. In the clinic, antibodies support disease diagnosis via ELISAs for the detection of proteins in biofluids or via immunohistochemistry on tissue specimens. Characterization of antibodies is necessary before pursuing any biological interpretation, and the role of biological validation concludes this chapter. Here, the validation of therapeutic antibodies will not be covered. Instead, the focus is on the validation of antibodies used for research applications.
The purpose of every antibody is for it to selectively bind to preferably one target epitope, even in the presence of other similar and more abundant epitopes. However, the affinity of antibodies is context and application dependent, meaning that while they can be suitable in their intended assay and sample type, they may be inadequate in other settings [109]. An antibody that has been generated to recognize a particular protein in its native conformation might not be able to capture a denatured version of the protein [137]. Modifications of the target protein such as PTMs or labelling molecules may also interfere with or influence antibody-antigen binding [138]. Polymorphism and cleaved proteins can further disrupt the intended binding interaction [139].

Antibody selectivity is dependent on sample composition, its complexity, and the presence or absence of the target and other similar proteins [140]. If the antibody captures its target protein from a mix of proteins, the binding is said to be on-target. If the antibody instead binds to other proteins, for example with similar epitope sequence or conformation, the binding is off-target. An antibody may even bind several proteins by direct or indirect co-enrichment [62]. Notably, antibody binding can be reproducible but not due to on-target capture but caused by repeated off-target binding. Such cross-reactivity is a major concern in both single- and dual-binder immunoassays and is essential to evaluate during assay development [83]. The same antibody-antigen binding properties hold true in reverse phase arrays; the choice of antigen presentation (full-length protein, PrEST, or peptide) can impact the recognition and binding of the endogenous antibodies. In summary, antibody-antigen binding can depend on conformation, concentration, and sequence of the target protein, but also on the composition and complexity of the sample.

Additionally complicating the matter is the fact that the performance of antibodies, even those that are validated, can still vary between production lots causing concern regarding antibody reliability and reproducibility. For example, polyclonal antibodies are produced by introducing an antigen into a host animal (e.g. rabbit, chicken, or donkey), eliciting the immune response. The immunization stimulates multiple B-cell clones to generate antibodies against the foreign antigen. Thus, the antibodies obtained from the sera of different animals consist of a mixture of antibodies that bind multiple, and possibly unique, sets of epitopes of a single antigen and with different degrees of specificity and affinity [138]. In contrast, monoclonal antibodies are produced by selecting and fusing a single B-cell clone from a host animal with an immortal cancer cell line, creating hybridoma cells originating from a single clone. These cells produce collections of identical antibodies that target one single epitope of the antigen [139, 141]. Thus, polyclonal antibodies are more susceptible to product lot-to-lot variation compared to monoclonals, as variation arises when immunizing multiple animals. Still, monoclonals are vulnerable to other factors; future batches can be lost if the cell line is contaminated or dies, and variation can be
introduced due to hybridoma genetic drifts [142]. In general, polyclonals with their mixture of antibodies are more tolerant to changes of antigen presentation such as denaturation, while monoclonals are attractive for detecting small molecular changes, such as PTMs [139].

A large number of monoclonal and polyclonal antibodies for research are available through vendors selling affinity reagents online. However, many of the >4.2 million antibody products [143] that are offered have not been fully characterized, or the vendor information is incomplete. Guidelines regarding the recommended application or sample type for antibodies are also often not provided [109]. Thus, a demand for transparency and efforts concerning antibody validation have been put forth by antibody users and providers. Open-access databases, such as Antibodypedia [144], collects data on commercially available antibodies; both information regarding appropriate applications and also user experience of antibodies are published there. Another effort is the Research Resource Identification Initiative (RRID) that provides unique identifiers for life science tools, such as antibodies [145]. The aim is to improve reporting standards and transparency in the scientific community.

Recommendations on application-specific antibody validation have been published, one of them by the International Working Group for Antibody Validation (IWGAV) [109]. The proposed “five pillars of validation” framework (Box 2) entails recommended approaches for providing evidence that an antibody binds to its target protein.

**Box 2: The five pillars of antibody validation.**

<table>
<thead>
<tr>
<th>FIVE PILLARS OF ANTIBODY VALIDATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>i. Genetic strategies*</td>
</tr>
<tr>
<td>ii. Orthogonal strategies</td>
</tr>
<tr>
<td>iii. Independent antibody strategies</td>
</tr>
<tr>
<td>iv. Expression of tagged proteins*</td>
</tr>
<tr>
<td>v. Immunocapture followed by MS</td>
</tr>
</tbody>
</table>

*Not applicable in human plasma

As antibody validation can be a long and cumbersome process, IWGAV recommends that “at least one of the pillars should be used as a minimum criterion for claiming that a particular antibody has been adequately validated for a specific application”. However, the inclusion of several strategies would strengthen the evidence. The application of these principles have been demonstrated on over 6,000 antibodies validated for Western blots [146].
In short, *genetic strategies* entail the elimination or reduction of protein by knockout of the target gene or silencing of the corresponding RNA. The process should thus eliminate or drastically reduce the protein level reported by the antibody unless there is cross-reactivity to another target that was not affected by the treatment. The approach is applicable for example when studying cell lines, however, it is not possible to use in plasma or other human samples.

*Orthogonal strategies* involve the comparison of antibody-based methods with antibody-free methods. For instance, the same samples are analysed with both antibodies and MS, and the correlation between measurements is assessed. A high correlation between orthogonal methods is a strong indication of specificity towards a common target. It is however challenging to identify why two methods may not agree, such as if the two methods measure different protein versions of the target. In the case of the antibody, it is likely more difficult to know which specific proteoform is recognized, while in MS the detected peptide might be mapped back to a common sequence that does not include the particular genetic variant. Yet another emerging orthogonal validation approach is the comparison of protein data with genetic data [147]. A protein quantitative trait loci (pQTLs) can be identified if there is a SNP directly associated with the detected profiles of the target protein (Paper I-II). If the SNP is located at or in close proximity of the target gene, it is defined as a *cis*-pQTL. However, if the SNP is outside of the coding target gene, or within another gene entirely, it is classified as a *trans*-pQTL. The presence of a *cis*-pQTL is a strong indicator of on-target binding. However, *trans*-pQTLs are more difficult to interpret. This approach is applicable if the antibody-antigen epitope is modified due to a pQTL, while antibodies binding to other unaffected sites on the antigen cannot be identified and validated in this analysis.

*Independent antibody strategies* encompass the comparison of measurements obtained by two or more antibodies directed towards the same protein but targeting different epitopes (Paper I-V). As described in Chapter I, in the HPA workflow antibody siblings are generated that can be utilized for this type of validation in multiple assays, including plasma profiling and immunohistochemistry. SIA and PEA that rely on dual capture bindings are also a type of independent antibody strategy.

*Expression of tagged proteins* includes the design of proteins that carries an affinity tag (Paper IV) or a fluorescent protein for which well-validated reporter systems exist. The tag allows detection that can be compared with the antibody detection, and if the signals or visual pattern coincide, it strengthens the evidence of on-target binding. Discrepancies between the two detection systems may point at possible cross-reactivity. Important to note is that the tag itself might result in a differential protein conformation compared to the endogenous variant. Likewise, the tagged protein might differ in function and localization compared to the native protein.
Immunocapture followed by MS is a powerful tool for identifying off-target binding [62] and is described in Chapter II. In summary, antibody validation is necessary for research reproducibility and there are multiple strategies for providing evidence of on-target binding.

**Biological validation**

This chapter has so far described several considerations when embarking on an immuno-based proteomics study: study design, biological and technical (both pre-analytical and analytical) sources of variation, experimental design, antibody validation, and data processing. Bearing in mind this multitude of challenges, it is not surprising that many discovery studies fail biological validation, which entails replication in an independent set of samples [148]. However, successful biological validation is the first step towards demonstrating the molecular mechanism behind a finding.

Taking a protein candidate from discovery to biological validation often starts by analysing a smaller sample set and then increasing the number of samples along the verification progresses [45]. For research investigating disease-associated proteins, the long-term aim is to bring forward findings with a clinical utility that could be used in close connection to hospitals, clinics, or health-care centres [149]. Therefore, the initially used discovery method needs to be adapted or translated into those routine applications in clinical practice where samples are collected, processed, and evaluated continuously. There are extensive requirements for assays that are to be applied for disease diagnosis and the performance needs to match or surpass the current gold standard [150]. Relating this to the Luminex platform, the detection of gastrointestinal pathogens using the MAGPIX instrument has already been FDA cleared [151], hence demonstrating the clinical utility of such a system.

In conclusion, there are many steps in the immunoassay pipeline to consider prior and post data generation to ensure validity, as well as translation to a clinical setting. There is a vast amount of applications for proteomics assays and selected topics will be touched upon in Chapter IV.
IV. Applications

Broad screenings of the plasma proteome or the annotation of particular phosphorylation sites on a single protein are only two out of numerous examples of proteomics applications. A (very) simple query in PubMed (2020-08-06) for proteomics, or plasma proteomics, or affinity proteomics yielded >135,000, >8,000, and >6,000 hits, respectively. As a comparison, using the same search terms, Google Scholar returned about 1,180,000, 813,000, and 790,000 hits. Despite conceptual discrepancies between the two databases, this modest test illustrates the impressive scope of literature within the field. To set a reasonable limit, this chapter will bring examples of emerging applications of plasma proteomic studies. This includes the role of proteins in the efforts towards personalized medicine, as well as research integrating omics-analysis and big data. An outlook on future challenges and opportunities concludes the chapter.

Personal proteomics profiles

The concept of personalized medicine refers to the tailoring of healthcare according to the individual, in order to enable improved patient outcome. Risk assessment, diagnosis, and treatment are based on the individual's personal molecular fingerprint - a unique combination of factors such as genes, lifestyle, and environment. Personalized medicine or precision medicine are closely related concepts with emphasis on patient treatment strategies, while precision health focuses on wellness and maintenance of health [49]. Thus, the same distinction will be followed in this chapter, although the mentioned concepts do overlap in their aim of pursuing personal-based approaches to improve disease detection, prevention, and treatment.

Intra- and inter-individual diversity

The dynamic nature of the human plasma proteome is reflected in protein levels, ratios, and interactions that change with time. Humans and their molecular composition are controlled throughout growth [152], ageing [153], and the passing through phases, such as pregnancy [154] and menopause [155]. In Paper II, we studied proteins in plasma and their association with age. Changes in protein levels can also arise because of naturally occurring cycles to keep homeostasis, such as the internal sleep-wake clock (the circadian rhythm) [156] or seasonal variation [33]. Due to these and many other intrinsic differences that occur within every individual (intra-individual variability), there is also a large difference between individuals (inter-individual diversity). At baseline, that is prior to any intervention, the intra-individual molecular variation is generally
lower than the inter-individual diversity. We explored the concept of baseline protein levels of circulating proteins in Paper I.

When comparing protein profiles between for example patients and healthy controls, it is likely that protein levels from individuals within each respective group will differ. This can be due to factors that are tied to the personal proteome baseline, rather than caused by disease. As one study observed, 56% of 92 proteins were significantly associated with non-disease related factors, most predominantly explained by age, systolic blood pressure, and weight [147]. Additional factors that have been observed to attribute to the levels of proteins in plasma include for instance sex, BMI, and lifestyle. Interventions such as medication can also have profound impacts on the levels of circulating proteins [157]. A person’s genetic make-up further contributes to the inter-individual diversity observed in plasma proteomes [100]. Recently, characterization of 3.622 proteins by aptamer-based assays in >3.000 individuals revealed as many as 1.927 genotype-protein associations [158].

Not only are there many elements that contribute to the variability between person’s plasma proteomes, but there is also a temporal component to consider, adding to the complexity. This could be conceptualized as a time-dependent and personal trajectory that may change in response to disease or other perturbations [159]. Consequently, finding disease-associated proteins in plasma is challenging as there is a dynamic heterogeneity both within and between individuals. This property can impede conventional drug development because finding a single treatment that has effectiveness in all individuals without considering this heterogeneity is not a trivial task. The notion of inter-individual diversity has indeed led to the movement towards personalized medicine [160].

**Personalized medicine**

Why do we need personalized medicine? Traditionally, medical practice has been population-based, following the “one size fits all” approach. Individuals that share disease diagnosis are prescribed the same type of medication or treatment regime, although they may not be effective or can even cause severe side-effects. Today, it is recognized that treatment needs to be tailored to the individual patient [161]. Among many striking examples already in practice is the early detection of women with a genetic predisposition to breast and/or ovarian cancer. Mutations in the **BRCA1** or **BRCA2** gene are associated with a 70% higher risk of developing breast cancer, and early diagnosis is important for enrolment in intensified screening programs for tumour detection [162]. Personalized strategies have also been adopted for pharmacogenetics that aims to adjust drug administration according to an individual’s genetic information, in order to improve treatment effectiveness [163]. Among the maybe most well-known cases is the treatment of breast cancer with Trastuzumab, a therapeutic monoclonal antibody [164]. Breast cancer patients...
are tested for human epidermal growth factor receptor 2 (HER2) protein overexpression or HER2 gene amplification [165]. In patients with overexpressed HER2, the drug can be a viable treatment option. Other patients may however not benefit from the medication, in the worst case leading to adverse effects, and should be directed to other more suitable medications. It is clear that matching patients with personalized strategies is needed to ensure better patient care. Personalized treatments based on pharmacogenetics has in particular been adopted for patient tumour profiling, and an exploratory study of cancer antigens is described in Paper III.

At its core, precision health aims to characterize a personal baseline of health for every individual. Continuous or frequent monitoring of molecular longitudinal data can further enable health and risk assessments, relative to a person’s own baseline. As described in Chapter III, with longitudinal sampling an individual can serve as its own reference, for example by comparing data from the “present” self with a “past” self [166]. Ideally, longitudinal data allows for monitoring fluctuations over time, detecting transition into a disease state, or evaluating response to treatment [161]. This implies that molecular signatures pointing at an individual’s diseases state exist, targeted treatments are available, and that we have knowledge of natural fluctuation of the baseline that can occur within a time window (days, weeks, or years) [49].

Longitudinal monitoring of individual baselines by plasma proteomics is emerging. In a recent study, the impact of weight loss and maintenance was investigated in plasma proteomes of 52 individuals [167]. The authors identified individual-specific protein levels that varied considerably between participants, and the findings suggest possible patient stratification based on personal plasma proteomes. The importance of measuring plasma protein levels at baseline and after weight loss/maintenance phases was also observed in an independent study [168]. Other studies where longitudinal plasma monitoring have provided new insights on individual health profiles include the characterization of personal autoantibody repertoires [129], correlations between blood biomarker data [169] in healthy individuals, and immune system development in newborns [170].

**Omic data studies**

Molecules in the human body interplay, buffer and respond to outside stimuli in complex pathways and networks to maintain homeostasis. Proteins are only one of many players in these multifaceted regulatory systems that comprise human health. In order to get a more complete picture of cellular processes, the field of omics aims to relate layers of data from different biological -omics fields such as proteomics, transcriptomics, genomics, epigenomics, and metabolomics, which correspond to the study of proteins, RNA, DNA, epigenetic modifications, and metabolites, respectively.
**Plasma proteomics and omics**

Sometimes, the global analysis provided by one -omic is enough to elucidate disease aetiology. Monogenic diseases are rare diseases that can be directly linked to a single gene, and thus detection can be facilitated by genome sequencing alone. However, most disorders such as cancer, diabetes, and cardiovascular diseases are believed to be highly complex in their establishment and progression, and co-morbidities can occur. These complex diseases depend on the interplay between multiple dynamic systems, and here omics studies may shed some light. With the advances in high-throughput and multiplexed technologies, omics studies can connect different -omes to provide a detailed description of many relevant biological processes, even on the level of an individual [171]. For example, food allergy has been described as an interplay between a person’s genetic information, environmental exposures, and lifestyle [172].

Seeing as the bulk genome is highly constant across one lifetime, the addition of proteomics data can further pinpoint ongoing processes in cells, tissues, and organs. Large-scale genome-wide association studies (GWAS) have identified genetic variants with functional consequences in complex diseases. However, as many of these variants are outside of protein-coding genes, it is challenging to find the mechanistic link between GWAS hits with disease phenotypes [18]. Combining genomics with proteomics can provide valuable new insights. As demonstrated by a comprehensive study of 1.000 individuals, about 20% out of 1.124 proteins had a pQTL, and 83 unique GWAS disease associations were found [173]. Thus, with pQTLs analysis, the connection between the genome and complex phenotypes can be elucidated via the plasma proteome. Proteomics data can also be complemented with transcriptomics, in order to infer the relation between RNA and protein expression [174]. Omics studies that include genomics, proteomics, and transcriptomics data can allow for both eQTL and pQTL analysis, providing deeper insights into molecular processes [158].

Given that circulating proteins do associate with disease phenotypes, blood remains a valuable source for prognostic biological information even as we move into omics analysis. With longitudinal monitoring, personal health trajectories could be predicted [159]. As a consequence, recent studies in the plasma proteomics field have combined protein and omics data to describe personal molecular profiles [47, 153, 175, 176].

**Generating big data**

Advances in computing power and high-throughput analytical methods have also resulted in a surge of big data collections, enabling high-dimensional personalized medicine [177]. Big data is often defined by three characteristics known as the 3Vs, which refers to the volume of data, the
variety of data types and structures, and the velocity (producing and processing data) [178]. Thus, these datasets can contain a vast amount of measurements, both per individual and analyte. Already the field of omics can generate big sets of biological data utilizing for example DNA, RNA or protein microarrays. Coupling that to longitudinal sample measurements, the amount of data points increases even further.

New “non-omics” [179] big data collections are emerging from sensors, smartwatches, phones, and other wearables [180]. These devices allow for continuous monitoring of vital signs such as blood pressure, glucose levels, and heart rate. The introduction of personal activity trackers has also engaged individuals to monitor their own health, pursuing well-being, or wellness. As a consequence, the field of plasma proteomics has also gained interest in profiling individuals in the manner of personal health monitoring [175]. Combined with clinical data and omics analysis, these new non-omics tools have potential value in the quest of delineating human health and disease. Furthermore, models that are based on integrated data may allow disease reclassifications, providing more nuanced disease groupings.

Future challenges

Although highly interesting findings have been presented in the context of plasma proteomics combined with other omics and non-omics data, it still remains challenging to implement personal omics-profiling into the clinic [179]. This can also be largely due to the sheer number of implicated variables that can mask or distort true associations. Chapter III has highlighted some of the hurdles that can impede the discovery of biologically meaningful disease marker. Here, the focus is instead on challenges that can be encountered when integrating big omics data, as well as concerns regarding individualized health monitoring. Some of the overarching considerations for personalized medicine and omics analysis are listed in Box 3.

Box 3: Omics data considerations.

<table>
<thead>
<tr>
<th>DATA</th>
<th>• Quality</th>
<th>• Integration</th>
<th>• Infrastructure</th>
<th>• Ownership</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HEALTH CARE</td>
<td>• Patient contact</td>
<td>• Over-testing</td>
<td>• Reclassification of disease</td>
<td></td>
</tr>
<tr>
<td>PATIENT</td>
<td>• Privacy</td>
<td>• Information overload</td>
<td>• Interpretation</td>
<td>• Cost and inequality</td>
</tr>
</tbody>
</table>

Omics studies incorporate data generated at multiple sites, by different people, and during different times. Thus, prior to any analysis, potential batch effects or systematic biases need to be addressed to ensure high-quality data. Failing to do so can lead to non-reproducible results, error propagation, and incorrect interpretation of biological observations. Data integration refers here
to the combination of different omics data in order to explain or predict disease risk, health trajectories, treatment outcomes, or other biological variables. The choice of appropriate data-integration method depends on the hypothesis of the biological mechanisms that generate a particular phenotype [181]. Is there an assumed linear relationship between variables? Or, are non-linear relationships expected, such as the presence of interactions, feedback loops, and compensatory mechanisms? By incorporating multiple omics data types into the appropriate model, these questions can be investigated [181]. Further, the generation of large data sets requires an infrastructure for dynamic knowledge repositories, where privacy and security of data storage can be ensured. In these repositories, data from one individual could be cross-referenced with similar health profiles in order to determine risk factors or get records of successful interventions or treatments [182]. When it comes to personal health monitoring, data privacy, safety, ownership, and accessibility needs to be carefully defined. Medical research has clear guidelines and rules when it comes to the ethical principles of studies involving human subjects [183]. However, similar solutions that take big data privacy into account are needed before implementing personal omics-profiling into the clinic [184]. With new infrastructures and data anonymization, a researcher might collaborate on big data where expertise from different omics fields is required.

From a clinical point of view, omic studies on healthy, asymptomatic individuals may lead to over-testing or overdiagnosis, as compared to when a patient seeks medical help when there is suspicion of disease [185]. With current diagnosis methods that all have limitations in specificity and sensitivity, over-testing will unavoidably by chance lead to positive results in some individuals without disease (false positives). This in turn may lead to overtreatment. However, with the development of appropriate methods that have improved reliability and reproducibility, the problem of false positives and negative may be reduced [186]. Finally, longitudinal monitoring where individuals actively can follow their health profiles will require platforms where a patient can consolidate with physicians. Health data needs to be informative and possible for patients and physicians to interpret while avoiding information overload. Personalized diagnosis and therapy need to be cost-effective and not a burden for the individual, in order to ensure equal treatment for everyone.

To summarize, there are many challenges with the advance of big data collections, omics collaborations, personalized medicine, and health monitoring. Strategies concerning data integration, interpretation, and sharing need to be assessed in every project. Analysis of accessible circulating proteins can add value to these omics studies by elucidating the interplay between molecules and drivers behind human health and disease.
V. Present investigation

The focus of the work presented in Paper I-IV has been on the application of antibody-based affinity assays in the context of health and disease. The aim of the studies has been formulated based on the SBA technology, as well as the availability of biological specimens, clinical data, and the resource of affinity binders provided by the Human Protein Atlas. Additionally, a central part of the studies has been follow-up experiments for antibody validation and biological verification.

The work presented here has entailed the study of proteins in blood-derived plasma and serum from clinically healthy individuals (Paper I), elderly individuals (Paper II), and lung cancer patients (Paper III). Furthermore, GPCR-RAMP protein complexes have been studied in cell lysates (Paper IV). Three different array formats have been utilized; the single-binder assay for profiling proteins (Paper I-II), the reverse-phase array for identifying IgG reactivities and characterizing antigens (Paper III), and the sandwich immunoassay for detecting protein-protein interactions (Paper IV). By applying these methods, we have identified proteins and autoantibodies that could be further studied to delineate markers of health and disease.

Paper I: Facets of individual-specific health signatures determined from longitudinal plasma proteome profiling

With the advent of precision medicine, it is becoming common practice to explore not only global trends of health and disease in a population but also to measure these on the level of an individual. While there is an obvious and needed focus on studying disease, clinically healthy individuals remain largely uncharacterized. To gain insights on personal plasma proteomes, we set out to explore longitudinal effects on circulating proteins in healthy individuals. Here, 101 individuals from the Swedish CArdioPulmonary bioImage Study (SCAPIS) were followed over one year. EDTA plasma samples were collected longitudinally as participants visited the clinic every third month. In total, four blood samples per individual were collected (Figure 5A). Out of the 101 participants, 99 completed all four visits.

We analysed each collected plasma sample with multiplexed antibody suspension bead arrays (SBA) (Figure 5B). Antibodies were provided by the Human Protein Atlas (HPA) or obtained from commercial vendors. From a set of 1,483 antibodies that targeted almost 900 unique proteins, we selected 734 protein profiles for downstream statistical analysis. The selection was
based on an antibody score we developed while assessing antibody performance and reliability. The scores considered an antibody's reproducibility in repeated assays, correlation to internal control beads, as well as independent antibodies sharing the same protein target. We also used orthogonal methods including MS and GWAS. The downstream analyses included associations to clinical traits, longitudinal trends of protein profiles, as well as comparing the genetic with the protein data (Figure 5C).

We deemed that 49% (359/734) of the proteins were stable across one year. Global analysis of protein profiles revealed that each participant had a stable and person-specific plasma proteome profile, which was further retained thought all visits. Thus, we observed a diversity of protein profiles between these clinically healthy individuals. Next, we explored the presence of protein profiles that co-varied in the study population. We identified that eight groups of 11-242 proteins were part of longitudinally preserved patterns. We annotated the groups for their biological function and association with clinical variables.

Comparing genetic data with protein profiles, we identified 15 cis-pQTLs for 14 unique proteins. Interestingly, we observed an enrichment of non-synonymous SNPs that linked to protein polymorphisms. Some of the selected findings were validated in a set of 3.000 samples from elderly twins collected by the Swedish Twin Registry [29].

Referencing each individual protein measurement to the study population mean, we systematically annotated the studied protein profiles into three groups: elevated or downregulated at baseline, had an increasing or decreasing trend over time, fluctuated up or down during particular visits (Figure 6A). Comparing the trajectories of protein profiles, we found that each individual had at least one protein that differed from the population in terms of baseline, trend, or fluctuation (Figure 6B-C). This finding supports our initial observation that there is a diversity among clinically healthy individuals and additionally showed the temporal component of the personal plasma proteomes.

In conclusion, we profiled circulating proteins of 101 clinically healthy individuals during one year with antibody assays. We observed both individual-specific protein expression, as well as global trends of longitudinally preserved protein groups. In our study, we found that plasma proteomes were predominantly stable during the study period. However, we also observed that intra-individual diversity needs to be considered together with longitudinal variability when assessing personal health profiles. Integrated with detailed clinical data, our study approach could allow for defining normal variation in healthy individuals and assist in detecting perturbations or transitions into a disease. Future work will entail the integration of other omics data obtained from the same participants, the addition of two more visits, as well as contrasting this wellness
cohort with other individuals from SCAPIS that have pre-diabetic characteristics, or cardiovascular disease, or non-alcoholic fatty liver disease.

Figure 5: Experimental design and data analysis workflow. (A) Samples were collected during one year at four visits. Each dot corresponds to one individual. (B) EDTA plasma was analysed with SBA assays. (C) The experimental design involved four separate assays where samples were analysed after each completed visit. For replication purposes, samples from subsequent visits were analysed together with all previously collected visits. Four SBAs were applied in the analysis. Bold labels indicate assays that included 96 replicated samples and underlined labels indicate assays where all samples were analysed in duplicate, both for technical validation. (D) Data analysis workflow included the analysis of protein profiles and associations to clinical traits, longitudinal variation, networks of co-variation, and GWAS. SBA, suspension bead array; GWAS, genome wide association study.
Figure 6: Longitudinal variability. (A) Venn Diagram reporting the number of proteins with the annotation criteria: elevated/decreased throughout one year (baseline), increased/decreased linearly with time (trend) or at one particular visit (fluctuation). Selected protein examples are on the right. Each grey line corresponds to one individual, while the red line highlights one individual with a selected protein profile. (B) Distribution of the annotation criteria per individual. (C) The sum of proteins that fulfill the annotation criteria per individual.
**Paper II: Profiles of histidine-rich glycoprotein associate with age and risk of all-cause mortality**

Ageing is a major denominator for many common disorders, such as cardiovascular diseases, different types of cancer, type 2 diabetes mellitus, and Alzheimer’s disease. Interestingly, little is still known about age-dependent molecular processes and how these are reflected on the levels of proteins that circulate in human blood. Clinical tests often rely on blood analysis as it is an accessible source for health and disease monitoring, and thus identifying any effects of age is valuable for the understanding of non-disease factors that impact on blood proteins. By using multiplexed antibody bead arrays, we explored proteins in serum and plasma in relation to age.

An initial screening was performed in a set of 156 individuals aged between 50-92 years from the TwinGene cohort. With 7.258 HPA antibodies corresponding to 6.370 unique protein targets, we found one protein profile (HPA045005) to be significantly associated with age ($P = 4.69 \times 10^{-5}$). Next, we set out to validate our finding and confirmed the observation in eight additional sample sets. The largest sample set consisted of sera from 3,000 elderly twins from the TwinGene study, and these were analysed with a new SBA of 384 antibodies. We compared our proteomic and genetic data and found a protein quantitative trait loci (pQTL) to histidine-rich glycoprotein (HRG) (Figure 7A-B). The most significantly associated genetic variant ($P = 2.35 \times 10^{-97}$) was a SNP encoding a non-synonymous mutation. The utilized SBA contained an addition antibody towards HRG (BSI0137, BioSystems International Kft). Interestingly, while HPA045005 profiles were highest for the CC genotype and decreased as the number of C alleles decreased, the BSI0137 profile had an inverse trend with lowest levels for the CC genotype (Figure 7C). The HRG profiles obtained using HPA045005 were found to be associated with both ageing and mortality, but that was not the case for profiles measured with BSI0137.

To further validate that our main antibody (HPA045005) indeed bound to and captured circulating HRG, a sandwich immunoassay was developed in house. Two working antibody pairs for the capture and detection of HRG were established: HPA045005 or HPA054598 for protein capture and biotinylated HPA054598 for detection. Both antibody pairs demonstrated concentration-dependent detection of a full-length recombinant HRG. To assess antibody selectivity, we used a large protein microarray containing >10,000 antigens, among them the antigen that was used to raise HPA045005. There, the antibody exclusively bound to its corresponding antigen.
Figure 7: GWAS analysis of HRG protein profiles in plasma. (A) Manhattan plot of HPA045005. The X-axis denotes chromosome and the Y-axis is $-\log_{10}(P\text{-value})$. The line corresponds to the significance cut-off that was set to $P=0.01$ after Bonferroni correction. (B) A zoomed-in figure (LocusZoom) of chromosome 3 and the locus associated with HPA045005. (C) Boxplots reporting the relation between HRG proteins HPA045005 and BSI0137 measured with SBA and the genotypes in the SNP.
With our study, we hope to contribute to the knowledge of age-dependent proteins as we demonstrated by the association of plasma profiles of HRG and age. HRG is a highly abundant plasma protein and has been ascribed different functions. Further work will be necessary to disentangle the mechanism behind the reported results, in particular for the two HRG antibodies which showed an opposite association of protein abundance and SNP. With our current data, we cannot propose a mechanism for why HRG profiles increase with age or how it relates to mortality. However, future studies could build on the work presented here in order to find the mechanistic link between HRG and ageing. These studies should preferably consider genetic variability of HRG and the molecular consequences of these on its function and interaction with other proteins.

**Paper III: Detection of autoantibodies against cancer-testis antigens in non-small cell lung cancer**

Cancer testis antigens (CTAs) are proteins primarily expressed in healthy testis and placenta. In cancer, these proteins can be detected in the affected tissue. CTAs have furthermore been classified as tumour antigens and could serve as targets in immunotherapy. To this end, we studied the occurrence of CTA autoantibodies in plasma of NSCLC patients with multiplexed bead arrays. By identifying CTAs that are present in a majority of studied cancer patients, we hoped to provide insights on which markers could potentially aid in early detection of tumours in screening efforts.

In our study, we investigated CTA reactivity in plasma of NSCLC patients from two cohorts. The primary sample set included 133 NSCLC patients and 57 cases with benign lung diseases, and samples were collected prior to surgery. The second sample set was used for validation and included independent samples from 34 NSCLC patients from another cohort. Seven patients had samples collected longitudinally. Patient samples from both cohorts were obtained from Uppsala University Hospital.

For the detection of circulating IgG antibodies in plasma, we designed suspension bead arrays presenting the CTAs in form of protein fragments produced by the Human Protein Atlas. CTA antigens that have previously been reported to be expressed in NSCLC were selected and included in the array. In total, 144 antigens were chosen for the analysis of CTA reactivity in plasma, while for the statistical analysis 120 antigens representing 112 unique CTAs remained.
In our assay, reactivity was detected for 69 of the 120 antigens in either NSCLC or controls from the primary cohort. No obvious association was found between reactivity levels and any of the available clinical data such as sex, age, and smoking. However, a total of 29 antigens demonstrated reactivity exclusively in NSCLC patient samples, and these candidates were selected for validation in the second NSCLC cohort.

Reactivity towards six of the 29 CTAs was confirmed in both NSCLC patient sample sets (Figure 8). This included the P antigen family member 3 (PAGE3), cancer-testis antigen family 47; member A (CT47A) genes, variable charge X-linked (VCX), melanoma antigen family B1 (MAGEB1), lin-28 homolog B (LIN28B) and chromosome 12 open reading frame 54 (C12orf54). The reactivities were observed in patients at a frequency of 1-4%, and these antigens were then further evaluated in seven patients for which longitudinally collected samples existed (Figure 9). We observed that the reactivities for each patient were conserved across time and sample collection.

![Figure 8: CTA reactivity in NSCLC patients and controls. Each dot represents one individual, and patients classified as reactive for a particular CTA are highlighted in red. The asterisk denotes that the protein fragment corresponds to several members of the CT47 family. Signal intensities are measured in MFI and converted to binned values where 0 corresponds to no reactivity and values above 1.5 are defined as reactive. MFI, median fluorescence intensity.](image-url)
In conclusion, we applied multiplexed suspension bead arrays to screen for circulating antibodies towards CTAs in NSCLC patients. We designed an exploratory array and identified several CTAs with reactivities exclusively found in patient plasma. Six of these antigens could be validated in an additional set of independent NSCLC samples. Future work could continue the comparison by staining tissue samples from the respective NSCLC patients to extend our understanding of where the autoantigens are located in NSCLC tumours. Moreover, technical confirmation should consider including different antigen representations, such as full-length proteins or overlapping peptides of the investigated proteins, as these would further support the validity of our findings.

Figure 9: Heat map of CTA reactivity in seven NSCLC patients with multiple samples collected over time. Signal intensities are measured in MFI and converted to binned values where 0 corresponds to no reactivity and values above 1.5 are defined as reactive. MFI, median fluorescence intensity.
**Paper IV: Multiplexed analysis of the secretin-like GPCR-RAMP interactome**

G protein-coupled receptors (GPCRs) is a large family of transmembrane proteins found in eukaryotes. The GPCRs mediate signalling through a number of complexes and pathways. Various physiological processes are controlled by GPCRs as they can bind odour molecules, neurotransmitters, and hormones. It is estimated that about 30% of marketed drugs act by binding to GPCRs [187].

Receptor activity modifying proteins (RAMPs) is another family of three proteins that interact with GPCRs at cell surfaces. RAMPs modulate the activity of GPCRs, influencing downstream signalling and ligand specificity. As many GPCR-RAMP interactions largely remain uncharacterized, we set out to develop an antibody-based assay for identifying GPCR-RAMP complexes. Because of previously unreliable GPCR antibodies reported on the market [188], we also put an effort into carefully validating the applied GPCR antibodies, the majority provided by the Human Protein Atlas.

To study RAMP-GPCR complexes, a set of 23 GPCRs and three RAMPs were selected for the development of a novel multiplexed protein-protein interaction assay built on suspension bead arrays (SBA). Each GPCR was expressed in combination with one RAMP, and epitope-tagged protein constructs were expressed in HEK293F cells (Figure 10A). Following transfection, the proteins were then solubilized. A total of 55 antibodies targeting the 23 GPCRs and nine antibodies targeting the three RAMPs were included in the SBA (Figure 10B). All antibodies were designed to bind to the native sequence of the proteins. However, to confirm protein expression and validate antibody specificity, we included four monoclonal antibodies against the engineered epitope tags HA, 1D4, FLAG, and OLLAS (Figure 10C). Finally, in a sandwich assay set-up, we mapped GPCR-RAMP complex formations via protein-specific antibody capture and anti-epitope antibody detection.

We applied several capture and detection antibody pairs to identify and validate GPCR-RAMP complexes. All antibody pair combinations for the detection of each GPCR and RAMP are summarized in Figure 11. Out of the 23 GPCRs, all except three appeared in complexes with at least one RAMP. We identified previously reported complexes and also found evidence for 15 GPCR-RAMP complexes that have not been described previously. Five GPCRs were selected for validation with proximity ligation assays to confirm expression in membranes. Four out of the five GPCRs were detected in complex with RAMP2, which was consistent with the SBA results.
Figure 10: GPCR-RAMP interaction assay. (A) The GPCRs and RAMPs were epitope-engineered with a tag at their N-terminal and C-terminal. (B) Antibodies were coupled to magnetic, colour-coded beads and pooled to create a multiplexed SBA. GPCR and RAMP constructs were co-expressed in mammalian cells, solubilized and the cell lysates were incubated with the SBA. Protein detection was confirmed via engineered epitope tags and anti-tag PE-conjugated antibodies. (C) Using a Luminex instrument, protein expression, antibody validation, and complex detection were determined. SBA, suspension bead array.
Figure 11: Mapping GPCR-RAMP interactomes. Each circle corresponds to one GPCR, and the lines show detected interaction to each of the three RAMP1 (grey), RAMP2 (green), and RAMP3 (orange). Capture antibodies are labelled with their bead ID number, and epitope tag antibodies are indicated with their name (HA, 1D4, FLAG, and OLLAS). The thickness of the line corresponds to the relative statistical significance where P ≤ 0.05 is given an arbitrary thickness of 1, P ≤ 0.01 a thickness of 2, P ≤ 0.001 a thickness of 3, and P ≤ 0.0001 a thickness of 4.

In summary, we developed a novel multiplexed protein interaction assay for the detection of GPCR-RAMP complexes. We show that the number of complexes between these two protein families extends beyond what has previously been appreciated. As drugs targeting GPCRs most often focus on a single receptor, our findings could have implications for drug development and the targeting of protein-protein complexes. Future work will entail the systematic investigation of even more GPCRs with the three RAMPs. With a portfolio of validated antibodies against native GPCR and RAMP sequences, our approach also has the potential to be translated into tissue samples where endogenously expressed GPCR-RAMP complexes can be studied.
Concluding remarks and future directions

This thesis has aimed to provide a glimpse into the world of proteins and their utility as markers of health and disease. The focus has been on the application of affinity-based assays, in particular the suspension bead array technology, as a tool for characterizing proteins and autoantibody profiles. Further, the chapters here have highlighted some of the many factors that are important to consider in affinity proteomics when it comes to reproducibility and validation.

The proteomic studies presented here can be viewed as a small piece in the large puzzle that makes up our knowledge of human molecular biology. Proteins may provide clues about disease aetiology, or they can be used as clinical biomarkers or drug targets. However, proteomics combined with other layers of omics-data will be necessary to build truly comprehensive maps of all factors that may come together and cause disease. For instance, multifaceted processes such as wellness, ageing, and cancer can manifest due to the accumulation of numerous molecular interactions, where proteins are one of several key players.

Focusing on proteins, there is a large diversity in protein expression that exists both within and between individuals. This wealth of information may be used to deepen our understanding of human biology, however, also poses a challenge in protein analysis. After all, protein activity and abundance can be influenced by analytical, biological, and environmental factors. Therefore, it remains challenging to identify robust markers that may be reliably used for medical decision-making. The characterization of personal protein and autoantibody repertoires may prove to be the next step forward in personalized disease monitoring, treatment, and prevention.

Here, Paper I-III describes an exploratory approach for studying proteins and autoantibodies in blood-derived serum and plasma. It is important to note that the applied assays do not provide absolute quantification of protein abundance, and the detected proteins may consist of off-targets, co-targets, or even protein complexes that circulate in the blood. Thus, to evaluate the performance of our assays, selected antibodies were further investigated following the recommendations by IWGAV. We validated antibody selectivity using independent antibody strategies and orthogonal methods. Additionally, we succeeded in replicating some of the biological findings in independent sample sets. Although these validation approaches may increase the confidence of on-target binding, the assay’s performance and the downstream biological interpretation still remain context and application-dependent. Future studies with new samples, metadata, or affinity-binders may strengthen the observations made here, or assess how generalizable the results are in other study populations.

In Paper IV, we developed a multiplexed dual-binder assay for the detection of GPCR-RAMP complexes in cell lysates. While the method was based on the SBA assay, the sandwich format
allowed for a higher degree of evidence of on-target binding. To validate our applied antibodies, proteins were expressed and engineered to carry affinity-tags. Thus, antibody-mediated detection of protein complexes was possible by antibodies that either bound to the endogenous protein sequence or the affinity-tag. We performed a systematic validation of the applied antibodies, and we confirmed the formation of protein complexes in cell membranes by PLA. However, further studies will be necessary to determine if our findings also hold true in a living cell, where proteins are neither overexpressed nor carrying affinity-tags.

In summary, the work presented in this thesis describes approaches for studying proteins in blood-derived samples and cell lysates. Further follow-up studies are encouraged to determine the broader potential and possible clinical utility of the findings presented here.
I would like to thank all the fantastic people who I have worked with these past years. Without you, this thesis would not have been possible. I also would like to express my sincere gratitude to all blood donors and everyone who contributed to sample collection and preparation – you make our research into reality.

I would like to acknowledge the KTH Center for Applied Precision Medicine (KCAP) funded by the Erling-Persson Family Foundation for supporting the work presented in this thesis, and the Knut and Alice Wallenberg Foundation funding for HPA.

I especially would like to thank my main supervisor Jochen, and co-supervisors Mun-Gwan and Peter. Jochen, thank you for welcoming me to your research group, for giving me the opportunity to be part of so many exciting scientific endeavours, and for guiding and inspiring me to grow as a scientist. During my master thesis and PhD studies, you always made time to help me and provided feedback, no matter how small or large the matter might be, which I am very grateful for. Thank you for your support, trust, and for always encouraging me to ‘be brave’! Mun-Gwan, thank you for introducing me to R and the world of data analysis, and for supporting me during my master thesis and as a PhD student. I always look forward to our monthly meetings where talk about life, science, and statistics! Peter, tack för dina kloka och insiktsfulla råd, för att du ställer de rätta frågorna, och för din positiva energi som du delar med dig av!

Peter and Jochen, I want to thank you both for the PAPP/Biobank Profiling (Profilong)/Affinity Proteomics/… group - Kätt barn har många namn! It is an truly inspiring and unique environment to work in, and I would like to thank all the fantastic members of the PAPP group: Annika och Matilda, tack för att ni forgyller tiden på jobbet! Att umgås med er innebär många glada skratt och diskussioner om allt mellan himmel och jord. Tack även för de otaliga gånger ni hoppat in och hjälpt mig i olika projekt. Matilda, tack för att jag fick handleda ditt exjobb, för att du är en riktig klippa man alltid kan lita på, och för dina kloka råd och tips. Annika, tack för din nyfikenhet, din värdefulla feedback på allt ifrån plottar och dataanalys, och för alla roliga diskussioner kring spel och filmer. Cecilia ET, mange tak för alla projekt och dataanalys som vi klurat på tillsammans, för våra trevliga kaffepronommenader och yoga-pepp! Jag är så glad att du valde att göra din postdoc i vår grupp så vi fick lära känna varandra. Vi ses snart igen, i Köpenhamn eller Stockholm! Cecilia M, tack för alla roliga samtal och labbhjälp, och för att du delar med dig av ditt lugn och positiva inställning! Claudia, thank you for sharing your expertise.
and insights in different projects, and for keeping the facility running! Leo, så roligt att du anslutit dig till vår lilla grupp, det ska bli spännande att följa din forskning. Julia, tänk att vi började på KTH tillsammans nu för 10 år sedan! Tack för att jag fick dela min doktorandresa med dig, för alla roliga och välbehövda fikapauser, R snack, kurser, och för din uppmuntran och kloka ord – nu tar vi tag i bokklubben igen! Cecilia H, jag är så glad över de projekt vi samarbetat i, de har alltid varit väldigt roliga och lärorika. Tack för alla gånger som du tagit dig tid att hjälpa mig och för hårligt häng utanför jobbet med konserter, ljusfestival och spelkväller på Dice! David J, Jennie, och Sofia, stort tack för er värme, sprudlande energi, engagemang och alla roliga stunder på och utanför jobbet! Är det inte dags för en brädspelskväll snart? Eni, tack för att du roddar och har koll på allt i labbet, för det massiva PrEST-plock vi utförde tillsammans förra sommaren, och för den värme du delar med mig av. Anna, August, Elisa, Jamil, Lovisa S, Ronald, Sára, Shaghayegh, och master thesis students, I would like to thank each of you for making our lab such a great place to work at! I would also like to thank all previous PAPP members, especially: Arash, Burcu, Elin, Julius, Kimi, Lucia, Mahya, Maja, Maria, Philippa P, Philippa R, Ragna, Sanna, and Ulrika. Ett extra stort tack till Elin för att du tog dig an mig som exjobbare och lärde mig allt om beads, antikroppar och PrESTar!

A big thank you to my other colleagues at SciLifeLab: Maria-Jesus, thank you so much for all the fun times we had together in the lab, for your kindness, and for always being so thoughtful. I especially want to thank you for your help during my master thesis and PhD, and for always letting me borrow your reagents and antibodies in the lab! Christian, tack för ditt glada humör och sällskap, dina tyska post-it lektioner och AWs! Tack också för att du fick mig att lära känna Zaneta och Alma – thank you all three for the adventures we had in SF! David K, thank you for sharing the trip to Brixen with me, it was a lot of fun! Thank you Andreas, Björn, Casper, Devin, Diana, Fredrik, Laura, Lovisa Å, Max, Philip and all past and current members of Alpha 2, CAP, Cell Profiling and QPrEST group for contributing with a fun, positive atmosphere and great science!

Thank you all members of the Human Protein Atlas and Atlas Antibodies for creating the valuable resource of antibodies and PrESTs that make our research possible. Mathias, tack för att du startade HPA och för all forskning som du möjliggjör. Speciellt Wellnessprojektet har varit mycket spännande och lärorikt, jag är tacksam över att ha få tagit del av det.

I would further like to thank all my collaborators and co-authors. Thank you Patrick and Dijana for the collaboration on CTAs, our fun meetings (med tårta!) and data analysis discussions. Dijana, tack för våra träffar på och utanför labbet, dina goda råd kring doktorandlivet och för ditt glada humör som alltid smittar av sig! I would like to thank everyone in the Castleman’s disease project for giving me a fantastic summer at the Utz lab at Stanford. PJ, David, and Rong, thank you for making this possible! Sarah and Allan, thank you for your company and
help in the lab, like real detectives we solved all Luminex and BioTek mysteries! Tom and Emily, thank you for introducing me to the exciting field of protein interactions and GPCRs. Patrik and Sara, tack för samarbetet under AGE-projektet, för de intressanta diskussionerna och möten som jag lyssnat in på. Linn, tack för samarbetet under Wellnessprojektet och din hjälp kring alla möjliga frågor, allt ifrån blodprov som förväxlats till olika normaliseringsmetoder.

Thank you Cecilia W and Cristina for the valuable feedback you gave me during my 2-year seminar.

A big thank you to Jochen, Mun-Gwan, Claudia, Philippa P, and Julia for reading and providing valuable input on the different parts of my thesis.

Tack till alla mina fina vänner utanför forskningsvärlden som jag får dela så många roliga stunder med. Tack till alla i TMK ( torsdagsmiddagsklubben) – vilket amazing gång ni är! Min bästa vän Sofie, tack för allt härligt häng och skratt, för att du alltid finns där och lyssnar, uppmuntrar, och sätter saker och ting i rätt perspektiv!

Slutligen vill jag taka min kära familj, Gordana, Ivica, Luka, Louise, August och Ada. Tack för all kärlek och värme, för att ni alltid finns där och lyssnar, stöttar, ger råd, och peprar! Jag ser fram emot när vi alla bor lite närmare varandra och kan ses oftare. Jesper, tack för din kärlek, stöd, skratt, trygghet, och för att jag får dela allt med dig ♥
References


Dean I, Information NCIaB. Blood Groups and Red Cell Antigens: NCIaB; 2005.


