olinska
itutet

From the Department of Medicine, Solna
Respiratory Medicine Unit
Karolinska Institutet, Stockholm, Sweden

BetweeN TwO LungS

PROTEOMIC AND METABOLOMIC APPROACHES IN INFLAMMATORY LUNG DISEASES

Tina Heyder

Karolinska
Institutet

Stockholm 2017

Between Two Lungs

PROTEOMIC AND METABOLOMIC APPROACHES IN INFLAMMATORY LUNG DISEASES

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Tina Heyder

Principal Supervisor:
Associate Prof. Åsa Wheelock
Karolinska Institutet
Department of Medicine, Solna
Respiratory Medicine Unit

Co-supervisor(s):
Prof. Johan Grunewald
Karolinska Institutet
Department of Medicine, Solna
Respiratory Medicine Unit

Prof. Roman Zubarev
Karolinska Institutet
Department of Medical Biochemistry and Biophysics
Division of Physiological Chemistry I

Assistant Prof. Jimmy Ytterberg
Karolinska Institutet
Department of Medicine, Solna
Rheumatology Unit
Centre for Molecular Medicine

Opponent:
Prof. Peter Sterk
University of Amsterdam
Department of Respiratory Medicine
Academic Medical Centre

Examination Board:
Associate Prof. Mikael Adner
Karolinska Institutet
Institute for Environmental Medicine
Unit of Experimental Asthma and Allergy Research

Prof. Sophia Hober
Royal Institute of Technology
School of Biotechnology
Division of Protein Technology

Prof. Jonas Bergquist
Uppsala University
Department of Chemistry
Analytical Chemistry and Neurochemistry

The public defence for the degree of "Medicinsk doctor", Doctor of Philosophy (Ph.D.) in Medical Science will be held on the 24th of August, 2017, at 9:00 am in the Rockefeller lecture hall (Nobelsväg 11, Karolinska Institutet, Solna Campus).

ABSTRACT

This doctoral thesis presents results on proteomic and metabolomic profiling of two distinct inflammatory lung diseases: sarcoidosis and chronic obstructive pulmonary disease (COPD). Chronic inflammatory lung diseases are debilitating conditions representing a major health problem throughout the world. However, up to this date disease pathobiology is not fully understood and clinical practise is lacking early diagnostic tools. The overall aim of the work has been the investigation of disease specific biomarkers as well as underlying molecular pathways. In addition, the characterization of clinical sub-phenotypes will provide deeper insights into the molecular mechanisms involved in disease pathophysiology. Moreover, for the purpose of profiling the immunopeptidome in sarcoidosis, an approach was developed to identify immunopeptides presented on antigen presenting cells with high sensitivity.

The first part of this thesis focuses on sarcoidosis, an inflammatory T cell driven disease that mainly manifests in the lungs. It is hypothesised that specific antigenic substances trigger onset of sarcoidosis. Because lung samples from healthy non-smokers typically contains 10–15×10^6 cells, there was a need for a sensitive approach to identify immunopeptides. **Project I** describes the optimization of an approach designed specifically to investigate the immunopeptidome on human leukocyte allele antigens (HLA)-DR from scarce clinical material. Epstein-Barr Virus immortalized B cells and antigen presenting cells obtained from bronchoalveolar lavage of sarcoidosis patients were utilised for investigating the HLA-DR immunopeptidome. The approach presents a valuable tool to profile and compare peptide repertoires in health and disease in order to reveal disease specific antigenic peptide(s) in sarcoidosis.

Project II is based on the hypothesis that sarcoidosis patients potentially exhibit an expansion of certain antibody clones, supporting the concept of induced activation against a disease specific antigen. Specific features of polyclonal antibodies could be utilized for disease characterization via blood testing. The IgG Fc-galactosylation status in matched bronchoalveolar lavage fluid and serum samples was very well correlating. In particular, the ratio between the main agalactosylated (FA2) and main di-galactosylated Fc-glycans (FA2G2) of IgG$_4$ could successfully distinguish both sarcoidosis phenotypes and controls.

The second part of this thesis concerns investigations of early stage COPD with a particular focus on identifying molecular gender differences related to distinct disease characteristics and clinical phenotypes between women and men. Disease related alterations in the airway epithelial proteome were characterised using two complementary proteomic platforms (**project III**); two-dimensional difference gel electrophoresis (2D-DIGE) and tandem mass tag-based shotgun proteomics (TMT-MS). Additionally, serum metabolites from the same

study cohort were analysed with non-targeted liquid chromatography-high resolution mass spectrometry (**project IV**). Both molecular investigations revealed specific alterations between early stage COPD patients and smokers with normal lung function mainly driven by the female population in the cohort. Gender-enhanced alterations in xenobiotic metabolism were revealed in the airway epithelium as well as changes in oxidative phosphorylation and protein processing in the endoplasmic reticulum. Non-targeted metabolomics and downstream-targeted metabolomic validations confirmed enhanced metabolic dysregulation in women with COPD, which was linked to oxidative stress. Together, this suggests metabolic activation and detoxification in the airway epithelium, as well as dysregulation in the delicate balance between oxidative stress and the anti-oxidative defence system, influence homeostasis and tissue damage in early stage COPD.

This work provides an optimized methodology that enables the investigation of the human immunopeptidome to reveal disease specific immunogenic peptides. Furthermore, proteomic and metabolomics approaches strive to add novel insights into biomarker discovery and disease mechanisms of inflammatory pulmonary diseases. The presented findings have the potential to provide clinical biomarkers and pharmaceutical targets to be used for early diagnosis and individualised treatment.

Popular Science Summary

Between Two Lungs presents a doctoral thesis studying two inflammatory lung diseases, sarcoidosis and COPD. Both diseases are chronic, however they show highly distinct clinical characteristics and therefore were studied separately. What brings the investigation in this thesis together are methodologies to identify and characterize proteins and metabolites and their alterations due to disease. The causes of both diseases are not completely understood, but genetic predisposition, gender and environmental factors present main risk factors in both cases. Symptoms for patients can be severe and life threatening in both instances. Therefore, markers that can be used for early diagnosis, as well as improved therapeutic strategies are in high demand to improve disease outcome and the life quality of patients.

Disease symptoms in sarcoidosis and COPD are distinct and their prevalence throughout the world also differs. Sarcoidosis is a rather rare disease, with the highest incidences in Scandinavian countries with approximately 1,200 new cases in Sweden every year. It is a disease which presents with acute inflammation and granuloma formation in the affected organs, most commonly the lungs. Breathing complications constitute a major issue for these patients and occur due to tissue scaring, which decreases the elasticity of the lung.

Certain features of sarcoidosis resemble autoimmune diseases, such as rheumatoid arthritis and multiple sclerosis. Autoimmunity is triggered by molecules (known as antigens) which cause the immune system to react against its own tissue, and it has been suggested that sarcoidosis is induced by exposure peptides that resembles the body's own proteins (auto antigens). In this thesis, a sensitive method was developed to identify such antigenic molecules in the lungs of individuals, both healthy controls as well as patients. This provides a platform for individual antigen profiling and comparison of peptide repertoires in heath and disease, and may reveal disease specific antigens and potential disease-specific triggers. This methodology can also be applied to other disease with unknown antigenic triggers and therefore presents a valuable tool in the hunt for specific antigens in complex, multifactorial diseases.

Further, within the scope of biomarker discovery for the diagnosis of sarcoidosis, we studied the antibody repertoire in blood and lung fluid. A marker that could be measured by a simple blood test that reflects the inflammatory status in the lungs, would be of great value in clinical practice. Emphasis was put on attached sugar moieties on antibodies (glycosylation), a specific modification that is known to change with inflammation. A ratio between two different glycosylation stages was identified as a potential marker for successfully distinguishing sarcoidosis patients from healthy controls, while simultaneously reflecting the degree of inflammation in the lung.

COPD, the second disease studied in this thesis, is characterized by lung inflammation and the narrowing of the airways, with obstructions resulting in difficulties in complete exhalation and therefore small amounts of air remaining in the lungs. COPD is the most common lung disease in the world and smoking presents the main risk factor for disease development.

Focus of this part of the thesis was the discovery of disease-specific biomarkers for early diagnosis and the identification of biological mechanisms that could be relevant for COPD development. In general, women are more susceptible to developing the disease and also present with different clinical features compared to men. Therefore, we specifically focused on gender differences while comparing healthy controls with COPD patients.

In order to study molecular changes attributed to disease, we characterized proteins in the airway epithelium, which presents the top cell layer in the airways and the first line of defence against any inhaled particles such as tobacco smoke. Additionally, we analysed metabolites in blood, which reflect all chemical processes that occur throughout the body. Distinct alterations in protein and metabolite profiles were observed in both female and male patients compared to healthy smokers. Discovered proteins and metabolites could therefore have potential as diagnostic marker for COPD.

We were also able to link the observed changes to processes related to oxidative stress in women. Oxidative stress essentially indicates an imbalance between the production of reactive toxic compounds (free radicals) and the ability of the body to counteract or detoxify their harmful effects. Disturbance of this balance may result in tissue damage. This dysregulation could play a key role in the destruction of lung tissue in COPD. Our data indicates that men and women react differently to tobacco smoke, and are not equal in their ability to tolerate inhaled toxicants. This might also explain why males and females show different symptoms upon clinical investigation.

This thesis presents different methods to study sarcoidosis and COPD, striving to identify disease-specific markers and reveal mechanisms that are of importance in the course of disease. We were able to provide tools developed to study specific antigens in inflammatory diseases, identify potential biomarkers and increase our knowledge of disease mechanisms in sarcoidosis and COPD. This might eventually contribute to improved diagnostic procedures, the development of new drugs, and eventually an improved quality of life for patients.

SCIENTIFIC PAPERS

This doctoral thesis is based on the following original articles, referred to in the text as project I-IV:

I. **Heyder T**, Kohler M, Tarasova NK, Haag S, Rutishauser D, Rivera NV, Sandin C, Mia S, Malmstrom V, Wheelock AM, Wahlstrom J, Holmdahl R, Eklund A, Zubarev RA, Grunewald J, Ytterberg AJ. Approach for Identifying Human Leukocyte Antigen (HLA)-DR Bound Peptides from Scarce Clinical Samples. *Mol Cell Proteomics* 2016: 15(9): 3017-3029.

II. **Heyder T*,** Wiklundh E*, Eklund A, Grunewald J, Zubarev RA, Lundstrom SL. Altered Fc-galactosylation status in IgG4 is a potential serum biomarker for chronic sarcoidosis. *Submitted manuscript.*

III. **Heyder T**, Yang M, Kohler M, Merikallio H, Forsslund H, Li CX, Karimi R, Eklund A, Grunewald J, Kaarteenaho R, Sihlbom C, Skold CM, Wheelock AM. Gender-enhanced alteration in airway epithelial proteome in COPD related to xenobiotic metabolism. *Submitted manuscript.*

IV. Naz S, Kolmert J, Yang M, Reinke SN, Kamleh MA, Snowden S, **Heyder T**, Levanen B, Erle DJ, Skold CM, Wheelock AM, Wheelock CE. Metabolomics analysis identifies sex-associated metabotypes of oxidative stress and the autotaxin-lysoPA axis in COPD. *Eur Respir J 2017.*

*These authors contributed equally to this work.

ADDITIONAL SCIENTIFIC PAPERS

Yang M, Kohler M, **Heyder T**; Forsslund H; Garberg H, Karimi R, Grunewald J, Berven F, Skold CM; Wheelock AM. Shotgun-based proteomics investigations of the chronic effects of smoking in the human bronchoalveolar lavage cell proteome. *Submitted manuscript.*

Yang M, Kohler M, **Heyder T**; Forsslund H; Garberg H, Karimi R, Grunewald J, Berven F, Nyren S, Skold CM; Wheelock AM. Proteomic profiling of lung immune cells reveals dysregulation of phagocytotic pathways in female-dominated molecular COPD phenotype. *Submitted manuscript.*

CONTENTS

ABBREVIATIONS

2D-DIGE	two-dimensional difference gel electrophoresis
APC	antigen-presenting cells
AUC	area under the curve
BAL	bronchoalveolar lavage
BALF	bronchoalveolar lavage fluid
BCR	B cell receptor
BHL	bilateral hilar lymphadenopathy
COPD	chronic obstructive pulmonary disease
CV-ANOVA	cross validated ANOVA
DC	dendritic cells
DL_{CO}	carbon monoxide diffusion capacity
ESI	electrospray ionization
ER	endoplasmic reticulum
EBV	Epstein-Barr Virus
FA2	truncated (agalactosylated) glycoform
FA2G2	digalactosylated glycoform
FDR	false discovery rate
FEV_1	forced expiratory volume in 1 s
FVC	forced vital capacity
GOLD	The Global Initiative for Chronic Obstructive Lung Disease
HILIC	hydrophilic interaction liquid chromatography
HLA	human leukocyte allele antigens
IgG	immunoglobulin
IL	interleukin
IP	immunoprecipitation
LC	liquid chromatography
LS	Löfgren's syndrome
m/z	mass-to-charge ratio
MHC	major histocompatibility complex
MS	mass spectrometry

MS/MS	tandem mass spectrometry
NK	natural killer
nLS	non-Löfgren's syndrome
OPLS-DA	orthogonal projections to latent structures discrimination analysis
p(corr)	correlation coefficient
PCA	principal component analysis
PRR	pattern-recognition receptor
PTM	post-translational modification
Q^2	goodness of fit
R^2	predictive power
ROC	receiver operating characteristic
ROS	reactive oxygen species
SD	standard deviation
SDS-PAGE	sodium dodecyl sulfate polyacryl gel electrophoresis
SUS	shared and unique structures
TAP	transporter associated with antigen processing
TCR	T cell receptor
TMT	tandem mass tag
VIP	variable influence on projection
WHO	World Health Organization

INTRODUCTION

Chronic inflammatory lung diseases represent a major health problem throughout the world and disease pathobiology is yet not fully understood. Within the frame of this doctoral thesis, two pulmonary diseases with unknown etiology were studied: sarcoidosis and chronic obstructive pulmonary disease (COPD). Particular focus was put on investigating molecular biomarkers and the underlying disease mechanisms via the application of established proteomic and metabolomics approaches, as well as the development of an optimized methodology for the identification of immunopeptides.

Sarcoidosis is a disease that is most prevalent in Northern European countries, with approximately 1,200 new cases diagnosed in Sweden every year [1]. It is characterized by granuloma formation in the affected organs that mainly manifests in the lungs. The majority of patients experience a resolving disease but some progress to chronic sarcoidosis, which has a higher risk of developing fibrosis and eventually respiratory failure. Mortality rates are between 1 and 5%, and mainly result from severe fibrosis and/or myocardial or central nervous system failure [2]. The well-established hypothesis that sarcoidosis develops as a result of genetically predisposed individuals that are exposed to specific antigens or inorganic particulates, which promote inflammation has potential to reveal the etiology of this disease [3].

COPD is a leading cause of morbidity and mortality worldwide [4], and the World Health Organization (WHO) estimates it will become the 3rd leading cause of death by the year 2020 [5]. It presents as a heterogeneous disease comprising of several distinct clinical phenotypes, which are characterized by an airflow obstruction which is not fully reversible [6]. Exposure to tobacco smoke represents a major risk factor for developing the disease. However, only 15-44% of life-long smokers develop COPD, and smoking alone cannot fully explain the disease origin [7, 8]. In addition, up to 25% of COPD patients have never smoked. Under-diagnosis and a failure to detect the disease early are major concerns in the clinical management of patients.

In order to grasp the complexity of disease etiology in chronic inflammatory lung diseases, molecular screening approaches and targeted methods were used to comprehensively characterize different molecular levels and compartments. Depending on the scientific question being addressed, specific molecules that are suspected to have an impact on diseases, such as immunopeptides and post-translational modifications on immunoglobulins (IgGs) in sarcoidosis were investigated. In COPD, where tobacco smoke is a major risk factor but further pathophysiological mechanisms are unknown, wide molecular screening and

integration of multiple "omics" platforms is the approaches of choice. The combination of unique and well characterized clinical material, state of the art analytical techniques, and novel bioinformatic and biostatistical tools represent the cornerstone of this work which strives to reveal the disease mechanisms in pulmonary inflammation. The results may eventually be formatted for the diagnostic of patients suffering from sarcoidosis or the initial stages of COPD, enabling clinicians to intervene at an earlier stage of the disease, thereby greatly reducing both the cost to society and more importantly improving the quality of life of patients.

THE IMMUNE SYSTEM

The Nobel Prize in Physiology or Medicine 1908: Ilya Ilyich Mechnikov & Paul Ehrlich
Recognition of their work on immunity

The immune system is a complex and tightly regulated network involving several organs, cells, tissues and signalling processes that serve to protect an organism against invasion by constantly evolving pathogens. Innate and adaptive immune mechanisms act in conjunction to recognize and eliminate foreign intruders that are harmful to the host. This summary presents a simplified tutorial of a classic immune response in humans, highlighting the major components, which are illustrated in Figure 1.

Innate immunity

The Nobel Prize in Physiology or Medicine 2011: Bruce A. Beutler & Jules A. Hoffmann
Discoveries concerning the activation of innate immunity

Innate immunity is a highly conserved mechanism that exists in all multicellular organisms, and it provides one of the first lines of defence against invading microbes [9, 10]. Defence mechanisms are composed of chemical, physical and microbiological barriers, as well as a number of specific immune cells, which initiate an acute inflammatory response. Cells of the innate immune system are present in all tissues and include granulocytes, mast cells, natural killer (NK) cells, dendritic cells (DCs), monocytes and macrophages [11] and they can immediate detect infiltration of a foreign pathogen. Tissue-resident macrophages and DCs, followed by circulating phagocytes (neutrophils and monocytes) recognize pathogenic surface molecules, so-called pathogen-associated molecular patterns (PAMPs) with their pattern-recognition receptors (PRRs) [12]. The best studied PRRs are the Toll-like receptors (TLRs) [13, 14]. Additionally, PRRs can also recognise similar structures known as damage-associated molecular patterns (DAMPs), which are released by stressed or necrotic cells and detected by phagocytes [15]. The interaction of these ligands with PRRs activates the transcription factor NF-κB, which further initiates the release of certain pro-inflammatory cytokines especially interleukin (IL)-1β, IL-6, IL-8 and tumour necrosis factor (TNF)-α to further activate and recruit cells to the site of inflammation [16]. Macrophages and neutrophils are able to directly ingest and eliminate pathogens and thereby prevent the release of harmful substances into the surrounding tissue. Eosinophils, basophils and mast cells, on the contrary, release antimicrobial substances and immune activating molecules that directly inactivate the pathogen.
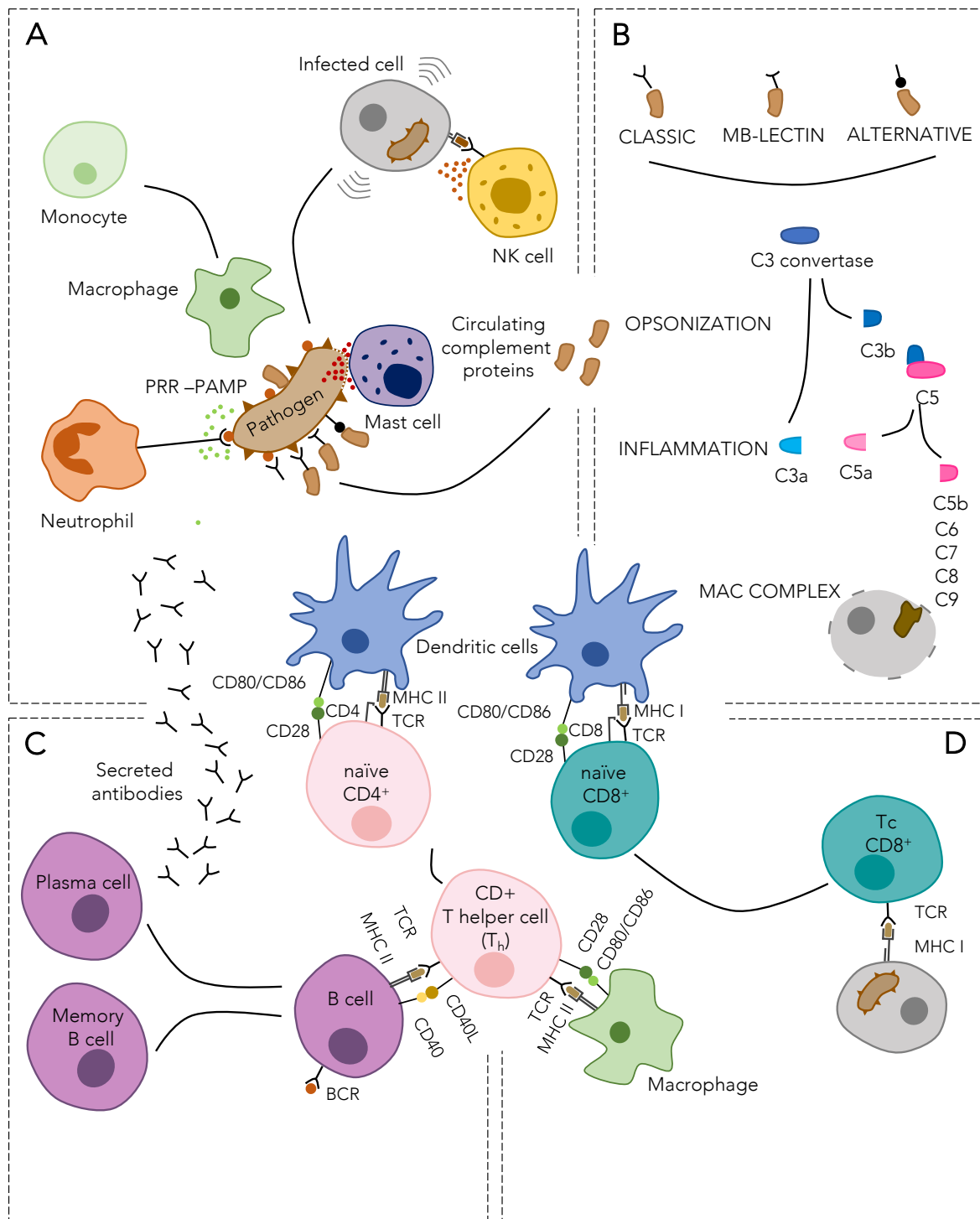
Figure 1| Illustration of the major components of the tightly regulated immune response. (A) First line of defence in innate immunity. (B) Schematic showing the major reactions in the complement cascade which are closely regulating innate immunity. (C) Humoral response of adaptive immunity involving B cells. (D) Cell mediated adaptive immunity activating cytotoxic T cells.

An additional part of the innate immune response is the onset of the complement cascade, which includes three pathways: namely the classical pathway, the lectin pathway and the alternative pathway [10]. These pathways induce the release of further pro-inflammatory mediators through the complement proteins C3a and C5a. Furthermore, complement protein C3b coats pathogens for enhanced recognition and uptake by macrophages and neutrophils (opsonisation). Alternatively, complement proteins C5b-C9 trigger the assembly of a membrane attack complex (MAC), which induces osmotic lysis of the pathogen [17]. The activation of the complement system also acts on mast cells inducing histamine release, which causes vascular dilatation to aid the recruitment of circulating immune cells to the site of infection [11]. Some pathogens invade human cells directly after entering the body. Natural killer (NK) cells are able to identify and target virus-infected cells and they release perforins that increase cellular membrane permeability thereby orchestrating the disruption of the infected target [18]. Eventually, the innate immune system recruits antigen-presenting cells (APCs), which together with NK cells, establish a link with the more advanced adaptive immune response.

Adaptive immunity

The Nobel Prize in Physiology or Medicine 2011: Ralph M. Steinman
Discovery of the dendritic cell and its role in adaptive immunity

Professional antigen presenting cells (APCs) such as macrophages, B cells and particularly DCs, have the unique ability to actively uptake antigen, process it and then present it on their surface, which is the key mechanism found in innate immunity that bridges the gap with the more specific adaptive immune response. Adaptive immunity is the second line of defence against invading pathogens and is more sophisticated than innate immunity. It can be subdivided into T cell-mediated (cellular) and antibody (humoral)/ B cell-mediated immunity.

Cellular response

Naïve T cells interact by their T cell receptor (TCR) with the surface of APCs through the major histocompatibility complex (MHC), consisting of the MHC molecule loaded with an antigenic peptide (antigen) of either endogenous (MHC class I) or exogenous (MHC class II) origin. After antigen uptake by DCs, the cells move to the local draining lymph node where T cell activation of CD4+ and CD8+ T cells occurs [19]. The requirement for the interaction of the TCR and the T cell surface protein CD8 or CD4 with the antigenic peptide together with the MHC molecule, prevents activation through free circulating antigenic peptides [11]. This step specifically requires the binding of co-stimulatory receptors for the complete

activation of naïve T cells. APCs express the co-stimulatory proteins CD80/CD86, which are recognized by the activating CD28 co-receptor on the T cell [20]. These co-stimulatory molecules are promoted during inflammation to fully activate T cells. Activated T cells produce interleukin (IL)-2, which induces clonal expansion and T cell proliferation [17]. CD4+ T cells (T helper cells, Th) are capable of recognising peptides presented by MHC class II, whereas CD8+ T cells (T cytotoxic cells, Tc) recognise antigens on MHC class I molecules (intracellular peptides, typically virus-derived). However, this mode of detection is not always systematic as cross-presentation often occurs [21]. The concept of presentation and recognition will be further described below.

The CD8+ T cell-mediated response is specialized for intracellular infections and is set in motion by the recognition of a cell that has been invaded by a pathogen [9]. This interplay activates the phagocyte to eliminate the pathogen and promotes a response against infections [17]. Infected cells are eliminated by direct induction of cell death, which is particularly effective in virus infections to prevent viral dissemination. The release of perforin at the immunological synapse enables the further release of granzymes and granulysin into the infected cell, which destroys the cell through triggering apoptosis [22]. Another pathway, which causes apoptosis is the so-called death-inducing signalling complex (DISC), which is activated by Fas-Fas ligand interactions [9]. CD8+ activation also releases cytokines that can directly affect infected cells, *e.g.* by intervening with pathogen replication.

CD4+ T cells play a central role in the adaptive immune response. The release of specific cytokines by activated CD4+ T cells supports macrophages in eliminating pathogens, activates cytotoxic T cells, and boosts the antibody producing ability of B cells, which is part of the antibody (humoral) B cell-mediated response (described below). T cells can generally be divided in different T cell subsets depending on the functional cytokines they produce. Traditionally, they are classified as Th1/ Tc1 and Th2/ Tc2 T cells but additional Th subsets were revealed adding Th17, Th22, Th9 and T regulatory cells [23].

Antigen presentation

The Nobel Prize in Physiology or Medicine 1996: Peter C. Doherty and Rolf M. Zinkernagel
Discoveries concerning the specificity of the cell mediated immune defence

The Nobel Prize in Physiology or Medicine 1980: Baruj Benacerraf, Jean Dausset and George D. Snell
Discoveries concerning genetically determined structures on the cell surface that regulate immunological reactions.

Processing and presentation of pathogenic derivatives occurs through different pathways depending on the origin of the pathogenic substance. Cytosolic pathogenic proteins are transported to the proteasome and degraded into small peptides. These peptides are then loaded onto the endoplasmic reticulum (ER) membrane bound protein transporter

associated with antigen processing (TAP). TAP actively pumps the associated peptides into the ER, where MHC I molecules attach to TAP through a protein called tapsin, and are subsequently loaded with peptides that fit into the binding grove. Next, MHC I peptide complexes are transported through the Golgi apparatus in an exocytic vesicle to the cell membrane and are expressed on the cell surface [9, 24].

Extracellular pathogens are engulfed by professional APCs into endocytic vesicles. These vesicles fuse with the lysosomal compartment containing enzymes that initiate the degradation of the pathogen into peptides. At the same time, MHC II molecules are synthesized in the ER and consist of an alpha (23kDa) and beta chain (35 kDa), and a binding cleft is formed in the interface between the two [9]. This cleft is loaded with the CLIP peptide of the Invariant chain protein (Ii) during synthesis, to prevent any unspecific peptide loading and promote transport in the endosomal vesicle [24, 25]. The MHC II complex is transported through the Golgi apparatus where the Ii protein is cleaved and the CLIP peptide remains in the binding cleft [10]. Endosomes containing the MHC II complex fuse with the endocytic vesicle to load the binding cleft with specific pathogenic peptides. CLIP peptide release from the binding cleft is initiated by the binding of the HLA-DM protein [10, 24]. MHC II molecules loaded with foreign peptide are subsequently transported to the cell membrane and expressed on the cell surface, where peptides are presented to circulating CD4+ T cells.

Peptides presented on MHC class II complexes are longer around 17 around amino acids in size [26] compared to peptides presented on MHC class I that are commonly around 8-10 amino acids in length [25]. MHC II peptides were specifically studied in this work in the context of investigating immunogenic peptides specific to sarcoidosis (project I). MHC II complexes have certain binding pockets at positions P1, P4, P6, P7 and P9 where peptides fit, and P1 appears to be the most important one for peptide binding [27].

T cell receptors (TCRs) are membrane-bound heterodimers composed of an α and β chain, each containing one constant (C) one variable (V) segment. The antigen binding site is formed by the Vα and the Vβ segments. The TCR complex also contains the invariant signalling proteins CD3 γ and δ, as well as the ϵ and ζ chains [9, 17]. TCRs only recognise as few as one to three residues of the displayed peptide, as shown by x-ray crystallography [28], however the flanking residues of the peptide determine binding strength and specificity [29]. In order to detect a myriad of constantly evolving pathogens, an enormous variety of specific TCRs are required. During T cell development, germline-encoded gene segments (V, D and J segments) are randomly recombined to generate vast receptor diversity [30]. An illustration of both MHC II molecule, TCR receptor and the interactive complex is given in Figure 2.
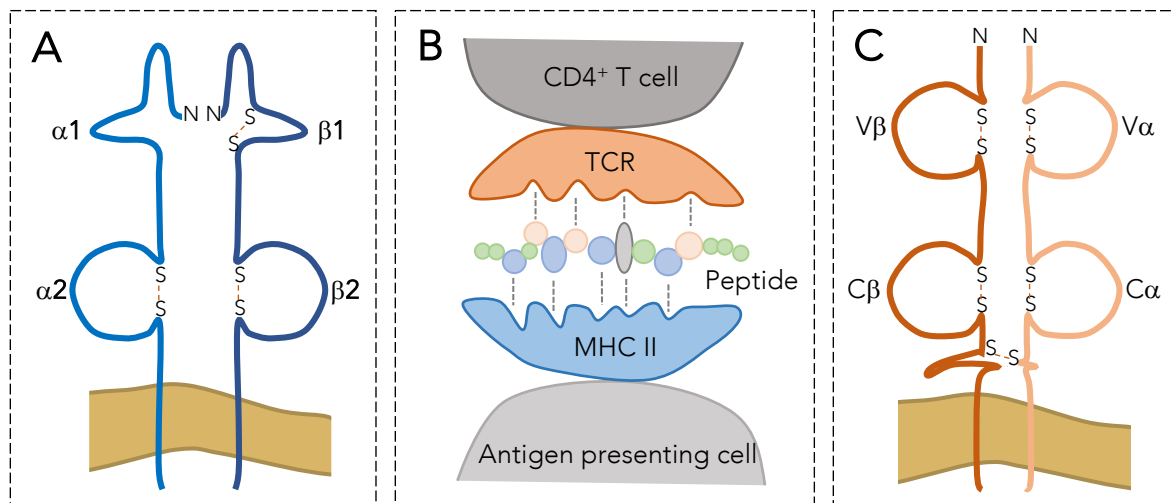
Figure 2| Illustration of the three components of antigen presentation and recognition. (A) Simplified structure of the MHC class II molecule composed of an α and β chain. (B) Model of an antigenic peptide presented by the MHC II molecule and recognized by the TCR. The amino acid sequence of the antigenic peptide presents certain side chains pointing towards the MHC II molecule and TCR, interacting with binding anchors in the complexes (indicated by the dotted lines) (C) Schematic structure of the TCR, constructed of an α and β chain, each with a constant and a variable region. The antigen binding site is composed of the Vα and Vβ segment.

MHC in humans

The major histocompatibility complexes are called human leukocyte antigens (HLA) in humans. HLA and non-HLA genes are located on chromosome 6 and make up the largest polymorphic region in the human genome [31]. HLA class I molecules are encoded by the gene loci A, B and C (HLA-A, -B and -C), whereas class II molecules are encoded by DR, DQ and DP (HLA-DR, -DQ and -DP) [32]. Most HLA genes exist in hundreds of allelic variants, one is inherited from each parent (*e.g.* HLA-DRB1*03,04). Therefore, a person can be homozygous or heterozygous for any given HLA gene. HLA typing is annotated by the inclusion of gene locus, allele group, and up to eight digits for the specific alleles. The first two digits specify a group of alleles, and the final digits the specific allelic variants inherited (*e.g.*, if a person has the HLA configuration HLA-DRB1*03,04, allele 03 is derived from one parent and 04 from the other. A schematic overview of HLA alleles and nomenclature of HLA types is shown in Figure 3.
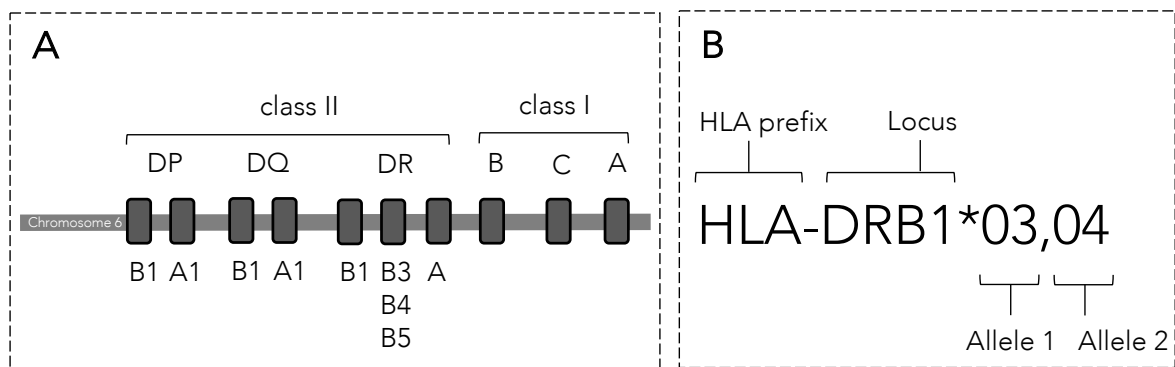
Figure 3| HLA nomenclature. (A) Schematic overview of HLA class I and class II alleles on chromosome 6 in the human genome. (B) Nomenclature of HLA types with an example of both allele variants in a heterozygous individual.

Humoral response

Parallel to cell-mediated immunity, humoral immunity occurs when mature B cells detect pathogenic particles through their B cell receptor (BCR). As with APCs, B cells get activated by phagocytosing pathogens and presenting peptides on their MHC class II molecules on the cell surface. The peptide, together with the MHC II molecule, is recognised by a $CD4^+$ cell via the TCR together with a secondary signal initiated via the binding of CD40L on the T cell with the co-stimulatory molecule CD40 on the B cell [10]. The $CD4^+$ cell releases specific cytokines inducing B cell proliferation and maturation into antibody producing plasma cells and eventually memory cells. The plasma cell subsequently releases antibodies upon antigen detection through their B cell receptor. Memory B cells are long-lasting cells and upon encountering a pathogen which has previously infected the host, they immediately proliferate and secrete high-affinity antibodies [9].

Antibodies

The Nobel Prize in Physiology or Medicine 1972: Gerald M. Edelman and Rodney R. Porter
Discoveries concerning the chemical structure of antibodies

The Nobel Prize in Physiology or Medicine 1987: Susumu Tonegawa
Discovery of the genetic principle for generation of antibody diversity

The Nobel Prize in Physiology or Medicine 1984: Niels K. Jerne, Georges J.F. Köhler & César Milstein
Theories concerning the specificity in development and control of the immune system and the discovery of the principle for production of monoclonal antibodies

Immunoglobulins are commonly referred to as antibodies and depending on their isotype (IgA, IgD, IgE, IgG, IgM), they have specific functions in the immune response. IgG is the most prominent isotype in circulation (75%) and can be subdivided into IgG_1 (60-70%), IgG_2 (20-30%), IgG_3 (5-8%) and IgG_4 (1-3%) [33]. The IgG molecule is composed of four polypeptide chains (2 heavy and 2 light chains) that form a Y-shaped molecule. Each of these

chains have a constant (C) region and a variable (V) region. The heavy chain is composed of three constant and one variable segment, whereas the light chain contains one of each segment. The variable segment together with the first following constant region constitutes the antigen binding (Fab) region. The second and third constant segments make the Fc region [9]. A simplified illustration of an IgG monomer is given in Figure 4A.

Fc-glycans

Most immune cells have a surface molecule (Fc receptor) with the ability to bind the Fc region of antibodies and communicate with the adaptive and innate parts of the immune response. Modifications in the Fc region of immunoglobulins by glycosylation have a strong impact on their interaction with Fc receptors, activation of the complement cascade and other effector functions [34, 35]. The IgG Fc-glycosylation profile is altered in autoimmune and inflammatory disorders [36-38]. The more complex the Fc-glycan is (particularly if the glycan is substituted with galactose and sialic acid), the less likely the IgG is to have a pro-inflammatory effect [35, 39-41]. The amount of galactosylated and/or sialylated Fc-glycans (or the ratio between agalactosylated and galactosylated glycans) is a potential marker for inflammation and/or activation of the immune system [42]. Glycosylation structure and nomenclature by Royle *et al.* [43] is summarized in Figure 4B and C.
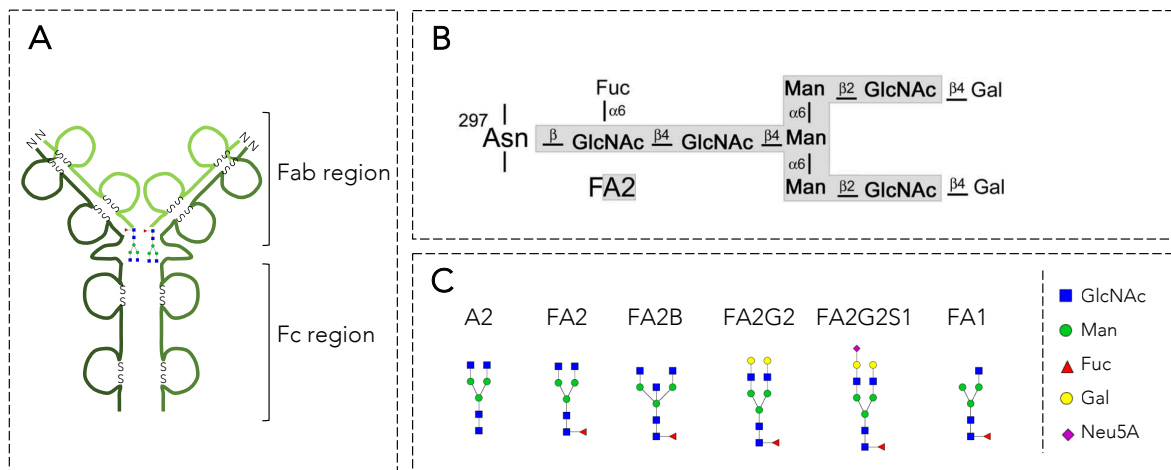

Figure 4| Schematic structure of the IgG protein and Fc-glycans. (A) Simplified structure of the IgG protein composed of a Fab- and Fc- region and indicated glycosylation sites. (B) Chemical structure of FA2G2, indicating the core heptasaccharide moiety (A2), in grey, linked to Asparagine (Asn) 297. F at the start of the abbreviation (FA2) indicates a fucose (Fuc) linked to the inner N-acetyl-glucosamine (GlcNAc). G2 indicates 2 galactoses (Gal) linked to the antenna N-acetyl-glucosamines. (C) Examples of the structural diversity and glycan nomenclature described by Royle *et al.* [43]. *Adapted from [44].*

Autoimmunity

The Nobel Prize in Physiology or Medicine 1960: Sir Frank Macfarlane Burnet & Peter Brian Medawar
Discovery of acquired immunological tolerance

The German immunologist Paul Ehrlich first described the concept of autoimmunity as "horror autotoxicus" [45]. The immune system is educated to react with foreign substances (non-self) and to protect own (self) cells. When this inbuilt tolerance is lost, benign autoimmunity progresses to pathogenic autoimmunity. The prevalence of such disorders is approximately five percent in Europe and North America, and they constitute the 10th leading cause of death worldwide [46, 47]. Furthermore, women are more frequently affected than men [48]. Molecular mimicry is one mechanism by which infectious agents (or other exogenous substances) may trigger an immune response against autoantigens [49]. Genetic studies have shown that most autoimmune diseases have a strong HLA association and there is a further contribution of certain HLA haplotypes to disease susceptibility, symptomatic presentation and prognosis [50, 51]. Beyond genetic susceptibility, as shown in twins [52, 53] and genome-wide association studies (GWAS) [54, 55], other risk factors contributing to development of autoimmunity include environmental factors [56], gender [57], smoking [58] and infections [59].

THE RESPIRATORY SYSTEM

Nobel Prize in Physiology of Medicine 1938: Corneille Jean François Heymans
Discovery of the role played by the sinus and aortic mechanisms in the regulation of respiration.

The airways

The main function of the respiratory tract is to enable gas exchange in the alveoli by oxygenising (O_2) the circulating blood, as well as removing carbon dioxide (CO_2). The respiratory system can be divided into the proximal conducting part and the distal respiratory tract [60]. The human lung has an asymmetric anatomy with two lobes on the left and three lobes on the right side. The main bronchus divides into branches (primary, secondary and tertiary bifurcations) in each lobe. The conductive zone of the lungs is composed of the nasal cavities, oropharynx, trachea, bronchi and bronchioles. The segmental bronchi branches into several generations, the intersegmental bronchi, the bronchiole and the terminal bronchiole, Figure 5A) [61]. The terminal bronchiole subdivides numerous times and evolves into the respiratory zone, which consists in total of seven generations (bifurcations), where the first three are respiratory bronchioles, followed by three generations of alveolar ducts, which then terminate in the alveolar sacs; ~23rd generation

(Figure 5B) The alveoli are surrounded by a filigree net of pulmonary capillaries providing the interface for gas exchange.

Airway epithelium

The airways have a continuous pseudostratified columnar epithelial lining from the nasal passages down to the alveolar sacs. The epithelium is coated with epithelial lining fluid, which forms a thin fluid layer that covers the mucosa of the alveoli, the small airways, and the large airways. The airway epithelium and epithelial lining fluid constitutes the interface with the environment and represents the first line of defence against harmful airborne agents such as microbial pathogens, toxicants and air pollutants [62, 63]. The airway epithelium is anchored to the basement membrane, which is the structural scaffold between the mesenchymal compartment and the surface epithelium. The main constituents of the basement membrane are Type IV collagens, intertwined with proteoglycans, laminins, integrins, entacins and other constituents of the lamina lucida. The extracellular spaces of the basement membrane are thought to be unique signal input devices that are likely to fine tune cellular function [64, 65]. Throughout the conductive and respiratory zones, the cellular composition and morphology of the epithelium changes. The large airways of the proximal region including nasal cavity, trachea, bronchi and bronchioles are mainly composed of basal cells forming the basal membrane, ciliated epithelial cells and mucous secreting goblet cells (Figure 5C-D). Ciliated cells are the most abundant cell type in the lung epithelial lining, and are covered with cilia, which together with mucus producing goblet cells are essential for mucociliary clearance [63]. The cuboidal epithelium in the bronchioles consists primarily of ciliated epithelial cells and secretory club cells, which are fewer in number and also smaller in height.

In response to exposure, secretory cells in the mesenchyme are activated and enable the clearance of inhaled substances. In addition, epithelial cells can trigger innate and adaptive immunity by releasing antimicrobial substances as well as pro-inflammatory mediators [66, 67].
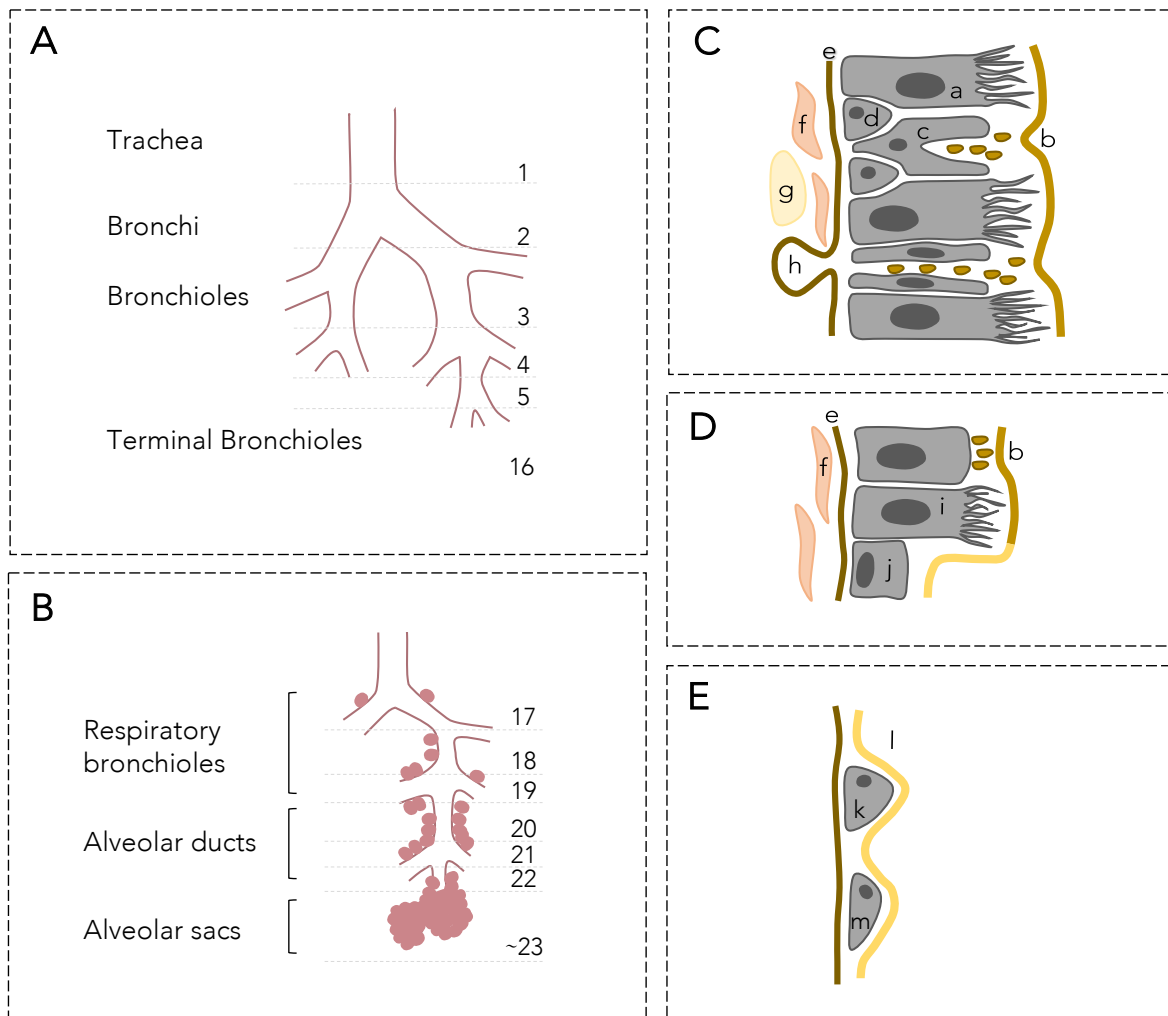
Figure 5| Anatomical structure of the airways and cellular composition of the airway epithelial lining along the respiratory tract. (A) Conducting zone of the lungs composed of trachea, bronchi, bronchioles until the terminal bronchioles. (B) Respiratory zone composed of total 7 bifurcations containing respiratory bronchioles, alveolar ducts and alveolar sacs. (C) The bronchial epithelium is present in the trachea and bronchi and contains ciliated epithelium (a), mucus (b) secreting goblet cells (c), basal cells (d) forming the basal membrane (e), smooth muscle cells (f). Additional cartilages (g) and submucosal glands (h) are present in this section of the airways. (D) The bronchiolar epithelium contains smaller ciliated epithelial cells (i) and club cells (j). (E) The alveolar epithelium contains type I pneumocyte (k) and surfactant (l) producing type II pneumocyte (m).

Lung toxicology

The airways are constantly exposed to a variety of harmful foreign substances such as air pollutants, chemical compounds, nanoparticles, vapours and aerosols. Any chemicals which an organism is exposed to that are extrinsic to the normal metabolism of that organism are defined as xenobiotics [68]. Cigarette smoke, which is a main exposure and risk factor in COPD has been estimated to consist of >7,000 chemical compounds from many different chemical classes with a variety of toxic mechanisms [69]. Xenobiotic metabolism plays a key role in both the activation and detoxification of several noxious stimuli, including polycyclic aromatic hydrocarbons and their nitro-derivatives, to maintain homeostasis in the
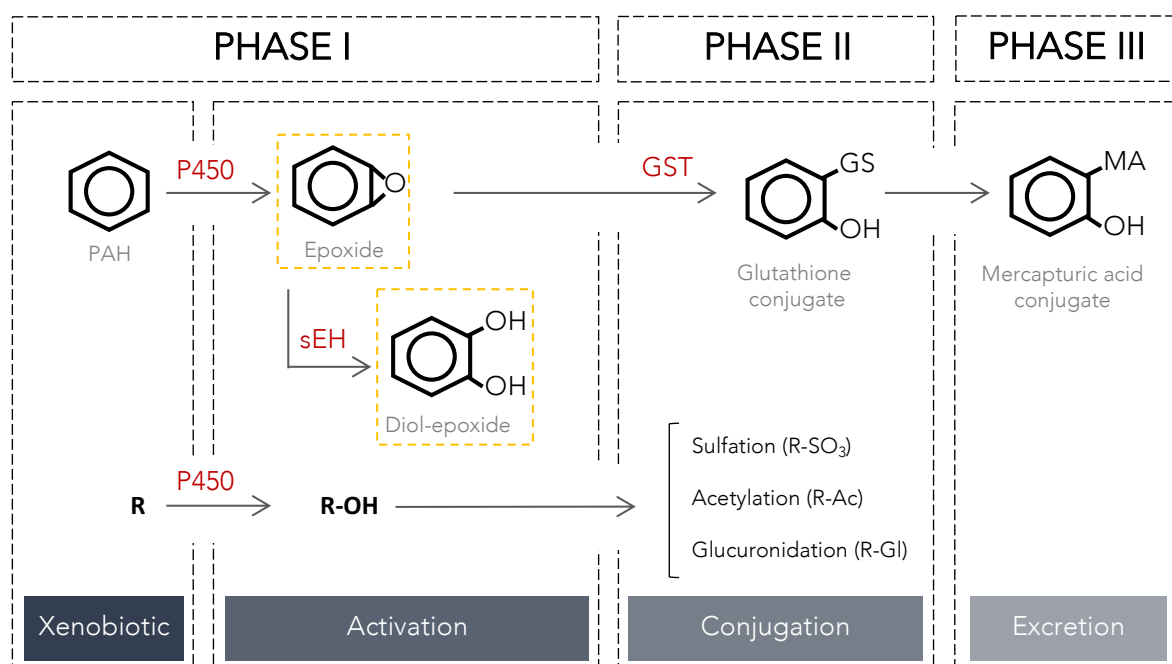
Figure 6| Xenobiotic metabolism reaction. Biotransformation can be divided into three phases: Phase I reactions including the activation of xenobiotics, generally through oxidation. Phase II reaction involves conjugation of activated compounds, and Phase III involves the potentially further conjugation and excretion of hydrophilic components. Epoxide hydrolase (©), Glutathione transferase (GST), Glutathione (GS), Mercapturic acid (MA).

Lung. Xenobiotic metabolism can be divided into phase I (activation) and phase II (conjugation) reactions. Cytochrome P450 oxygenases are the main family of enzymes responsible for catalysing the activating oxidation of polycyclic aromatic hydrocarbons into epoxide intermediates, which can then be converted by epoxide hydrolase to highly carcinogenic diol-epoxides [70]. After the initial activation, the conjugation (phase II) occurs through attachment of glutathione, sulfate, glycine, or glucuronic acid. After conjugation, xenobiotic compounds may be further transformed by the processing of glutathione conjugates to mercapturic acid conjugates to increase the solubility of the compound [71], and thereby increase the extractability of the toxicant. Simplified reactions are summarised in Figure 6.

INFLAMMATORY LUNG DISEASES

The presented thesis includes studies on two inflammatory pulmonary diseases: sarcoidosis and COPD. Both disorders exhibit an inflammatory, chiefly Th1-mediated immune response in the lungs with unknown causes and disease mechanisms Clinical manifestations of both diseases represent a large spectrum of various disease entities with a range or yet to be defined clinical phenotypes.

Sarcoidosis

Sarcoidosis is an inflammatory disease of unknown etiology mainly affecting the lungs, but it may also involve several other organs such as the eyes, heart, skin and central nervous system. The histological hallmark is the formation of non-caseating epithelioid cell granulomas in affected organs. Other causes of granuloma formation, *e.g.* mycobacterial or fungal infections, have to be excluded before confirming a sarcoidosis diagnosis. In the lungs, an accumulation of CD4$^+$ T cells is often observed. The incidence and presentation of sarcoidosis varies between different ethnic groups and the highest prevalence has been observed in northern European countries [2]. In Sweden approximately 1,200 new cases are diagnosed every year and more than half of all patients will have a disease course lasting more than two years, which by definition is regarded as chronic [1]. While the main clinical symptoms are a non-productive cough and dyspnoea, fatigue and arthralgia are common features with major influence on the patient's quality of everyday life [72, 73]. Approximately one third of Swedish patients with sarcoidosis will have acute disease onset with high-grade fever, erythema nodosum and/or ankle arthritis or periarticular inflammation in combination with bilateral hilar lymphadenopathy (BHL) with or without concomitant parenchymal infiltrates on chest radiographs [74, 75]. This subgroup is defined as having Löfgren´s syndrome (LS), first described by the Swedish physician Sven Löfgren [76]. These patients usually have a favourable prognosis, with spontaneous resolution of disease symptoms, characteristically without permanent tissue scaring. LS is strongly linked to HLA-DRB1*03, which is in turn associated with a good prognosis. Ninety-five percent of LS patients with HLA-DRB1*03 will resolve within two years after onset, as opposed to DRB1*03 negative LS patients who only have a 50% chance of resolving within this time [75]. HLA-DR3$^+$ sarcoidosis patients show the unique feature of an accumulations of T cells in the lungs with the specific T cell receptor segments Vα2.3 , [77] and Vβ22 [78]. This finding suggests the presentation of a disease specific antigen that might be responsible for the disease onset in sarcoidosis [79]. The subgroup with an insidious onset and higher risk of developing a chronic disease course is referred to as non-Löfgren (nLS) sarcoidosis. NLS patients who develop chronic disease have a higher prevalence of the HLA-haplotypes DRB1*14 or DRB1*15 [80, 81]. Other HLA types have been associated with specific clinical phenotypes, *e.g.* DRB1*04 which can be linked to ocular sarcoidosis [82]. Recent genetic studies have revealed distinct differences between LS and non-LS, with LS patients being genetically more homogenous, and showing gene associations not present in non-LS, and vice versa [83].

The diagnosis of sarcoidosis is based on the combined detection of non-caseating epithelioid cell granulomas in biopsy specimens from affected organs, typical from chest X-rays, and clinical symptoms compatible with sarcoidosis. The diagnosis may be supported by findings with spirometry and analysis of the diffusion capacity of carbon monoxide (DL$_{CO}$), as

well as with an accumulation of CD4$^+$ T-lymphocytes in the BAL fluid resulting in a CD4$^+$/CD8$^+$ T-cell ratio >3.5 [84]. Also, some biomarkers in blood, such as elevated serum angiotensin converting enzyme activity and increased serum calcium levels, may strengthen the diagnosis. Based on the findings seen in chest X-rays, the stage of sarcoidosis can be defined as follows: Stage 0 – normal findings; stage I – BHL, stage II – BHL and pulmonary infiltrates; stage III – pulmonary infiltrates without BHL; stage IV – signs of fibrosis and volume reduction [85].

Chronic obstructive pulmonary disease

Chronic Obstructive Pulmonary Disease (COPD) is a heterogeneous disease characterized by persistent respiratory symptoms and airflow limitation that is due to airway and/or alveolar abnormalities. Causes traditionally considered to induce disease include the significant exposure to noxious particles or gases [86] and emerging evidence also implicates developmental factors and early life events. COPD presents as a multifaceted condition comprising chronic bronchitis, bronchiolitis and emphysema of varying degrees. The disease is a leading cause of mortality and morbidity worldwide [4], and the WHO predicts it will become the 3rd leading cause of death by the year 2020 [5]. Major symptoms include dyspnea, coughing and exercise limitations, whereas clinical manifestations vary widely between individuals and genders. Smoking is a major risk factor for disease development, however not all smokers develop COPD, and smoking alone cannot therefore fully explain disease etiology. Other risk factors are genetics, environmental exposures, asthma, premature birth and persistent respiratory infections in early childhood [87, 88]. Traditionally, COPD has been considered a disease primarily affecting male smokers, however the mortality rate in females has increased [89, 90]. Additionally, differences in molecular and clinical phenotypes have been reported. Men appear to experience a higher degree of emphysema relative to smoking history [91-94], whereas female smokers exhibit a significantly faster annual decline in lung function after the menopause [89, 95], even after correction for smoking habits. Diagnosis of COPD is based on the presence of airflow obstruction, which is assessed by spirometry. The essential spirometry values required for a COPD diagnosis are the ratio between forced expiratory volume in 1 second (FEV$_1$) and forced vital capacity (FVC), *i.e.* FEV$_1$/FVC, as described above. According to the Global Initiative for Chronic Obstructive Lung Disease (GOLD), COPD is staged on severity based on spirometry values for FEV$_1$ (% of predicted). In recent years, the symptoms and frequency of exacerbations (*e.g.*, acute worsening of respiratory symptoms) have also been considered in the GOLD criteria signified as disease severity A-D [96, 97]. The most significant changes in the latest update have been the separation of spirometry assessment and evaluation of symptoms and exacerbations. A simplified overview of the staging is given in Figure 7.
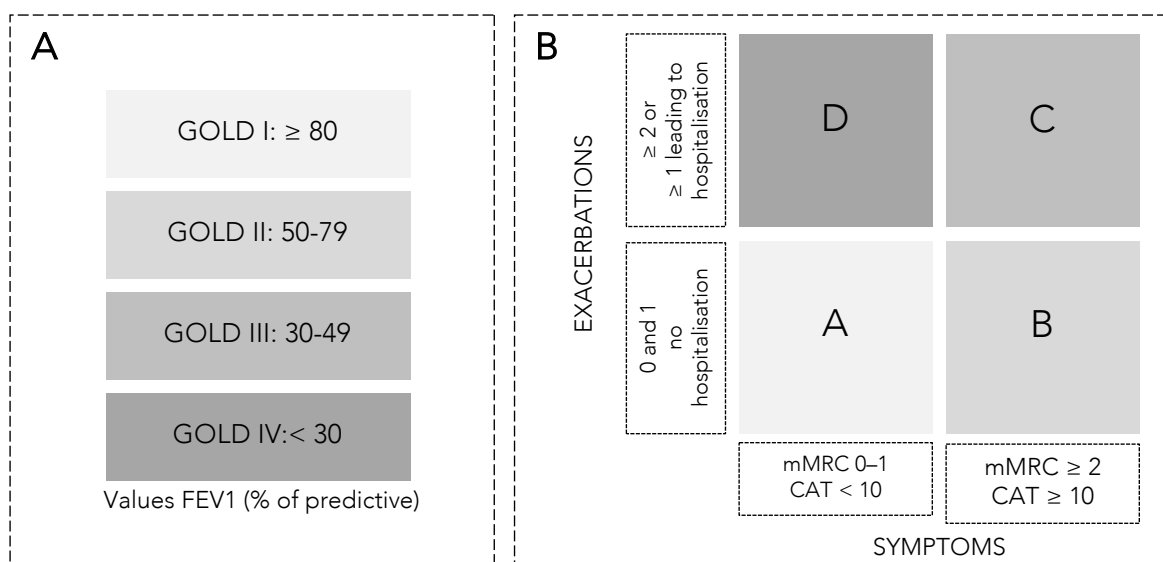
Figure 7| GOLD standards for staging of COPD. (A) GOLD standards defining severity of COPD based on airflow limitation assessed by FEV1 (% of predicted). (B) Refined assessment tool ([96, 97]) using four categorical groups based on number of exacerbations and the modified British Medical Research Council (mMRC) questionnaire and COPD Assessment Test (CAT).

CLINICAL PROTEOMICS & METABOLOMICS

The investigations presented in this thesis work are based on proteomic and metabolomics approaches, both of which were performed using mass spectrometry (MS) with preceding sample preparation and chromatographic approaches. This makes MS a key methodology throughout the projects. The following sections will give an overview of "omics" and their application in medical science, utilising workflows as well as an introduction to MS.

The omics concept

Omics represents a scientific field comprising of analytical methods (suffix –omics) that are used for the comprehensive characterization of specific molecule classes (suffix –ome) in a biological system (*e.g.*, cells, fluid etc.). The main omics approaches include studies of genes, RNA, proteins and metabolites, which together present the omics cascade (Figure 8). Different techniques are applied depending on which part of the omics cascade is the focus of the scientific question. Genomics represents the investigation of the DNA content of a system, which contains information on "*what can happen*". Transcriptomics in contrast gives in insight into "*what might happen*" due to the encoded RNA. Protein synthesis subsequently occurs and proteomics describes a field that focuses on the functional role of proteins and "*what currently happens*". One step further down in the omics cascade is metabolomics, which involves the analysis of small metabolites that indicate "*what has*

happened" in biological processes. Finally, all these molecular processes come together in a biological phenotype (phenome), which can be either unaltered (healthy) or modified (disease) [98, 99].
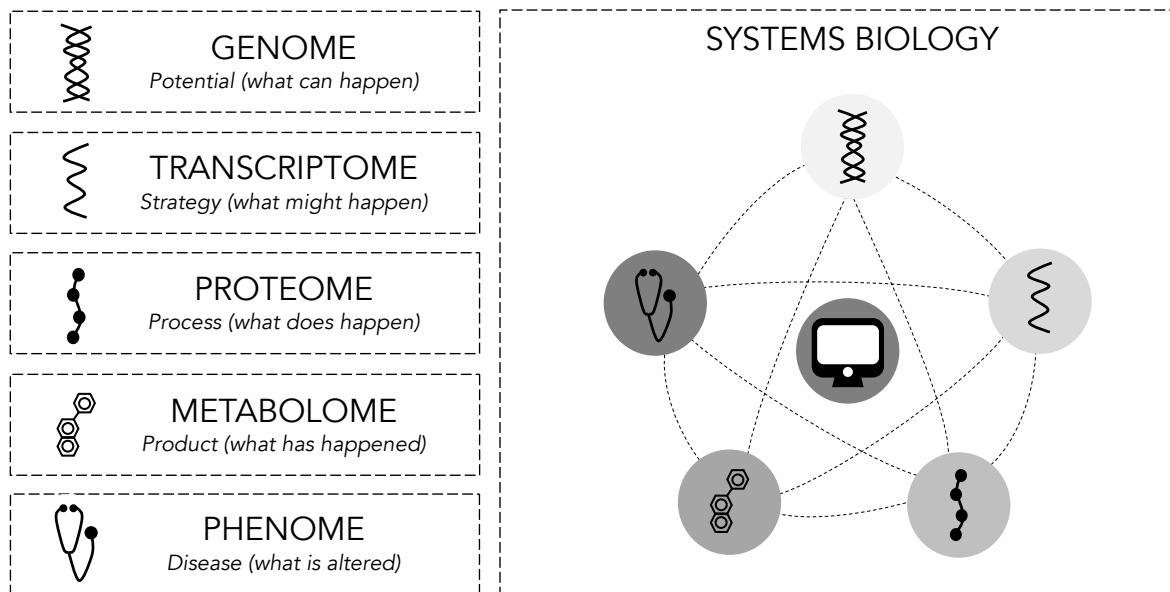

Figure 8| The integration of the omics cascade using computational and statistical tools represent the core in systems biology. The cascade includes, but is not limited to, the genome (DNA), transcriptome (RNA), proteome (proteins), metabolome (metabolites). The status of the molecular composition results in a phenotype (phenome), which after alteration may turn from healthy to a diseased status.

Clinical omics

Clinical omics comprises of many analytical methodologies used in medical research to answer a variety of questions using patient material. These approaches have the ability to evolve basic science towards clinical applications with the aim to reveal markers of disease and disease mechanisms which can be used to identify new targets for pharmaceutical drug development. Continuous improvements in mass spectrometry instrumentation, computational data analysis and bioinformatics as well as complete sequence databases for many species, drives omics research towards large-scale analysis and systems biology.

Systems biology often refers to analyses where the whole system (organism) is considered. The Seattle Institute for Systems Biology, one of the pioneering institutes in the field, describes it as follows: "Systems biology is based on the understanding that the whole is greater than the sum of the parts". It presents an interdisciplinary field between biology, the life sciences and computational methodologies that strive to investigate the complexity and interplay of biological processes. In omics research, the integration of multiple omics platforms has become more evident, and it offers a powerful tool with which to investigate biological questions at multiple molecular levels together with clinical parameters to gain deeper and broader insights into diseases pathologies [99].

Proteomics

The strategies employed in a proteomic experiment to identify and characterise proteins can be fundamentally different. Top-down approaches describe the analysis of intact proteins or large protein fragments. Alternatively, proteins can also be digested into small peptides prior to analysis, which is referred to as bottom-up proteomics. Proteomic experiments can include multiple, individual technologies depending on the objective. Generally, proteomic approaches can be classified as gel based and shotgun (gel free) proteomics (Figure 9). Proteomic workflows commonly start with protein enrichment/purification of the sample material (cells, tissues, fluids, etc.) using extraction, fractionation or separation, or depletion approaches.

Gel based proteomics include gel electrophoresis for protein separation in either a one- or two-dimensional design. The separation of proteins with traditional sodium dodecyl sulfate polyacryl gel electrophoresis (SDS-PAGE) is based electrophoretic mobility of charged proteins through a porous gel matrix, under a simultaneously applied electrical field, which separates by the total molecular weight. By introducing isoelectric focusing (IEF) prior to gel electrophoresis, proteins are separated in two dimensions: the isoelectric point and the molecular weight [100]. During isoelectric focusing, proteins migrate along an electrical field to their isoelectric point (netto charge 0) [101]. A breakthrough in gel electrophoresis was achieved with the introduction of commercially available immobilized pH gradients for enhanced reproducibility [102] and by mass and charge matched fluorescent minimal labelling [103]. Together, this resulted in a method called two-dimensional difference gel electrophoresis (2D-DIGE), which has the capability to normalize for technical variation by separating reference samples in every experimental run. Two-dimensional gel electrophoresis (2D-PAGE) multifluorescence analysis profits from two major advantages: high sensitivity and broad dynamic range [104]. Image analysis is performed on fluorescent scanners with matching lasers for excitation and filters for emission spectra of the corresponding fluorescent label. Software based image analysis can reveal altered protein spots of interest, which can subsequently be excised, digested with trypsin and identified using modern mass spectrometry approaches like LC-MS/MS, which will be described in more detail below. Novel methodologies like cumulative time-resolved emission two-dimensional electrophoresis (CuTEDGE) technology, facilitate in-depth proteomics characterization of very scarce samples using gel based proteomics [105]. However, limitations in the sensitivity and dynamic range of 2D-PAGE drive proteomics towards more gel free methods.
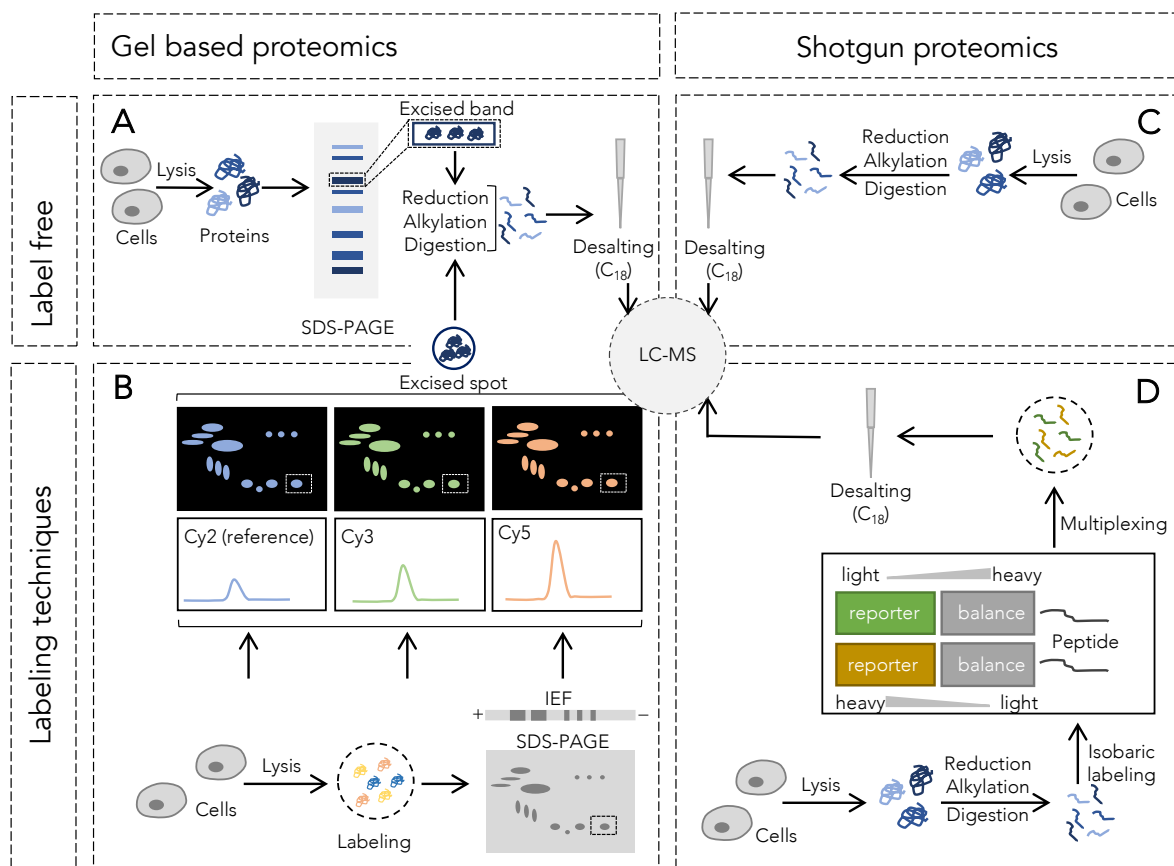
Figure 9| General concepts of different proteomic approaches prior to mass spectrometry (LC-MS). (A) Gel based label free proteomics represented in a one-dimensional gel electrophoresis (SDS-PAGE). (B) Semi-quantitative proteomics using 2D-DIGE. Quantification is enabled by measuring the emission intensity of fluorescent tags attached to the protein (Cy2, 3 and 5). (C) Shot gun proteomics using a label free approach with protein digestion and straight measurement with mass spectrometry. (D) Shot gun proteomics using isobaric labelling illustrated for two samples.

Proteins can also be directly applied to mass spectrometry methods without separation by gel electrophoresis, which is referred to as shotgun proteomics. Proteins obtained from a gel or in solution are most commonly digested with trypsin, which generates tryptic peptides. Shotgun proteomics almost invariably involves the enzymatic degradation of proteins to peptides by trypsin [106]. The cleavage of the C-terminal at arginine or lysine residues leads to the generation of peptides in the preferred mass range for effective fragmentation by tandem mass spectrometry (MS/MS) and places highly basic residues at the C termini of the peptides. This generally leads to an informative high mass y-ion series and it makes tandem mass spectra easier to interpret [107]. Concentration steps and further purification might be added before separation by liquid chromatography and analysis with mass spectrometry.

One research area within proteomics focuses on the characterisation of post-translational modifications (PTMs). PTMs include the covalent linkage of a functional group to a protein by enzymatic or chemical reactions or as a result of proteolytic activity. These modifications can occur during or after protein biosynthesis and are known to have essential functions in in

protein and signalling networks in cells. PTMs have been linked to numerous developmental disorders and human diseases, highlighting the importance of PTMs in maintaining normal cellular homeostasis [108]. In project II, we specifically studied glycosylation, which represents the linkage of sugar moieties, and has a large variety of possible glycosidic linkages (described in more detail in the section "Fc-glycans")

Metabolomics

Metabolomics describes the analysis of metabolites (*e.g.*, enzymes, amino acids, organic acids, fatty acid etc.) in cells, tissue or body fluids. Sample preparation commonly includes enrichment and cleaning steps prior to analysis. Approaches can be divided into targeted and non-targeted methodologies [109]. Non-targeted metabolomics using mass spectrometry are commonly coupled online to gas chromatography (GC) or liquid chromatography (LC) systems for separation of complex mixtures prior to analysis. In targeted metabolomics, the investigated metabolites are known prior to analysis and are used to answer a specific scientific hypothesis. This means targeted metabolomics have a greater analytical depth compared to non-targeted. Metabolomic approaches used in the frame of this thesis (project IV) were performed with LC-MS based non-targeted methods for universal detection of metabolic analytes in clinical samples, with subsequent validation of selected candidates by targeted approaches.

Chromatography

The Nobel Prize in Chemistry 1952: Archer John Porter Martin and Richard Laurence Millington Synge
Invention of partition chromatography

Mass spectrometry used in metabolomics and proteomics is commonly coupled to online chromatography systems for separation of complex mixtures prior to analysis. Chromatography is a technique used to separate complex mixtures based on different strong interactions with analytes in a mobile phase and in a stationary phase. Chromatography approaches can be classified into gas chromatography (GC) and liquid chromatography (LC). GC is based on the different boiling points of molecules which are dependent on volatility and polarity. High-performance LC (HPLC) is a widely used analytical platform for non-targeted metabolomics and shotgun proteomics [109, 110]. Specifically, micro capillary and nano-LC (nl/min flowrates) are common components of MS instrumentation. Additionally, ultra-performance liquid chromatography (UPLC) can be used instead of conventional HPLC.

Liquid chromatographic techniques can be distinguished based on the binding of the compounds that are to be separated. Reversed phase (RP) is the most common application

in liquid chromatography used in proteomics and metabolomics, and it is based on the hydrophobic interactions between analyte and the stationary phase (C_{18}) [111]. During loading, the mobile phase contains low organic components to enable binding, while a gradual increase of the organic concentration induces the elution of compounds from the stationary phase. Metabolites that are analysed using RP separation are mainly medium sized polar and non-polar metabolites, such as organic acids including amino acids, fatty acids and phospholipids.

A complementary chromatography method used specifically in metabolomics for separation is hydrophilic interaction liquid chromatography (HILIC), which is commonly used for the identification of hydrophilic metabolites such as organic acids with a greater polarity, amino acids, nucleotides and intermediates of central energy metabolism [112, 113]. The separation is based on hydrophilic interactions. Highly organic mobile phases enable the binding of polar metabolites with a higher affinity for the stationary phase than hydrophobic compounds. The concentration of organic solvent is slowly reduced, thereby eluting polar metabolites from the column.

Mass spectrometry

Nobel Prize in Physiology in Chemistry 2002: John B. Fenn and Koichi Tanaka
Development of soft desorption ionisation methods for mass spectrometric analyses of biological macromolecules

Mass spectrometry is indispensable in proteomics and metabolomics research due to its sensitivity and analytical speed. The introduction of soft ionization techniques further enabled the analysis of thermal instable molecules such as proteins and metabolites, which makes MS a preferred analytical tool in this area. The general components of a mass spectrometer include ion source, mass analyser (m/z) and detector [101]. Electron spray ionization (ESI) is the most common form used for the ionization of analytes directly from a solution, and can therefore be coupled online to a liquid chromatographic system (LC) of choice (RP, HILIC etc.). ESI enables the analysis of complex sample mixtures since chromatographic separation of molecules with different retention times increases the number of components that can be selected and fragmented. After ionization, analytes enter the mass analyser where they are separated and recorded by the detector, which is often an electron multiplier. The mass analyser is essential for the instrumental performance regarding mass resolution, mass accuracy, mass range, sensitivity and the capability for tandem MS [114]. The most common mass analysers for proteomics are: quadrupole (Q) and ion trap (quadrupole ion trap [QIT], linear ion trap [LIT] or linear trap quadrupole [LTQ]) [115].

The fast evolution of mass spectrometry based omics increasingly demands instrumentation with better performance characteristics including resolution and mass accuracy, in order to analyse samples and mixtures with a greater complexity. Makarov *et al.* [116] introduced the Orbitrap mass analyser, which has become standard in high resolution mass spectrometry instrumentation, with higher sensitivity and faster data acquisition than alternative machines [117]. Combined with an electrospray ionization and LC system it presents the most common instrumental set up in shotgun proteomics (Figure 10A).

MS data is normally acquired using data-dependent acquisition, where the most intense ions (precursors) in every MS cycle that have not been previously selected are used for fragmentation (Figure 10B). Fragmentation occurs by collision energy, which breaks the peptide into amino acid fragments from which the sequence can be deduced, either by matching the fragmentation pattern to those found in existing databases with regards to observed and expected fragmentation ions or manually by *de novo* sequencing. Tandem mass spectrometry presents a powerful tool for identification and furthermore, it can provide information about PTMs. Different fragmentation techniques are available and commonly used examples include collision-induced dissociation (CID), electron-capture dissociation (ECD), electron-transfer dissociation (ETD), and Higher-energy C-trap dissociation (HCD).
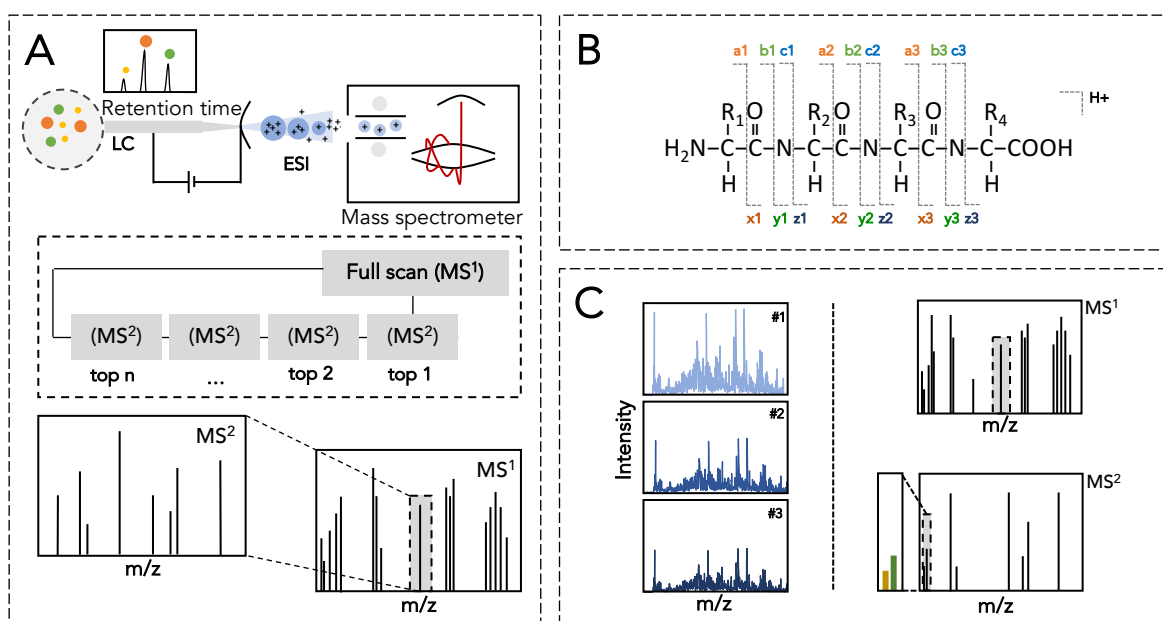

Figure 10| Steps in a classical mass spectrometry work flow. (A) Peptides are separated by liquid chromatography (LC) and ionized with electro spray ionization (ESI) before they enter the mass spectrometer. Peptide ions are measured at high resolution in data dependent mode. After each full scan (MS^1), the most intense peptide ions (precursor ions) that have not yet been selected are fragmented for identification by MS^2 spectra. (B) Nomenclature of fragment ions of a tryptic peptide (Roepstorff and Fohlman *et al.* [118]). (C) Quantification of peptides and proteins based on signal intensity between runs (label free, left) or of report ion intensities of stable isotope labels in MS^2 (right).

Quantification

In contrast to purely qualitative identification, proteomics and metabolomics can also be quantitative and used to investigate differences in the abundances of biomolecules when comparing study groups, which is essential in clinical omics where alterations can occur due to disease. Relative or absolute quantification has increasingly become the focus of proteomics experiments and has largely replaced the initial goal to only generate accurate and complete lists of identified proteins [119]. Relative quantification can be label free or based on labelling techniques that employ either metabolic, conjugated, colorimetric, or enzymatic labels [117].

Quantification using labels also enables the inclusion of a reference pool to normalise against technical variation by multiplexed experimental designs. DIGE (gel-based) and tandem TMT (MS-based) are two labelling techniques used in parts of this thesis. The underlying chemistry is described in more detail in the methodology section. In 2D-DIGE, CyDye fluorochromes are covalently linked to lysine amino acids found in the protein sequence and quantification is performed through the detection of the relevant emission wavelength. For compatibility with MS-based identification of selected protein spots, minimal labelling strategies are commonly used where limited amounts of dyes are appointed [120]. TMT mass spectrometry uses isobaric labels for quantification followed by shot gun proteomics. Currently, a maximum of 10-plex experimental designs are feasible. TMT labels are indistinguishable in the MS full scan due to mass matched isobaric labels. Fragmentation yields reporter ions of different masses enabling quantification based on different intensities of every reporter ion (Figure 10C) [121].

MULTIVARIATE DATA ANALYSIS

Proteomic and metabolomics experiments acquire large data volumes, and traditional univariate statistics (*e.g.,* t-tests) where each variable is analysed in isolation do not account for the co-variance of the multi-dimensional data sets. A multidimensional dataset can be viewed as a matrix containing *N* rows (study subjects as observations) and *K* columns (measured analytes as variables) and is referred to as an *X*-matrix. Multivariate statistical modelling reduces the dimensionality of the complex data structures by projections of the dominating trends in the multidimensional dataset onto a few representative latent variables called principal components (Figure 11A) [122]. Major modelling techniques can be divided into unsupervised and supervised methods. The need for multivariate data analysis becomes increasingly apparent in omics based research, which generates large data sets. There is a need for in depth studies that may facilitate the elucidation of the etiology of complex trait

diseases. However, multivariate data analyses are not widely used in the medical community. Terminology and model statistics need scrutinizing in order to evaluate the quality of the resulting models. The following section will give a basic overview of the major components of multivariate modelling within the scope of this work.

Data processing: Mean-centering and scaling

In order to adjust the different dynamic concentration ranges of analytes and to avoid the modelling being driven by single variables with high individual variance, imported data is processed before the actual modelling. Two common methods are mean-centering and scaling. Mean-centering subtracts the average value across all individuals from every data point. Scaling to unit variance involves dividing every data point by the standard deviation (SD) across all subjects (Figure 11B).

Modelling: PCA & OPLS-DA

Observations and variables are displayed in separate plots, with the scores plot visualising the group separation of the subjects and the loadings plot visualising the relationship of the original variables (*e.g.*, proteins and metabolites) to the scores (*i.e.*, subjects). The starting point of any multivariate data workflow is principal component analysis (PCA). The most important readouts in PCA are the provision of an overview of the data quality (*e.g.*, batch effects) and the identification of outliers (Hoteling's Square T^2 presenting 95% confidence interval, Figure 11C) [123].

In order to extract meaningful biological information from complex clinical omics data sets, more advanced models are usually required. In contrast to PCA modelling, orthogonal projections to latent structures (OPLS) analysis is a supervised method [124]. In a supervised analysis, the group labels are defined, as opposed to the unsupervised PCA analysis, which is completely data driven. A *Y*-matrix is linked to the initial *X*-matrix, which includes information defining the groups that will be compared in the analysis, *e.g.* information about gender, health *vs.* disease, etc. (Figure 11D). This form of analysis is designed to extract intra- and inter-group variance by separating structured noise unrelated (orthogonal) to the predictive variance of interest [124]. Discriminant analysis OPLS (OPLS-DA) is a form of OPLS, which utilizes the classification of subject groups (*e.g.* group A *versus* B or male *versus* female). Additional OPLS-DA regression analysis can also be included in the modelling. In this case the initial data matrix is modelled against a quantitative data *Y*-matrix, for example through the inclusion of clinical information. This way of modelling is used to investigate the inner correlation of the data with another data block.
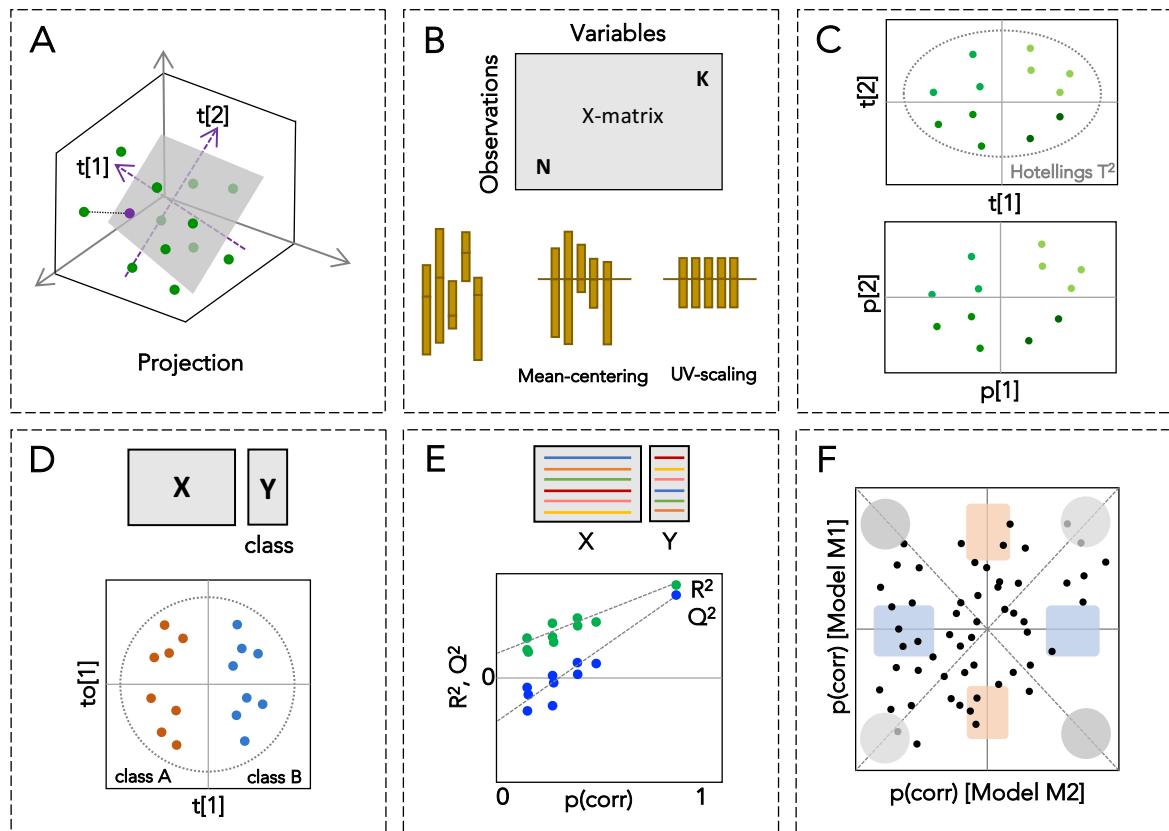
Figure 11| Fundamental steps in multivariate data analysis. (A) Principle of reduction of dimensionality by projecting a multi-dimensional space onto a two-dimensional space defined by two principle components (p[1] and p[2]). (B) Prior to analysis, the data containing X-matrix is processed by mean-centering and univariate scaling. (C) Principle component analysis (PCA) showing the scores plot (t[1] and t[2]) and corresponding loadings plot (p[1] and p[2]). (D) Orthogonal projections to latent structures discriminant analysis (OPLS-DA) model displaying one predictive (t[1]) and one orthogonal (to[1]) component for classification analyses (group A *versus* B) by including additional class information (Y-matrix) into the modelling. (E) Permutation to validate robustness of an OPLS model. (F) SUS analysis displaying the shared (grey) and unique features (orange and blue) of two models (Model M1and Model M2).

Variable selection

Variable selection is an essential step in multivariate modelling in order to extract biologically relevant information regarding biomarkers and dysregulated pathways from proteomic and metabolomic screenings. The variable selection step is based on the importance of a marker that drives the group separation, however, generally applicable cut offs are yet been established. Methods used for variable selection in this thesis include Variable Influence on Projection (VIP>1.0) [124] and the scaled loadings |p(corr)| equal to a *p*-value of <0.05 [125]. A combination of both is also commonly used and can be displayed in a so-called volcano plot.

Model statistics

Multivariate models are evaluated based on their model statistics, which represent the quality of the model and its value for data interpretation. Generally, two major values are reported in this context: the goodness of fit (R^2) and the goodness of prediction (Q^2). The goodness of fit reports the proportion of variance in the data set that can be described by the generated model. Additionally, the goodness of prediction value is created based on cross validation and reflects the predictive power of the model, which can reveal if the model is dependent on individual subjects. Cross validation is an iterative process where a proportion of the data is removed (*e.g.*, 1/7th of the subjects in 7-fold cross validation). A new model is created based on the remaining data (here, 6/7th of the subjects), and the performance of the model in correctly classifying the subjects included in the removed part of the data is predicted by the new model. The Q^2 value represents the average performance over a large number of iterations. Finally, the statistical significance of the group separation of supervised models is reported as cross validated ANOVA (CV-ANOVA) *p*-values [126]. Validation and evaluation of multivariate models based on model statistics is essential for the quality of data analysis since visual interpretation can be misleading. Overfitting is a risk in OPLS modelling and the determination of a correct number of predictive components is essential. Overfitting by adding more components to describe the data can be detected by a large discrepancy between R^2 and Q^2 values. An over fitted model may result in well discriminated score plots, but a model lacking predictive power.

Permutation

Permutation tests estimate the statistical significance of an OPLS model and constitute a further validation step for assessing model quality. This validation is based on comparing the goodness of fit (R^2) and the predictive power (Q^2) of the original model with several new models in which the order of the *Y*-observation (class information) is randomly permutated, while the *X*-matrix remains unchanged. A common way to present the permutation results is to plot matched R^2 and Q^2 values for the original, un-permutated model as a function of the number of correctly assigned subject labels for every permutated model. Common criteria for evaluating the permutation experiment require that all permutated R^2 and Q^2 values are below the un-permuted R^2 and Q^2 values, and moreover, that the regression line of Q^2 intersects below the origin (Figure 11E). Failure to do so indicates strong influences of a limited number of subjects on the original model, and presents an over fitted or false positive analysis.

Shared and unique structures

Shared and unique structures (SUS) analysis is a tool within multivariate modelling to further investigate similarities and differences in the variables driving different classification models. By plotting the scaled loadings (p[corr]) for two OPLS-DA models (*e.g.*, models following gender stratification) each classifying two groups (*e.g.* health and disease), a SUS plot is generated (Figure 11F). The information that can be extracted includes unique features for each model or shared characteristics. Variables that are unique for either of the models cluster at high p(corr) values along the axes of the plot. Shared features between the models can be found clustering along the positive diagonal, while opposing features are found along the negative diagonal.

PRESENT INVESTIGATIONS

OBJECTIVES

Overall objective of the thesis

The presented doctoral thesis includes work on two chronic inflammatory lung diseases: sarcoidosis and COPD. The overall objective of the work has been the discovery of disease specific biomarkers and antigenic peptides as well as the investigation of underlying molecular disease mechanisms and dysregulated biological pathways. Furthermore, the development of a sensitive methodology for the successful identification of immunopeptides was also part of the overall objective. By applying a specifically designed immunopeptidomic approach, as well as state of the art proteomic and metabolimic techniques, we aimed to contribute novel knowledge to disease pathophysiology in both diseases.

Specific research objectives

I. To optimize the approach to identify HLA-DR bound peptides to a sensitive methodology that enables the personalized profiling of immunopeptides from as little as 10×10^6 BAL cells, which corresponding to the minimum number of BAL cells obtained from healthy individuals

II. To reveal disease specific serum biomarkers for sarcoidosis that are able to target the inflammation in the lungs and have the potential to be formatted for use as non-invasive clinical test for sarcoidosis.

III. To investigate alterations in the airway epithelial proteome in smoking-induced COPD, with emphasis on gender differences to reveal disease specific biomarkers and underlying molecular pathways linked to COPD pathobiology.

IV. To investigate systemic metabolic shifts in serum in early stage COPD, with a specific focus on gender differences to identify diseases specific serum markers and molecular pathways involved in disease pathology.

METHODOLOGICAL CONSIDERATIONS

The specific methodology used in the four projects is described in detail in each original scientific article. The current section is intended to expand on these methods sections, providing more in depth information and motivations regarding the choice of methodology.

Clinical sampling

Ethical approvals

All studies are based on human material from clinically well characterized healthy controls and patients. The following ethical permits provide the legal foundation of the investigation presented in this doctoral thesis. The use of human BAL cells as well as BAL fluid in projects I and II for the analysis of HLA-DR peptides, as well as the investigation of the IgGome, was approved by the Stockholm Regional Ethical Board (no. 2005/1031-31, 2009/20-32, 2011/35-32). Projects III and IV present parts of a multi-omics systems biology approach investigating early stage COPD, by examining the airways epithelial proteome and serum metabolome. This work was approved by the Stockholm Regional Ethical Board (no. 2006/959-31/1, 2007/743-32, 2008/600-32, 2009/1353-32). All subjects provided their informed written consent.

The Karolinska COSMIC cohort

The proteomic and metabolomic investigation in project III and IV were performed on subjects from the Karolinska COSMIC (Chronic obstructive pulmonary disease and smoking from an omics perspective) cohort (www.clinicaltrials.gov/ct2/show/NCT02627872). The Karolinska COSMIC study is a three-group cross-sectional study including healthy never-smokers (Healthy), smokers with normal lung function (Smokers) and COPD patients (GOLD stage I-II/ A-B; FEV1, 50- 100% of predicted; and FEV1/FVC <0.7). A focus was put on early stage COPD patients regarding the aim to investigate molecular alterations related to disease onset, and early diagnostic markers. The age span of 45-65 years of age was chosen based on the apparent gender differences in disease onset, decline in lung function, and frequency of COPD in this age group [89, 95]. COPD patients were further divided into current-smokers (COPD) and ex-smokers (COPDxS; >2 years since smoke cessation). Smokers were matched for smoking history (>10 pack years, >10 cigarettes/day past 6 months) and subjects refrained from smoking >8 hrs prior bronchoscopy to control for acute smoking effects, as confirmed by exhaled carbon monoxide monitoring [127]. Further criteria for subjects included in the study was the absence of neither inhaled nor oral corticosteroids, no history of allergic asthma and no exacerbations for at least 3 months

before bronchoscopy. Information about menopause and hormonal replacement therapy was also recorded to investigate hormonal affect in the female subgroup. All comparisons were performed on either smoking or non-smoking groups to correct for smoking as a confounding factor. Extensive clinical characterisations were performed, including lung density measured by computed tomography, lung function (FEV$_1$, FVC) measured by spirometry, the percentage of goblet cells in the airway epithelium determined by AB-PAS$^+$ staining, cell differentials of lung immune cells, T cell phenotyping (CD4$^+$, CD8$^+$, CD69$^+$) identified by FACS analysis, smoking history (pack years), current smoking (cigarettes/ day) as well as general life style information and medication information collected through questionnaires.

Sampling by bronchoscopy

Bronchoscopy presents a central clinical procedure for the sampling of human material from the lungs used in this work. Bronchoscopy is a medical procedure that enables the visual inspection of the conducting airways. A flexible fiberoptic bronchoscope is introduced through the nose or mouth into the lower airways under local anaesthesia. In addition, it provides possibilities to collect various types of specimens from the proximal and distal parts of the lung. Thus, mucus can be retrieved as well as cells collected from the bronchi by brush biopsies. Tissue specimens may be taken with forceps from the airways and/or lung parenchyma. By performing bronchoalveolar lavage (BAL), cells and soluble components from the bronchial- and alveolar lumen can be sampled. In spite of its invasive nature, bronchoscopy is a very safe technique and can be performed on an out-patient basis also for research purposes on moderately sedated patients or healthy volunteers. Bronchoscopy and BAL were performed according protocol at our clinic as previously published [128, 129].

In this thesis, a number of specimens were collected during bronchoscopy to facilitate investigation of molecular alterations between healthy individuals and patients diagnosed with sarcoidosis or early stage COPD. BAL fluid (BALF) was collected after instillation of 5 aliquots of 50 ml phosphate buffered saline (PBS) into a segment of the right middle lobe. The fluid was immediately recollected by aspiration. BALF consists of a cellular and a soluble fraction. The dominating cells in healthy volunteers, both in number and percentage, are usually alveolar macrophages, and these are most pronounced in smokers. The BALF in healthy individuals also contains lymphocytes, a small percentage of neutrophils, and occasionally basophils and eosinophils [129]. In the soluble fraction, many different solutes reside such as proteins, lipids, and peptides and it is an interesting matrix to analyse when studying the effects of smoking and inflammatory lung diseases. Additionally, bronchial brushings were collected with serological brushed during each bronchoscopy from up to six distinct sites in the conducting airways, to sample the airway epithelium.

Lung measurements: Spirometry and DL$_{co}$

Spirometry is a method used to measure lung function and represents an easily accessible tool to detect airway restriction and obstruction [130]. By spirometry, lung volumes are determined. A schematic figure of recorded values during spirometry is given in Figure 12A. Patients with restrictive lung diseases including fibrotic diseases (*e.g.*, sarcoidosis) have a reduced total lung capacity due to scarring and therefore loss in flexibility and the ability to expand the lungs [131] (Figure 12B). Lung obstruction as in COPD is characterized by the inability to exhale air in a normal fashion, and COPD is currently defined as the inability to exhale 70% of the vital capacity in the first second of exhalation (FEV$_1$/ FVC< 0.7) [132, 133]. (Figure 12C).

Another tool to detect lung restriction and obstruction is diffusing capacity for carbon monoxide (DL$_{CO}$). This measurement determines how much oxygen is transmitted from the alveoli into the blood stream and based on the ratio of the partial pressure between inspired and expired carbon monoxide.
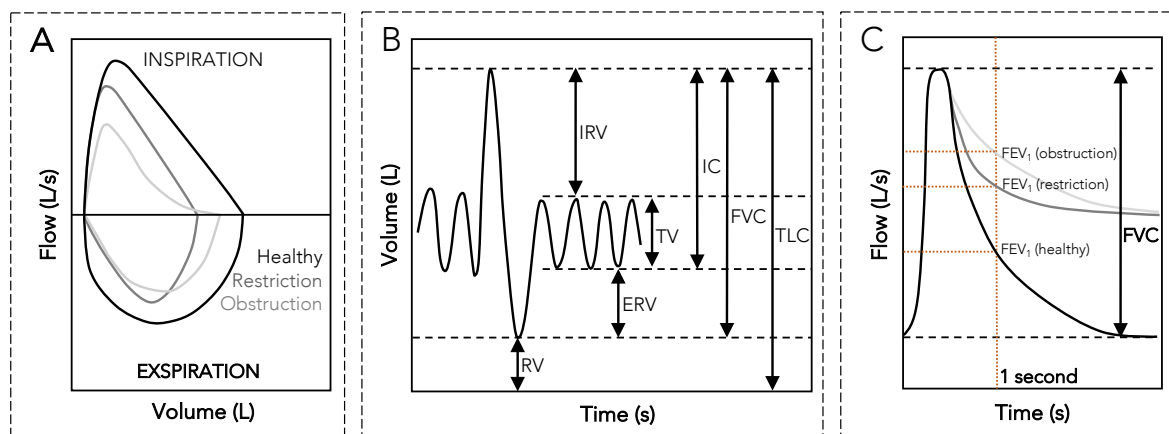

Figure 12| Spirometry measurements. (A) Demonstrating a flow-volume curve from a dynamic spirometry measurement from a healthy individual (black), a patient with lung restriction (dark grey) and a patient with lung obstruction (light grey). (B) Lung volumes recorded during static spirometry: total lung capacity (TLC), forced vital capacity (FVC), inspiratory capacity (IC), tidal volume (TV), expiratory reserve volume (ERV), inspiratory reserve volume (IRV) and residual volume (RV). (C) Forced expiratory volume in 1 second (FEV$_1$) in healthy, obstructed and restricted situations.

Meeting the analytical goal

The complexity and depth of the human proteome and metabolome is yet not fully understood. Dynamic ranges of the human proteome span over 10 orders of magnitude [134], whereas the human metabolome spans up to 11 orders of magnitude [135]. Modern mass spectrometry barely reaches 5 orders, which leaves an undiscovered "dark corner" of the proteome and metabolome consisting of low abundant molecules, which might present the most interesting targets for biomarker discovery [135, 136]. To intervene with such limitations and meet the analytical goals, specific sample preparation and protocol adjustments throughout the four projects were required. An overview of the investigated material, target molecules and analytical methodologies is presented in Figure 13.
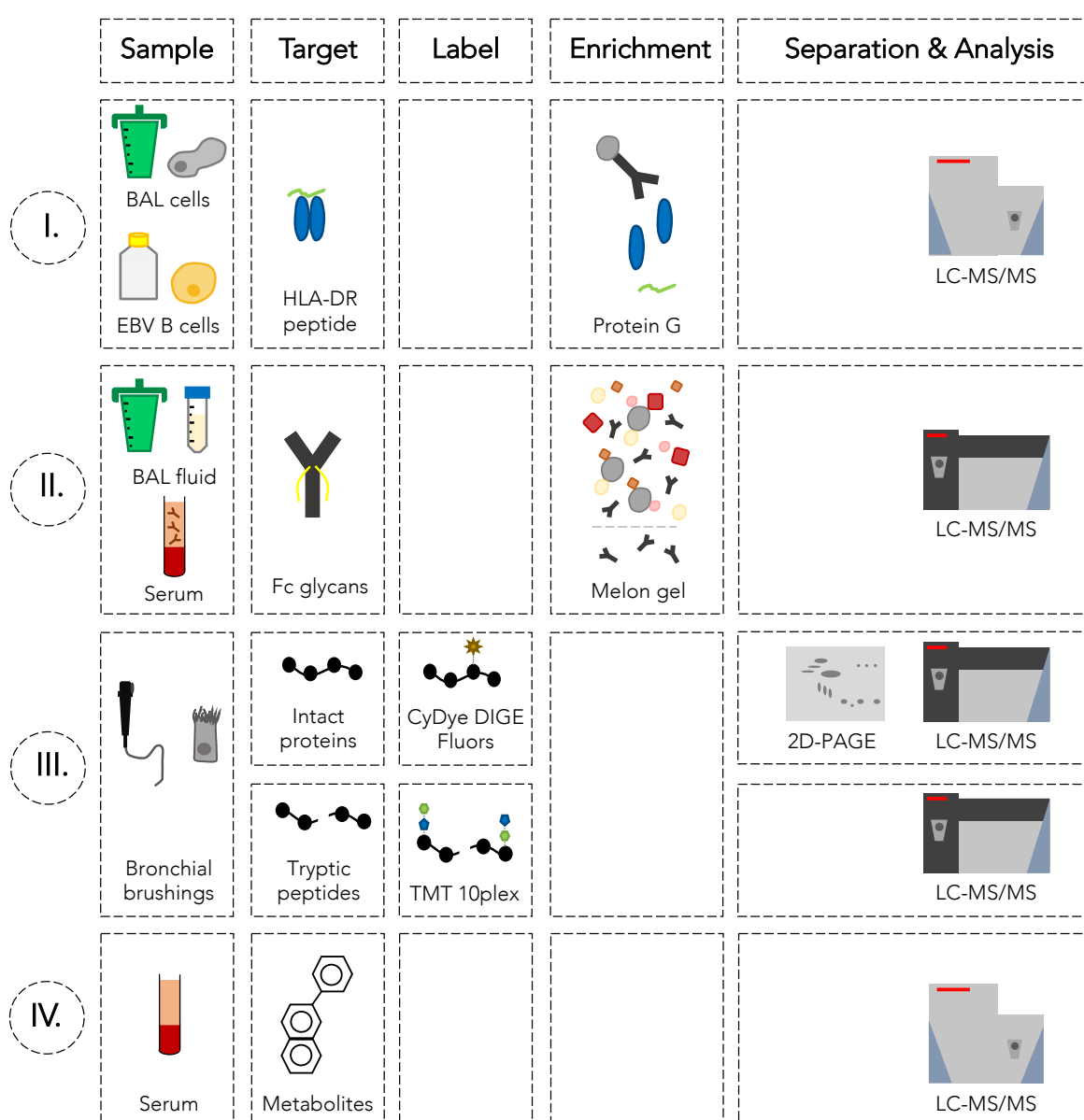

	Sample	Target	Label	Enrichment	Separation & Analysis
I.	BAL cells / EBV B cells	HLA-DR peptide		Protein G	LC-MS/MS
II.	BAL fluid / Serum	Fc glycans		Melon gel	LC-MS/MS
III.	Bronchial brushings	Intact proteins	CyDye DIGE Fluors		2D-PAGE / LC-MS/MS
		Tryptic peptides	TMT 10plex		LC-MS/MS
IV.	Serum	Metabolites			LC-MS/MS

Figure 13| Overview of the approaches used in projects I-IV. The projects differ regarding the sample material investigated, target molecules in the centre of the analysis, labelling methods (exclusively used in project III), enrichment steps (immunoprecipitation [IP] and Melon gel depletion) prior to analysis, separation and the mass spectrometry instrumentation.

Affinity approaches: Protein G and melon gel

Project I and II comprise of enrichment and depletion steps which remove highly abundant proteins and isolate target molecules respectively. The more a sample can be enriched with respect to a particular molecular preference, the more likely it is that the analytical methods succeed in identifying low abundance proteins. Project I included two consecutive steps to reduce the sample complexity: (1) enrichment of crude membranes by two solubilizing and ultracentrifugation steps and (2) immunoprecipitation to isolate HLA-DR complexes. Both of these steps were major adjustments required to access the low abundant immunopeptides. Furthermore, these steps reduce interfering contaminants, which preponderate when working with small sample sizes and scaled down protocols. This study presents a qualitative proteomic approach with the primary aim to identify immunopeptides in order to generate personalized immunopeptide maps in health and disease. Immunoprecipitation experiments that enrich for HLA-DR bound peptides were performed using the Thermo Scientific Protein G UltraLink Resins. HLA-DR complexes were incubated with anti HLA-DR (L243) [137] antibody and subsequently captured using protein G resin in a stationary phase. Peptides are commonly eluted from the complexes using acetic conditions. In project I, additional washing steps with 150 mM NaCl, 50 mM ammonium bicarbonate followed by 150 mM NaCl, 50mM ammonium bicarbonate, 5% acetonitrile removed nonspecific bound molecules. Elution yield were improved by using high organic buffers in addition to acetic conditions (50% acetonitrile and 5% formic acid) [138].

In contrast to the protein G affinity purification of HLA-DR complexes, the enrichment of total IgG from serum and BALF in project II was performed by depletion using the Thermo Scientific Melon Gel IgG Purification System. Human serum especially contains highly abundant proteins like albumin, which can mask any low abundant proteins which may be of interest in the analysis (*i.e.,* herein IgG$_4$ and the corresponding Fc-glycans on IgG$_4$). Melon Gel Resin contains a ligand that retains highly abundant proteins in serum, ascites and culture supernatants, while collecting total IgG in the flow-through fraction in mild buffer conditions without the need of elution steps. This makes the approach directly applicable for any further downstream processes. Melon gel overcomes the drawbacks of protein A or G based affinity purification, which required elution and is selective for certain IgG isotypes. This makes the approach highly suitable for the characterization of total IgG by mass spectrometry. Due to high saline (NaCl, 0,9%) concentrations in the BALF, isolation of IgGs from the fluid had to be adjusted. After an initial melon gel enrichment, samples were desalted using the Thermo Fisher Scientific Zeba™ Spin Columns (7K MWCO, 2 ml) followed by a second melon gel treatment, which gave the best, most reproducible results in the optimisation experiments.

Quantification in gel based and liquid based proteomics

Quantification of the IgG Fc-glycopeptides in project II was performed via label free proteomics and using a profiling approach [44, 139]. Thus, glycopeptides were identified by their characteristic retention times in the ion chromatograms of the monoisotopic ions (<10 ppm deviation from the theoretical mass values). Sixty-three glycopeptide variants of Fc-glycans N-linked to tryptic peptides were screened. Glycopeptide ion abundances were integrated using the ion intensities in the ion chromatograms and normalized to total content of Fc-glycosylated IgG from IgG_1, $IgG_{2/(3)}$ and IgG_4 peptides, respectively

Project III focused on the investigation of the airway epithelial proteome in early stage COPD by two complementary proteomic approaches: 2D-DIGE and TMT mass spectrometry. Both approaches aim to universally characterise the proteome. Protein separation, denaturation, reduction, alkylation and digestion were either performed in gel (2D-DIGE) or in solution (TMT-MS). Quantification of proteins in both methods was relative based on specific labelling techniques.

Quantitative intact proteomics by 2D-DIGE was performed using the GE healthcare Amersham™ CyDye minimal DIGE fluorochrome labelling before protein separation. There are three available CyDye labels (Cy2-blue, Cy3-green and Cy5-red). The triplex is matched with regards to charge and size (~500 Da). This is essential to facilitate a superimposable separation pattern in the 2-dimensional separation based on charge and size, in order to allow quantification and identification of the same protein spot during co-separation of three samples in the same gel (triplex experimental design). Designating one label to an internal reference pool enables correction of technical variation between 2D-gels [140]. CyDye fluors have NHS ester reactive groups, and are covalently bound to the epsilon amino group of lysine in proteins via an amide linkage [104] (Figure 14A). The quantity of 400 pmol CyDye fluors recommended for 50 µg of protein is limiting the reaction. Hence, this method is referred to as minimal labelling. Only 1-2% of the total number of lysine residues are labelled, guaranteeing labelling with a single flour chrome per protein moiety. The single labelling strategy thereby prevents smearing or spot trains due to variable numbers of Cy-labels on the same protein species. Due to the limited amounts of clinical material in project III, we compared the dynamic range by working with half the amounts of samples (25 µg protein). The loss in sensitivity was around 20% in our adapted workflow. 2D-DIGE quantifies the intact protein based on fluorescent intensities and therefore allows the detection of PTMs and isoforms. Altered protein spots, however, generally result in the identification of multiple proteins and careful selection based on pH, MW range and protein sequence coverage is required for true positive identification in this approach.
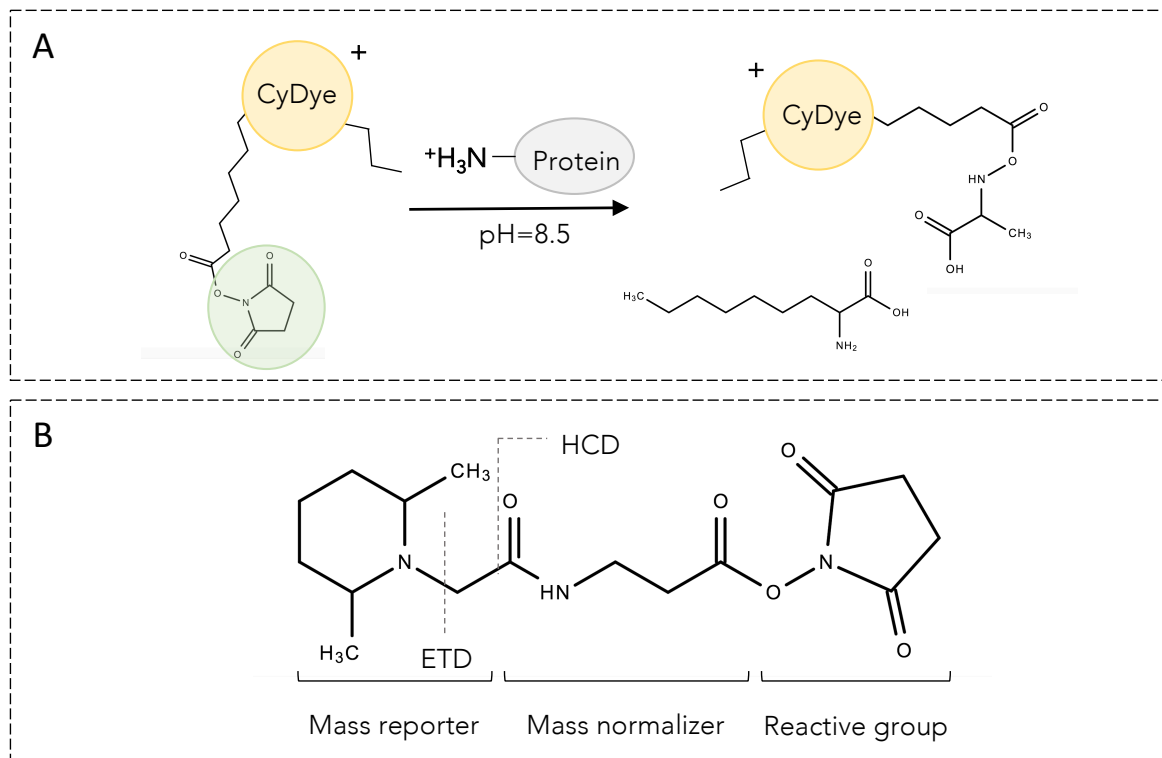
Figure 14| Labelling reactions of (A) CyDye Fluors attached to a protein and (B) the general structure of labels used in TMT-MS.

Labelled shotgun proteomics in project III were performed using the Thermo Scientific TMT 10plex™ mass tags. As in DIGE, TMT-MS is also based on multiplexing samples with distinct labels and thereby providing the possibility to include an internal reference pool to normalize for technical variation between experiments. Each label within one set has the same mass (*i.e.*, isobaric) and the general chemical structure is illustrated in Figure 14B. Based on the same total mass, the same peptide throughout all samples in one set appears in the same peak in the MS full scan. After fragmentation, the distinct reporter ions for each sample (*i.e.*, 126-131Da) in the low mass region of the MS/MS spectrum can be used for relative quantification. TMT-MS based proteomics have a higher sample through put compared to DIGE. PTMs and protein isoforms are difficult to detect with TMT-MS, however membrane associated proteins are captured instead.

ClusterMHCII software

Most MS/MS search engines, such as Mascot, are "protein oriented", *i.e.* are based on determining whether the protein is present in the sample or not and not designed for the identification of intact peptides, where only a few peptides per proteins are identified. This results in lower and sometimes sub-threshold scores compared with "normal" proteomics experiments. Therefore, scores of identified HLA-DR bound peptides in project I were
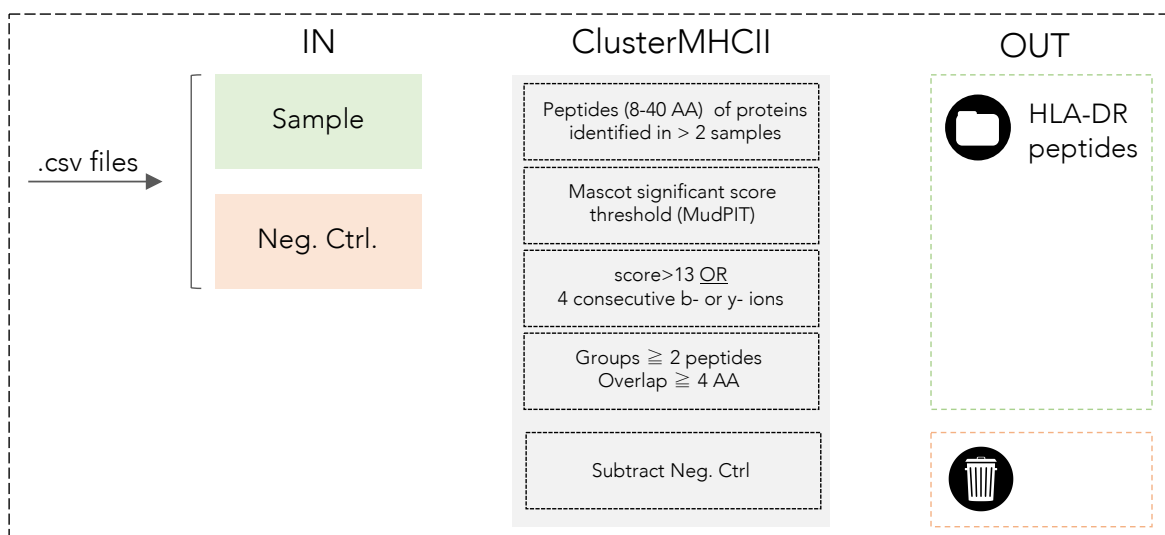
Figure 15| Flowchart of the ClusterMHCII algorithm showing two input directories for sample and negative control, filtering criteria and results generated in the out directory consisting of a HLA-DR peptide list and peptides of the negative control passing the filter.

recalculated using Percolator and filtered using the ClusterMHCII software an algorithm written in Perl to remove false positive identifications and nonspecific peptides. The Percolator score is an adjusted Mascot score with a threshold of 13 and corresponds to a false discovery rate (FDR) of 5%. The filtering only keeps peptides that belonged to proteins identified with significance scores (Percolator and MudPIT) in at least one sample, with either a peptide score higher than 13 or at least four consecutive b- or y-ions. Additionally, peptides are required to form clustering nests with groups of at least two peptides with an overlap of at least four amino acids. As a final filter, all peptides identified in the flow through of IP are subtracted. A simplified workflow of the main filter criteria of the software is illustrated in Figure 15 The software is available online (http://pkki.mbb.ki.se/tiki/tiki-index.php?page=ClusterMHCII).

Missing values

High throughput methods in proteomics and metabolomics produce large scale datasets. However, due to technical limitations such as imperfect identification of proteins or metabolites and the limited sensitivity of peptide detection technologies, missing values is a commonly occurring issue in mass-spectrometry based methods [141, 142]. Current technologies allow for the detection of only one-third to one-half of all coded proteins and thus leave a significant number of proteins left below the limit of detection (LOD) or lowest limit of quantification (LLOQ) [143]. However, missing values remain problematic in multivariate data modelling. In order to overcome this limitation in projects III and IV, imputation by k-nearest neighbours were used (k=10), which is described as a routine approach in metabolomics [144]. Proteins detected in ≥75% of the samples in all sub-groups

were included. The algorithm used estimations of missing values sequentially from k nearest neighbour values with the least missing rate.

Methods of choice

Overall, numerous quantitative proteomic approaches are available. It is, after all, a matter of analytical goals and scientific questions that are being addressed which decide the optimal method of choice. Pre-processing of samples by enrichment or depletion mainly depends on the abundancies of target analytes. Fewer experimental steps generally achieve higher reproducibility and keep technical variation to a minimum. 2D-DIGE was considered the "gold standard", since it was first popularized for the simultaneously quantification of large numbers of proteins. However, large scale studies are labour-intensive and time-consuming. Continuous improvements in liquid chromatographic separation methods with higher resolution exceed 2D gel electrophoresis [121]. Labelling techniques demand expensive fluorescent or isobaric tags and intensive sample preparation, which increases technical variability. Doubts about quantitative accuracy of labelling techniques have been raised [145, 146]. However, multiplexing provides high sample throughput and the ability to normalize for technical variation with internal standards is beneficial. Recently, there has been a movement towards label free quantification. MS spectra peak areas do not correlate with the quantity of distinct peptides in samples but correlate with chemically identical peptides. Thus, the relative quantification of the same peptides between runs can be utilized. However, no internal reference pool can be included in the same approach when compared to shotgun proteomics based on labelling techniques. The aforementioned improvements in chromatography further enable separation with very low run-to-run variation [121]. All methods bring advantages and disadvantages and need to be tailored to the goal of the analytical experiment together with the research funding available. A combination of several approaches as described in projects III and IV might give studies more depth and statistical power to forward the scientific questions at hand.

The benefit of multi-omics investigations

Project III and IV were carried out using mass spectrometry-based methodologies to investigate the same study cohort, but targeted very distinct molecules and anatomical compartments. Proteomic analysis of the airway epithelium (project III) gives insight into ongoing processes in the first line of defence against inhaled toxicants, such as tobacco smoke. Investigations of the serum metabolome (project IV) represents the molecular down-stream of the proteome in the omics cascade and a picture of the final systemic events. The non-invasive nature of serum metabolomics analyses has the advantage of being more applicable to the clinical setting, potentially providing the clinical phenotype that defines the disease, and can be particularly revealing regarding clinical sub-phenotypes. Non-targeted metabolomics was the main approach used in project IV to perform a broad molecular

screening of serum metabolites. Targeted metabolomics were further applied to confirm the markers of interest revealed by the non-targeted metabolomics. Complementary screening at multiple molecular levels in multiple anatomical locations represents a powerful tool in systems medicine, and correlation between molecular events at the site of inflammation in the lung, and systemic events may be useful in the development of diagnostic or prognostic tools. Biomarker discovery in diseases which lack broad understanding of the mechanisms that trigger disease onset and progression, such as COPD, may especially profit from multi omics approaches and system biology investigations.

Comments on statistical methods

A range of univariate statistical analyses were used throughout the different projects based on the different hypothesis tests and statistical requirements of the data sets. The tests utilised included T-test (normal distributed data), Mann Whitney (skewed distribution), and the Kolmogorov-Smirnov test (non-parametric test). Correction for multiple hypothesis testing was performed using FDR according to Benjamini and Hochberg [147], or according to Storey (q) [148].

In addition, receiver operating characteristic (ROC) curves were employed, to determine the threshold laboratory value that separates a clinical diagnosis from healthy controls. In a ROC curve, the true positive rate (sensitivity) is plotted as a function of the false positive rate (100-specificity) for different cut-off points of a parameter. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish and indicates the overall diagnostic accuracy of a technique and/or biomarker.

Multivariate analyses were performed with SIMCA software version 14 (Umetrics AB, Umeå, Sweden) applying both unsupervised principal component analysis (PCA) and supervised orthogonal projections to latent structures (OPLS) discrimination analysis, as well as inner correlation analyses using partial least squares regression (PLS). Clinical datasets can be difficult to analyse due to a high heterogeneity in the protein and metabolite profiles and variable selection in order to optimize models can be challenging. In projects III and IV, we performed a new approach for variable selection based on absolute p(corr) values [125]. P(corr) represents a scaling of the loadings from a multivariate model to a correlation coefficient between the model and the original data matrix. Models providing maximum group separation were generated and variables with a specific p(corr) cut off corresponding to $p<0.05$ were selected for subsequent models. This method provides efficient variable selection of all relevant analytes. More stringent variable selection can then be performed in an iterative manner until a final biomarker model is revealed. For pathway analysis, the

former, more inclusive variable selection method was utilized in order to maintain all proteins and metabolites of interest for pathway enrichment analysis.

RESULTS AND DISCUSSION

Project I. Approach for identifying HLA-DR bound peptides

Main results

The optimized approach (Figure 16) used to identify HLA-DR peptides from low cell numbers was composed of cell lysis, immunoprecipitation (IP) of HLA-DR complexes, elution of immunopeptides, desalting and peptide concentration, identification with mass spectrometry and eventually peptide filtering using the in house developed ClusterMHCII software. Additionally, negative controls in the analysis (peptide presented in the flow through after IP) were added to the protocol.
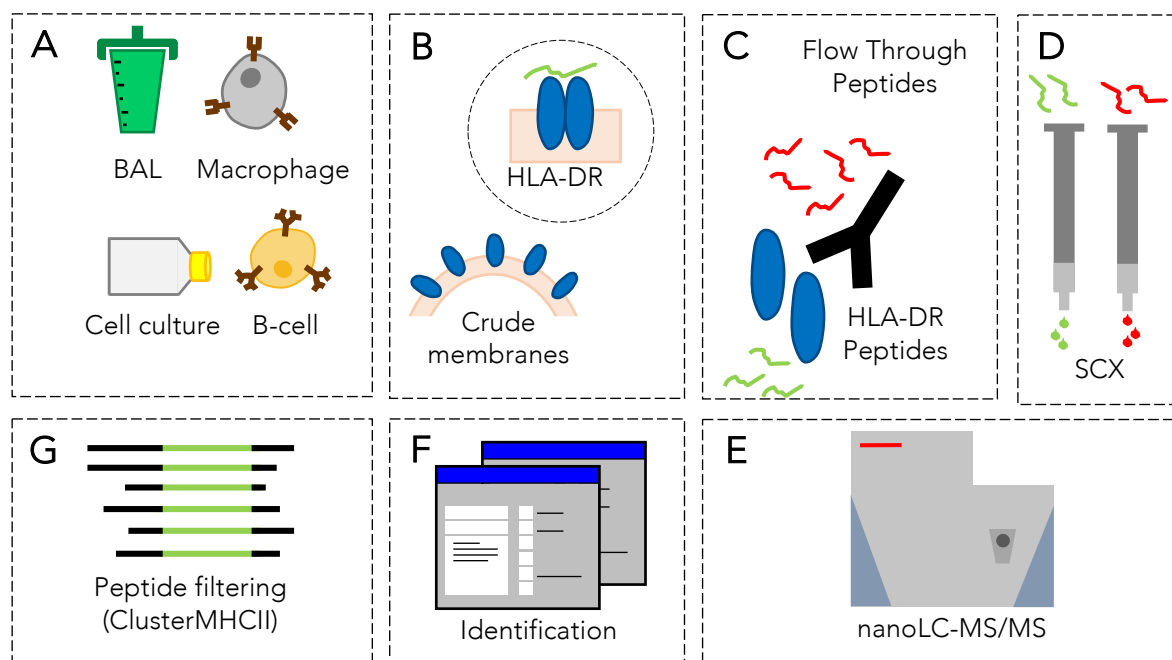
Figure 16| Schematic overview of the optimized approach to identify HLA-DR bound peptides from scarce material composed of: (A) EBV B cells and BAL cells, (B) crude membrane extraction, (C) immunoprecipitation (IP), (D) purification of eluted peptides (HLA-DR bound peptides) and peptides in the flow through of IP followed by strong cationic exchange (SCX) purification, (E) analysis using nanoLC-MS/MS, (F) peptide identification and (G) filtering using the ClusterMHCII software.

Specifically, changing the solubilizing detergent from the commonly used CHAPS [149-153] to the more MS compatible n-dodecyl –D-maltoside (DM) [154] increased spray stability during analysis. Improved washing steps of protein G reduced contamination as observed in lower polymer peaks in the MS spectra. In order to minimize non-specific binding, crude membranes were isolated prior to IP. IP was further improved and simplified by working with smaller volumes and elution yields were increased using acetonitrile at high concentrations. Mass spectrometry identification was enhanced by recalculation of peptide scores using Percolator and filtering results using the ClusterMHCII software, an algorithm written in Perl.

Steps of the final procedure were evaluated by analysing HLA-DR peptides from an EBV B cell line (sets of 4 replicates of 10×10^6 cells) comparing different conditions: the final protocol using crude membranes (*M*), isolation from total cell lysate (*T*), a freezing step in the procedure (*M**) and an incubation time with antibody to capture the HLA-DR complexes (*M2h*). Negative controls using an irrelevant as well as no antibody were also included in the evaluation. Additionally, peptides identified in the flow through of IP were subtracted from the results. In order to compare the different conditions, both the number of unique peptide sequences per million cells and the number of identified scans per million cells (spectral counting) were calculated.

Isolating HLA-DR bound peptides from crude membranes obtained a median of 11.95 scans/million cells (9.85 unique sequences/million cells) compared to a median of 4.50 scans /million cells (3.75 unique sequences/million cells) in the cases when working with total cell lysates (Figure 17A). Additionally, both freezing and a shorter incubation time (2h) with anti HLA-DR antibody reduced the yields of HLA-DR bound peptides from 11.95 scans/million cells to 2.00 and 2.80 scans/million cells respectively (Figure 17B). Furthermore, the overlap of proteins and peptides identified from total cell lysate and crude membranes with negative controls (proteins and peptides in the flow through of IP) was investigated. At the protein level, the overlap between both conditions was similar, however at the peptide level, the overlap was 7.5% using crude membranes compared to 15.5% from total lysate (Figure 17C). To take the reduced complexity of crude membranes into account we estimated relative protein abundancies and compared the protein ranks between the two conditions with the five highest ranked proteins in the negative control (flow through) shown in Figure 17D.
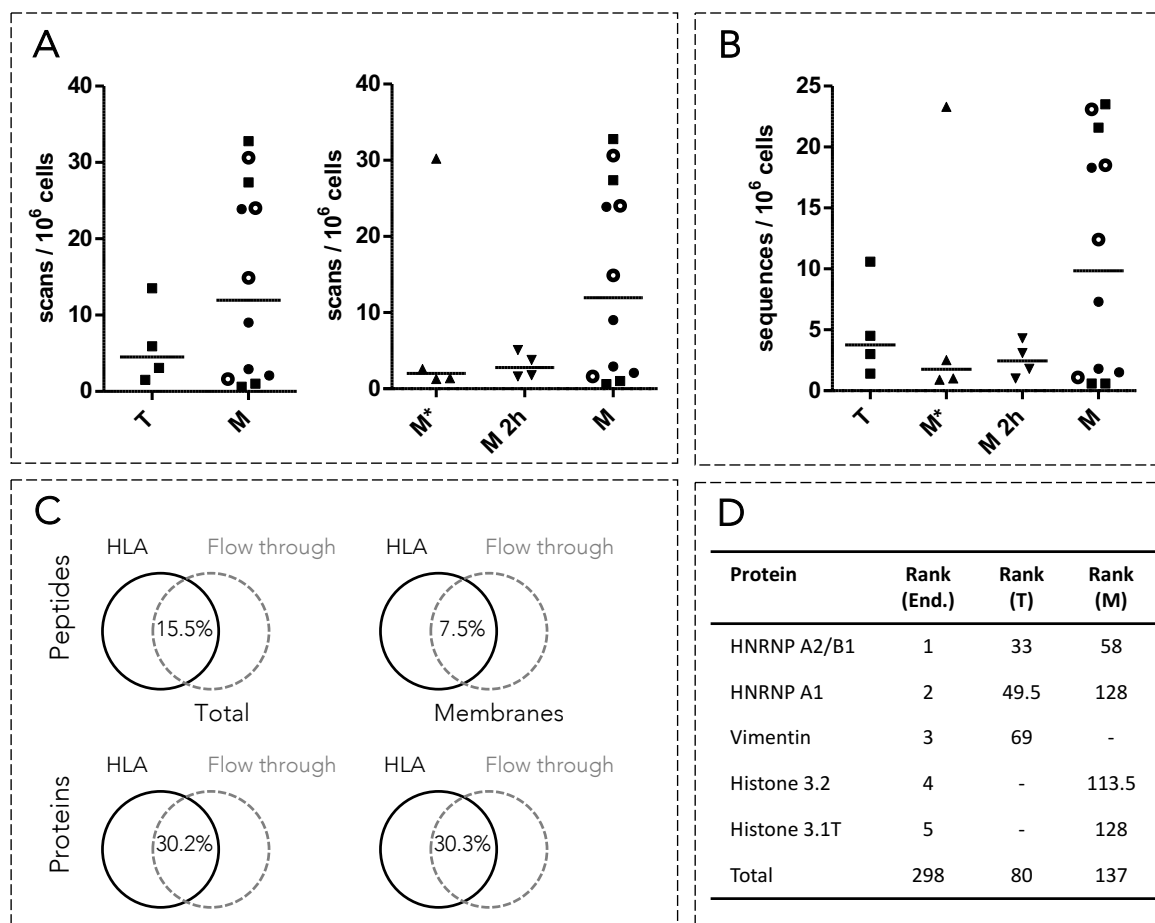
Figure 17| Evaluation of specific steps during the isolation of HLA-DR peptides from EBV B cells. (A) Identified scans per million cells comparing freeze-thawing membrane complexes (*M**), short incubation time with anti-HLA-DR antibody (*M2h*) and purifying peptides from membrane complexes (*M*) compared with a total cell lysate (*T*). The different symbols in M indicate 3 different sets of 4 replicates. (B) Number of unique peptide sequences per million cells for all conditions. (C) Overlap of peptides and source proteins between HLA-DR peptides and flow through peptides in purifying peptides from a total cell lysate and crude membranes. (D) Top 5 ranked proteins in flow through peptides and the corresponding ranks of the same protein in the HLA-DR peptide data set, identified from total cell lysate and crude membranes. *Adapted from [138].*

The optimized approach using crude membranes and an overnight incubation with anti-HLA-DR antibody was applied to human material in the form of alveolar macrophages obtained from bronchoalveolar lavage. In total, 1,434 unique peptide sequences belonging to 384 core sequences from 215 proteins were identified from seven patients. In three patients (PAT5–7), immunopeptides were investigated by multiple technical replicates: two replicates of 10×10^6 cells (PAT5), two replicates of 10×10^6 cells and one with 50×10^6 cells (PAT6) and four replicates of 10×10^6 cells (PAT7). On average, we identified 9.4 peptides/million BAL cells (Figure 18A, B). The overlap of peptides between the experimental replicates was 36.8% for PAT5 and 42.0% for PAT6. In PAT7 21.6% of the peptides were identified in all fractions, 40.8% were identified in at least three replicates, and 65.6% were identified in at least two replicates (Figure 18C).
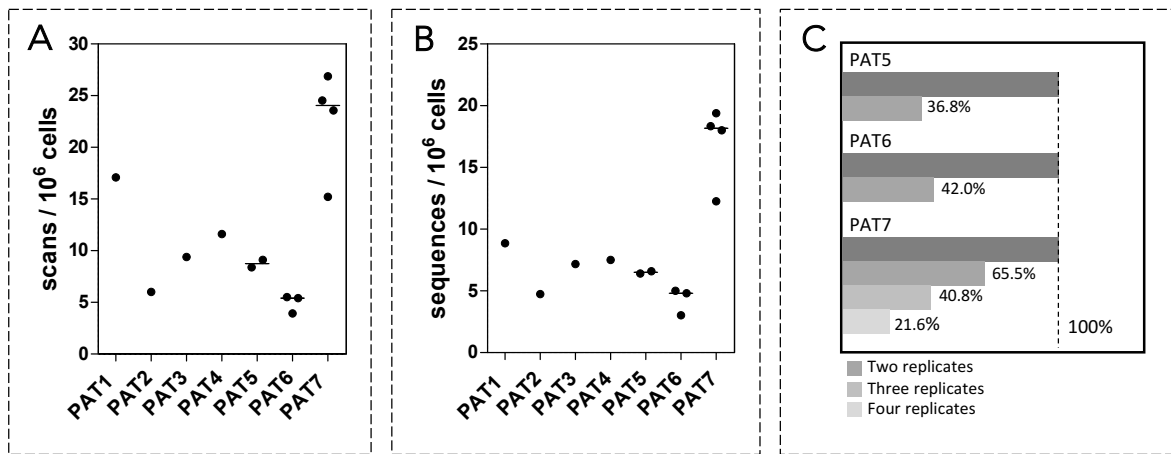
Figure 18| Identification of HLA-DR peptides from BAL cells of seven patients including technical replicates for PAT5-7. (A) Number of identified scans per million BAL cells for the seven patients (PAT1–7). Technical replicates were analysed for PAT5-7. (B) Number of unique sequences per million of BAL cells for the same set. (C) Overlap of identified peptides between two to four technical replicates (PAT5-7). *Panel A and B adapted from [138].*
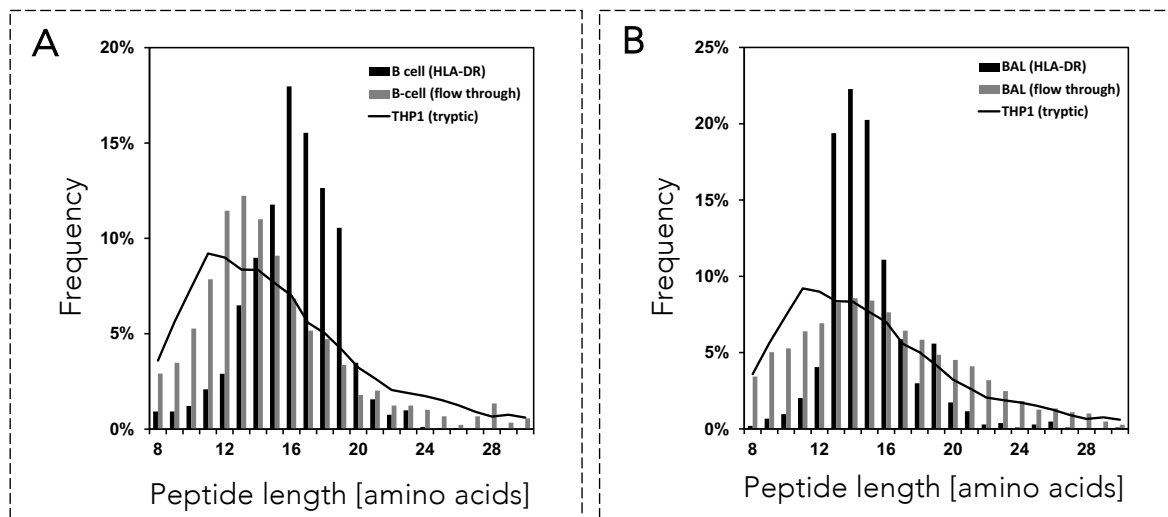

Figure 19| Peptide length distributions. (A) Peptide length distributions of identified HLA-DR bound peptides from EBV-B cells (condition M) compared with flow through peptides (from EBV-B cells) and tryptic peptides from a THP-1 cell line. (B) Peptide length distributions for the BAL data set (PAT7) including HLA-DR binding peptides, flow through peptides (from BAL cells) and the tryptic data set. *Adapted from [138].*

Finally, the chances of the identified peptides being HLA-DR derived was investigated by comparing distinct characteristics such as peptide length distribution and peptide binding motifs. Peptide length distributions of HLA-DR peptide identified from EBV B cells (condition *M;* Figure 19A) and BAL cells (PAT7; Figure 19B) were compared against the length of distributions of tryptic peptides and peptides identified in the IP flow through. HLA-DR bound peptides from EBV B cells and BAL cells showed a much narrower distribution around 16 amino acids for EBV B cells (SD 2.7) and 15 amino acids (SD 2.6) for BAL cells compared to peptides identified in the matching flow through and tryptic peptides.

Binding motifs of HLA-DR peptides of EBV B cells and BAL cells as well as peptides from the IP flow through were generated using the NNAlign prediction algorithm v1.4 (www.cbs.dtu.dk/services/NNAlign/) [155]. Predicted motifs were largely consistent with reference motifs for single alleles published from Nielsen *et al.* [156] (Figure 20). PAT7, which presents a heterogeneous subject regarding HLA-DR variant (DRB1*03,15), showed a binding motif consistent with a combination of the reference motifs of the single alleles. In order to compare our results to previously published studies and strengthen the results, we compared yields to two studies on similar cell types. The number of identified HLA-DR bound peptides compared to studies on DCs [149] and BAL cells [153] and was increased 100- and 13-fold respectively.
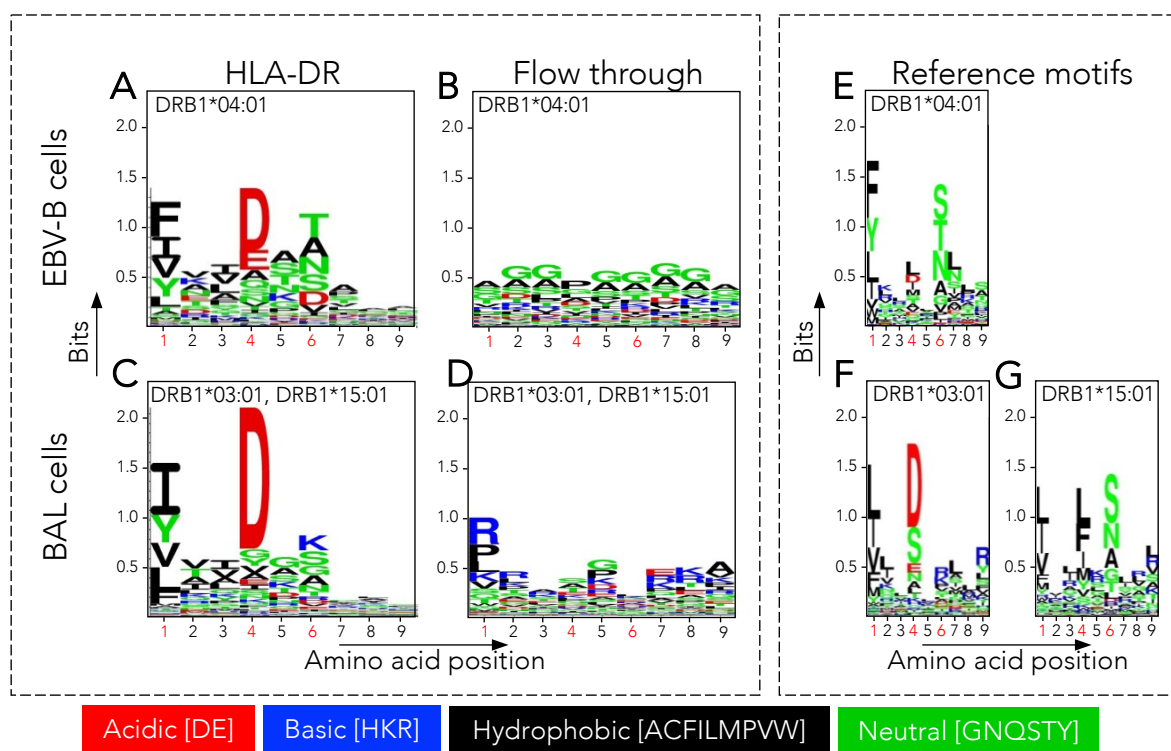
Figure 20| Predicted binding motifs for HLA-DR derived peptides using the NNAlign prediction server. (A) HLA-DR peptides from EBV-B cells (condition M), (B) flow through peptides from EBV-B cells, (C) HLA-DR peptides from BAL cells (PAT7), and (D) flow through peptides from PAT7. Panels to the left show reference motifs available at (www.cbs.dtu.dk/biotools/MHCMotifViewer-1.0/Home.html; [156]): (E) HLA-DRB1*04:01, (F) HLA-DRB1*03:01, and (G) HLA-DRB1*15:01. *Adapted from [138].*

Reflection and Discussion

With the introduction of specific changes to previously published methodologies [151-153] for the investigation of HLA-DR bound peptides, we were able to increase sensitivity and thereby enable the investigation of the immuno peptidome in low sample amounts. This is essential for the successful access of the HLA-DR peptide profiles in the lungs of individual healthy subjects. Investigations into healthy individuals are a unique opportunity when working with BAL fluid to enable comparisons of peptide repertoires in health and disease in

order to identify disease specific antigen(s). Ethically, access to healthy tissue in most diseases is not allowed therefore only comparisons between diseased phenotypes is possible [151].

Enrichment of crude membranes prior to IP originated from initial problems working with clinical material and polymer contamination made MS analysis impossible. Isolating membranes added a purification and enrichment step, which was able to reduce polymers to an acceptable degree. Previous studies performed investigations into the HLA-DR peptides in total cell lysates from large amounts of starting material (in the range of grams) [151, 152]. However, working with small amounts gives interfering contamination more opportunities to mask low abundance peptides of interest or even make the analysis impossible.

By experiments using an EBV B cell line, we validated different parameters throughout the protocol and the best results were achieved by isolating crude membranes and incubating the samples with antibodies for IP overnight. Working with method development highlights the importance of experimental design and validation of every step throughout the procedure. Every continuous decision of experimental readouts and subsequent adjustments are essential for successfully developing a solid protocol. In the first developing steps, technical replicates to ensure reproducibility and the control of conditions changes are key. Working with cell lines for example, and having adequate access to material (biological replicates) is not always possible but it enables the reduction of biological variance. However, parallel validations in target material (*i.e.,* BAL cells in this project) should not be disregarded, since conditions might also be tailored to every type of material.

The optimized approach presents an experimental method with a degree of variability in peptide recoveries from technical replicates. Isolation of HLA-DR peptides, especially from low numbers of cells, include a number pre-analytical steps required to access HLA-DR peptides. Higher reproducibility could be achieved by further improvements using more automated steps and continuous modes (*e.g.,* spin columns for IP and SCX). This would furthermore increase sample throughput, which is especially interesting for clinical studies in larger cohorts. In this context, quantification of the HLA-DR peptides could also be considered, however to date, this has been difficult to implement. A final comment in this section of reflection and discussion concerns the sample availability of BAL fluid. All patients included in this work were heterozygous patients meaning the expression of two variants of HLA-DR molecules in a ratio of 1:1 are present. This might indicate that the results from 10×10^6 cells of heterozygous patients rather reflect peptide repertoires of 5×10^6 cells of homozygous patients, which would present an even more sensitive approach than initially claimed for BAL cells.

Project II. Fc-galactosylation status of IgG$_4$ in sarcoidosis

Main results

A significant correlation between agalactosylated Fc-glycans and disease revealed the potential of these post-translational modifications as disease markers (Figure 21A). Multivariate modelling was performed to investigate the overall trend and the intra-individual correlation of BALF and serum profiles. PCA analysis shows a separation (R^2=0.46, Q^2=0.35, two components) between BALF and serum as well as intra-individual differences according to disease and/or age (Figure 21B).

To deeper investigate which Fc-glycans are driving the separations, supervised multivariate modelling was performed via OPLS-DA. The resulting analysis which included sample types from all groups, significantly separated BALF from serum (R^2=0.90, Q^2=0.85, *p*[CV-ANOVA]=1.3×10^{-26}). From the 64 Fc-glycans we screened for, 31 correlated with 95% confidence with serum (ten from IgG$_1$, twelve from IgG$_{2/(3)}$ and ten from IgG$_4$) compared to only five Fc-glycans that correlated with the BALF (three from IgG$_2$ and two from IgG$_4$). This may indicate a higher complexity of the Fc-glycan composition in serum than in BALF. This was further confirmed using univariate statistics with higher ranked low abundant forms like the monoantennary (A1) and bisected (B) Fc-glycans in serum.
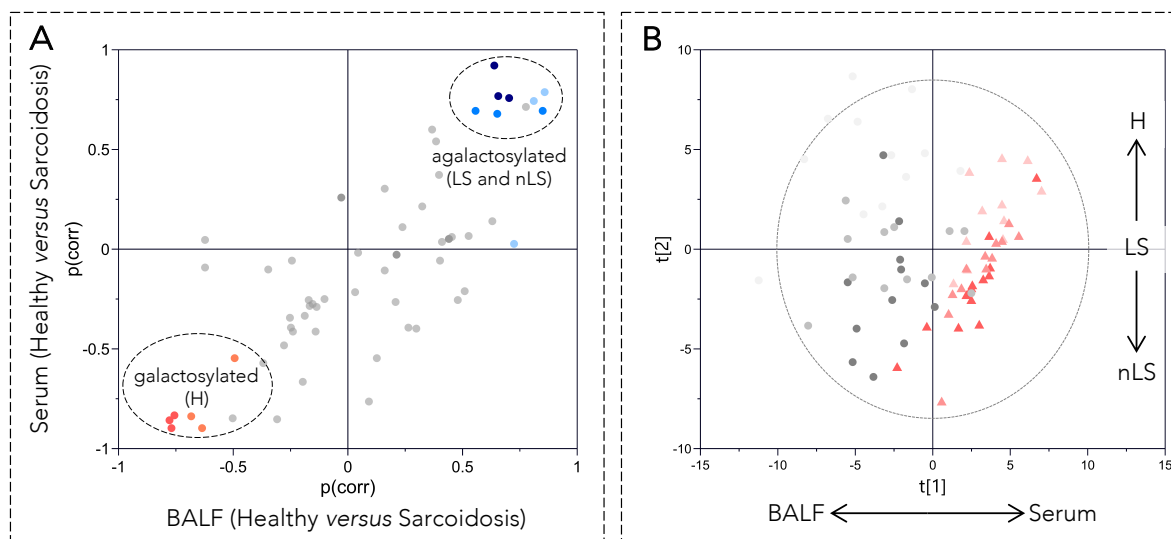

Figure 21| Overall trends of Fc-glycans in the study cohort (A) SUS plot showing galactosylated Fc-glycans correlating with a healthy state (H), whereas agalactosylated with sarcoidosis (LS and nLS). (B) PCA scores plot of the Fc-glycan profiles measured in the serum and BALF samples. Serum and BALF profiles are separated along component 1 (x-axis). Healthy and disease (LS and nLS) separate along component 2 (y-axis).

Furthermore, the identification of disease specific Fc-glycans to classify healthy controls and sarcoidosis patients was investigated. OPLD-DA modelling was able to classify controls from sarcoidosis patients (R^2=0.28, Q^2=0.24, p[CV-ANOVA]=4.4×10^{-12}). Interestingly, the study groups align along the predictive component with Löfgren patients clustering between healthy subjects and non-Löfgren (nLS) patients (healthy: tcv[1]=2.9±2.4, LS: tcv[1]=-0.9±2.1), nLS: tcv[1]=-2.1±2.1). This is particularly striking since nLS present a more severe and chronic form of the disease. From uni- and multivariate data analyses, it was evident that the Fc-glycan ratio between galactosylated and agalactosylated glycans was significantly different between patient subgroups and healthy individuals. The log transformed ratio between the main agalactosylated form (FA2) and the main digalactosylated form (FA2G2) correlated best between matrixes for IgG_4 (R^2=0.95), Figure 22A. Correlations or the other IgG variants were slightly lower (IgG_1 [R^2=0.87, $IgG_{2/3}$ [R^2=0.76], Figure 22B and C).

Aging has an effect on glycosylation in the body. In order to compensate for the differences in age between healthy and disease cohorts and to target differences due to inflammation/disease we applied an age correction (via Glycoage index, [157]) on log(FA2/FA2G2) of the IgG isotypes. Following age correction, the classification between groups still remained significant particularly for IgG_4. At a cut-off of 0.22, the parameter could distinguish healthy from sarcoidosis patients with a sensitivity of 87% and specificity of 67% and had a likelihood ratio of 2.61 (Figure 23A). Healthy subjects and sarcoidosis patients were distinguished in the serum with an AUC of 79±8%, or separately nLS: 87±7% and LS: 71±11% presenting a potential serum biomarker for chronic sarcoidosis (Figure 23B-D).
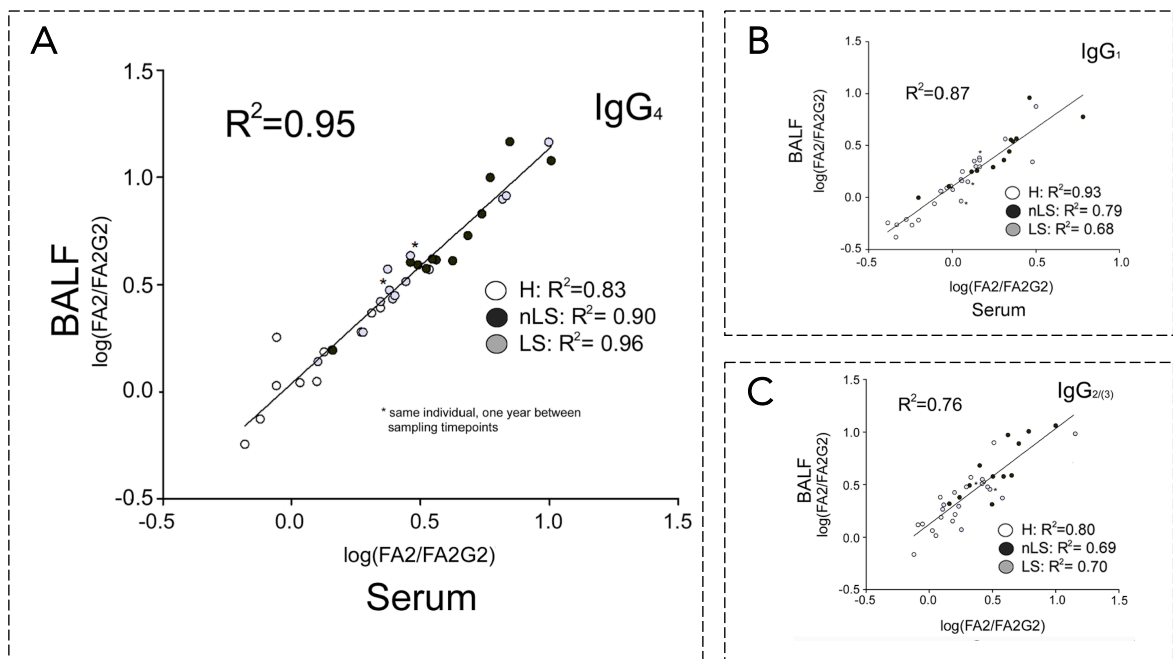

Figure 22| Intra-individual correlation between log(FA2/FA2G2) in serum *versus* BALF in (A) IgG_4, (B) IgG_1 and (C) $IgG_{2/(3)}$ respectively. As indicated in the plots with (*), for the correlation analysis we used both intra-individual data time points for the LS-patient that was sampled twice, a year between the sampling dates.
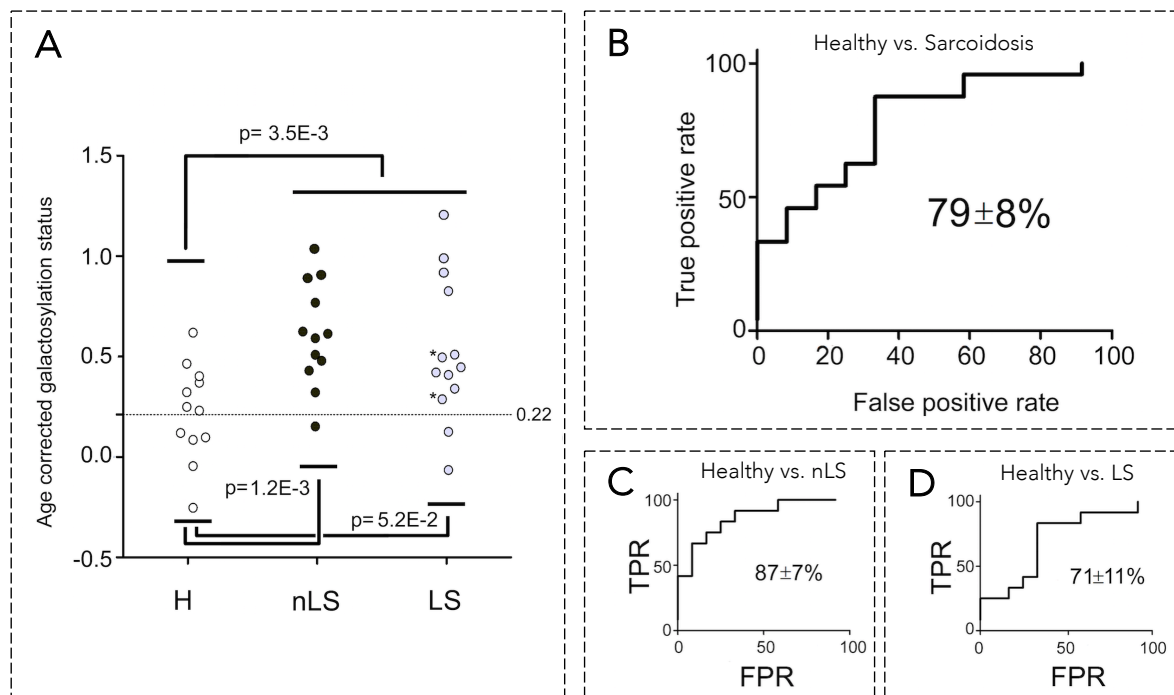
Figure 23| The predictability of the age corrected serum galactosylation status of IgG$_4$ for sarcoidosis. (A) Age corrected galactosylation status in healthy (H), non-Löfgren syndrome (non-LS) and Löfgren syndrome (LS). The data points of the individual sampled at two occasions are indicated with *. For significance and ROC-curve analysis the mean between the sampling points for this individual were used. At a cut-off of 0.22, the parameter could distinguish healthy from sarcoidosis patients with a sensitivity of 87% and specificity of 67% and had a likelihood ratio of 2.61. (B) ROC-curve for distinguishing healthy individuals *versus* the sarcoidosis patients. (C) ROC-curve for distinguishing healthy individuals *versus* non-LS patients. (D) ROC-curve for distinguishing healthy individuals *versus* LS patients. Area under curve ± Standard Error.

Reflection and Discussion

Sarcoidosis is an inflammatory lung disorder and it is still challenging to distinguish sarcoidosis from other interstitial lung diseases as well as to predict disease course and to distinguish sub-phenotypes such as nLS and LS. The fact that the inflammatory status in the lung could be captured by a simple blood test has enormous implications for clinical practice in sarcoidosis diagnostics. The significant correlations of Fc-glycan profiles of IgG isoforms between serum and BALF show that it is possible to capture the inflammatory status of the lungs from the blood.

In order to reveal disease specific markers, Fc-glycan profiles between healthy and the two sub-phenotypes of sarcoidosis were compared. The log transformed ratio between the main agalactosylated form (FA2) and the main digalactosylated form (FA2G2) were confirmed as biomarker candidates to distinguish the study groups. However, age correction in this cohort due to different means in the age of healthy individuals and sarcoidosis patients limited the marker discovery to isoforms where correction factors have been published. Additionally, the disease specificity needs to be compared with other inflammatory diseases

to exclude alterations that are only due to general inflammation and ensure specificity for Sarcoidosis. Alterations in IgG Fc-galactosylation have been reported in numerous studies, particularly the galactosylation status of IgG_1 in rheumatoid arthritis has been extensively studied [158-161].

This work presents a pilot investigation and bigger study cohorts are required to investigate the identified marker more carefully. Additionally, gender differences and/ or similarities would be of great interest, especially since autoimmunity is more prevalent in women compared to men, and sarcoidosis shares autoimmune characteristics. Overall this project opened a new area for promising investigations regarding biomarker discovery and underlying disease mechanisms in sarcoidosis.

Project III and IV. Gender-enhanced alterations in the airway epithelial proteome and the serum metabolome in early stage COPD

Main results

Investigations into epithelial proteins using two proteomics approaches (TMT-MS and 2D-DIGE), as well analysis of serum metabolites using a non-targeted mass spectrometry based approach revealed distinct differences between smokers and early stage COPD. Smokers classified as significantly different from COPD patients in all platforms (Figure 24). Gender stratified models furthermore showed that the classification was mainly driven by the female population. Gender enhanced alterations of the epithelial proteome and serum metabolome between Smokers and COPD groups were driven by distinct markers.
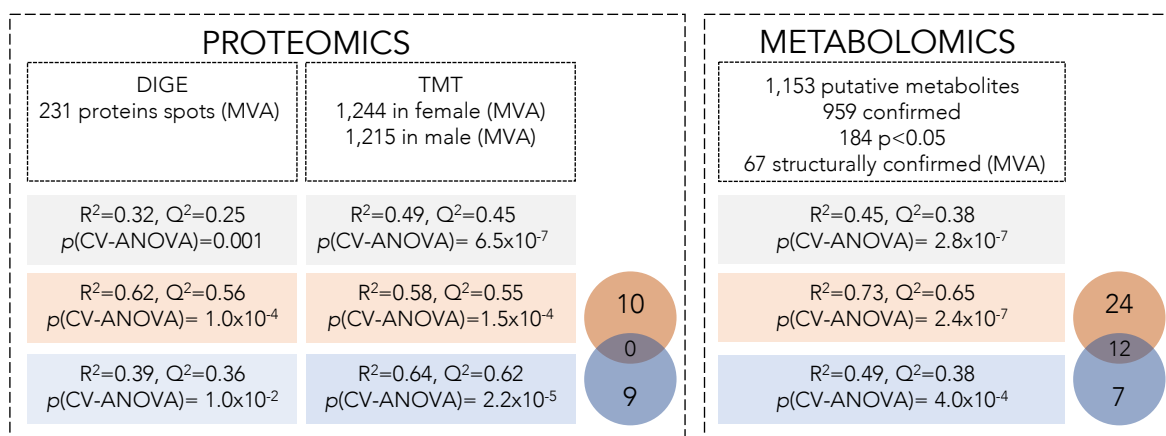

PROTEOMICS

DIGE 231 proteins spots (MVA)	TMT 1,244 in female (MVA) 1,215 in male (MVA)
R^2=0.32, Q^2=0.25 p(CV-ANOVA)=0.001	R^2=0.49, Q^2=0.45 p(CV-ANOVA)= 6.5x10^{-7}
R^2=0.62, Q^2=0.56 p(CV-ANOVA)= 1.0x10^{-4}	R^2=0.58, Q^2=0.55 p(CV-ANOVA)=1.5x10^{-4}
R^2=0.39, Q^2=0.36 p(CV-ANOVA)= 1.0x10^{-2}	R^2=0.64, Q^2=0.62 p(CV-ANOVA)= 2.2x10^{-5}

10
0
9

METABOLOMICS

1,153 putative metabolites 959 confirmed 184 p<0.05 67 structurally confirmed (MVA)
R^2=0.45, Q^2=0.38 p(CV-ANOVA)= 2.8x10^{-7}
R^2=0.73, Q^2=0.65 p(CV-ANOVA)= 2.4x10^{-7}
R^2=0.49, Q^2=0.38 p(CV-ANOVA)= 4.0x10^{-4}

24
12
7

Figure 24| Overview of multivariate models of shared (grey) and gender specific classifications (orange female, blue male). Number of driving proteins and metabolites for the gender model and their overlap are shown in the Venn diagrams.

Specifically, proteomics using TMT-MS identified a total of 1,244 proteins in the female Smoker and COPD groups and 1,215 proteins in corresponding male groups. Univariate statistics identified 123 proteins which were significantly altered between female Smoker and COPD groups ($p<0.05$), of which 4 passed FDR correction ($q<0.20$). In the corresponding males, 167 proteins were significantly altered ($p<0.05$) and 123 proteins with $q<0.20$. Ten of these proteins ($p<0.05$) were shared between the genders. Multivariate modelling was able to significantly classify COPD patients from Smokers with 9 proteins providing 60% predictive power ($R^2=0.64$ $Q^2=0.60$; $p[\text{CV-ANOVA}]=1.0\times10^{-3}$). Further investigations of the enhanced differences between females and males showed multivariate modelling stratified by gender. In females, a subset of 10 proteins provided significant ($p[\text{CV-ANOVA}]=1.5\times10^{-4}$) separation between the Smoker and COPD groups with 55% predictive power (Figure 25A, B; $R^2=0.58$, $Q^2=0.55$). In males, a subset of 9 different proteins provided classification ($p[\text{CV-ANOVA}]=2.2\times10^{-5}$) with 62% predictive power (Figure 25C,D; $R^2=0.64$, $Q^2=0.62$). In the complementary proteomic analysis using 2D-DIGE, a total of 231 protein spots could be quantified. Univariate analysis revealed 26 protein spots (14 identified proteins) significantly altered between the Smoker and COPD group in the female cohort ($p<0.05$, $q<0.30$) and 23 (7 identified proteins) protein spots in corresponding males ($p<0.05$, $q<0.30$). Only one protein spot was shared between genders, however this could not be identified.
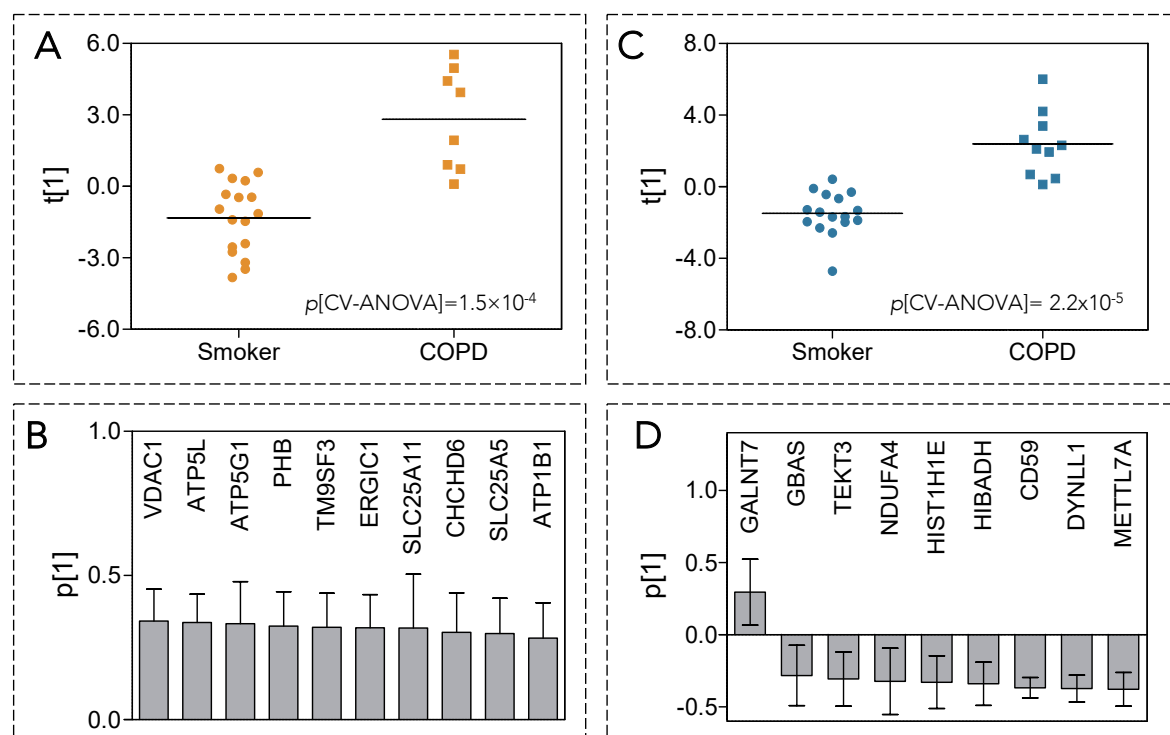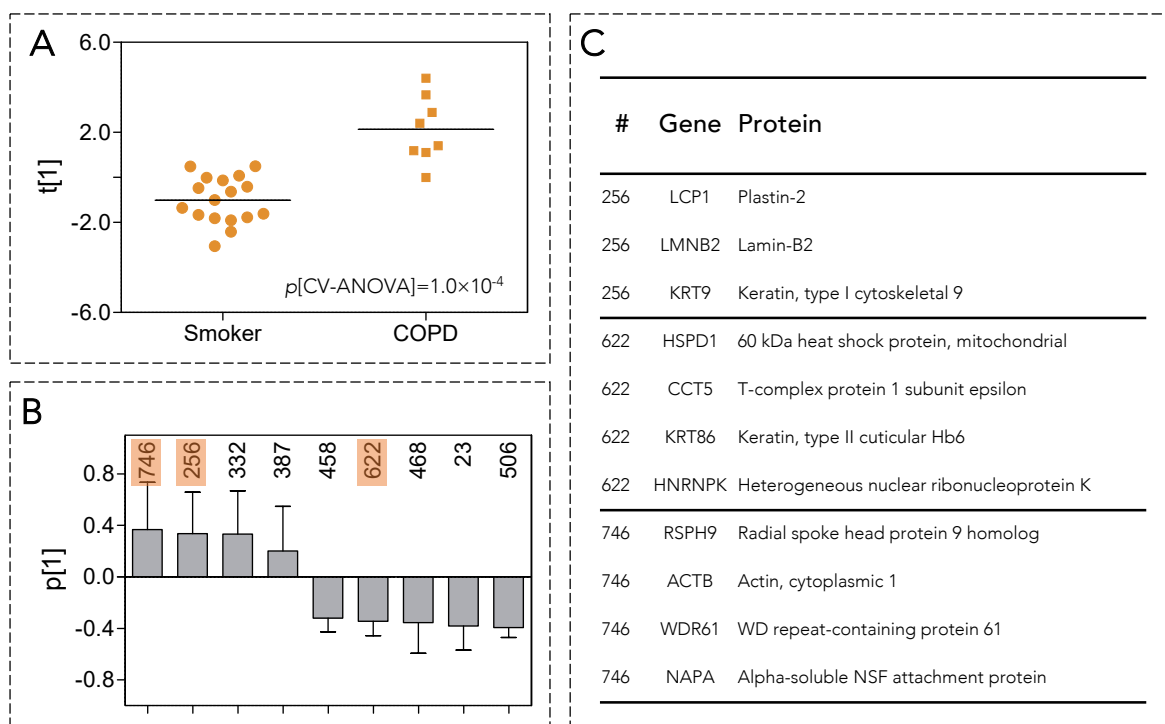

Figure 25| Optimized orthogonal projections to latent structure discriminant analysis multivariate models of proteins analysed with TMT-MS. (A) Scores plot of female smokers *versus* females with COPD ($R^2=0.58$, $Q^2=0.55$, $p[\text{CV-ANOVA}]=1.5\times10^{-4}$). (B) Corresponding loadings of the most prominent proteins driving the separation. (B) Scores plot of male smokers *versus* females with COPD ($R^2=0.64$, $Q^2=0.62$, $p[\text{CV-ANOVA}]=2.2\times10^{-5}$). (D) Most prominent proteins identified by TMT-MS for driving the separation in males.

Supervised multivariate modelling furthermore confirmed the female sub-phenotype identified in the TMT platform ($R^2=0.62$, $Q^2=0.56$; p[CV-ANOVA]= 1.0×10^{-4}, Figure 26 A-C).

Profiling of circulating metabolites in the serum of smokers and early stage COPD patients identified 1,153 putative metabolites, of which 959 passed quality control. Out of these metabolites, 184 passed a significance threshold of $p<0.05$ and were selected for structural validation and 67 were confirmed by MS/MS and/or matching the reference standards. Multivariate modelling was performed based on these confirmed metabolites. OPLS-DA could significantly distinguish Smokers from COPD patients with 38% prediction ($R^2=0.45$, $Q^2=0.38$; p[CV-ANOVA]=2.8×10^{-7}). Gender stratified models revealed a stronger separation of COPD patients and Smokers in females ($R^2=0.73$, $Q^2=0.65$; p[CV-ANOVA]=2.4×10^{-7}, Figure 27A, B) than in males ($R^2=0.49$, $Q^2=0.38$; p[CV-ANOVA]= 4.0×10^{-4}, Figure 27C, D). 36 metabolites were driving the separation in women whereas only 19 were most prominent in classifying male Smokers from COPD patients. The gender specific models overlapped with 12 of the most prominent metabolites driving the classification in females and males.

Figure 26| Optimized orthogonal projections to latent structures discriminant analysis multivariate models of proteins analysed with 2D-DIGE. (A) Scores plot of female smokers *versus* females with COPD ($R^2=0.62$, $Q^2=56$, p[CV-ANOVA]=1.0×10^{-4}). (B) Most prominent proteins identified by 2D-DIGE for driving the separation. (C) List of proteins from successfully identified protein spots.
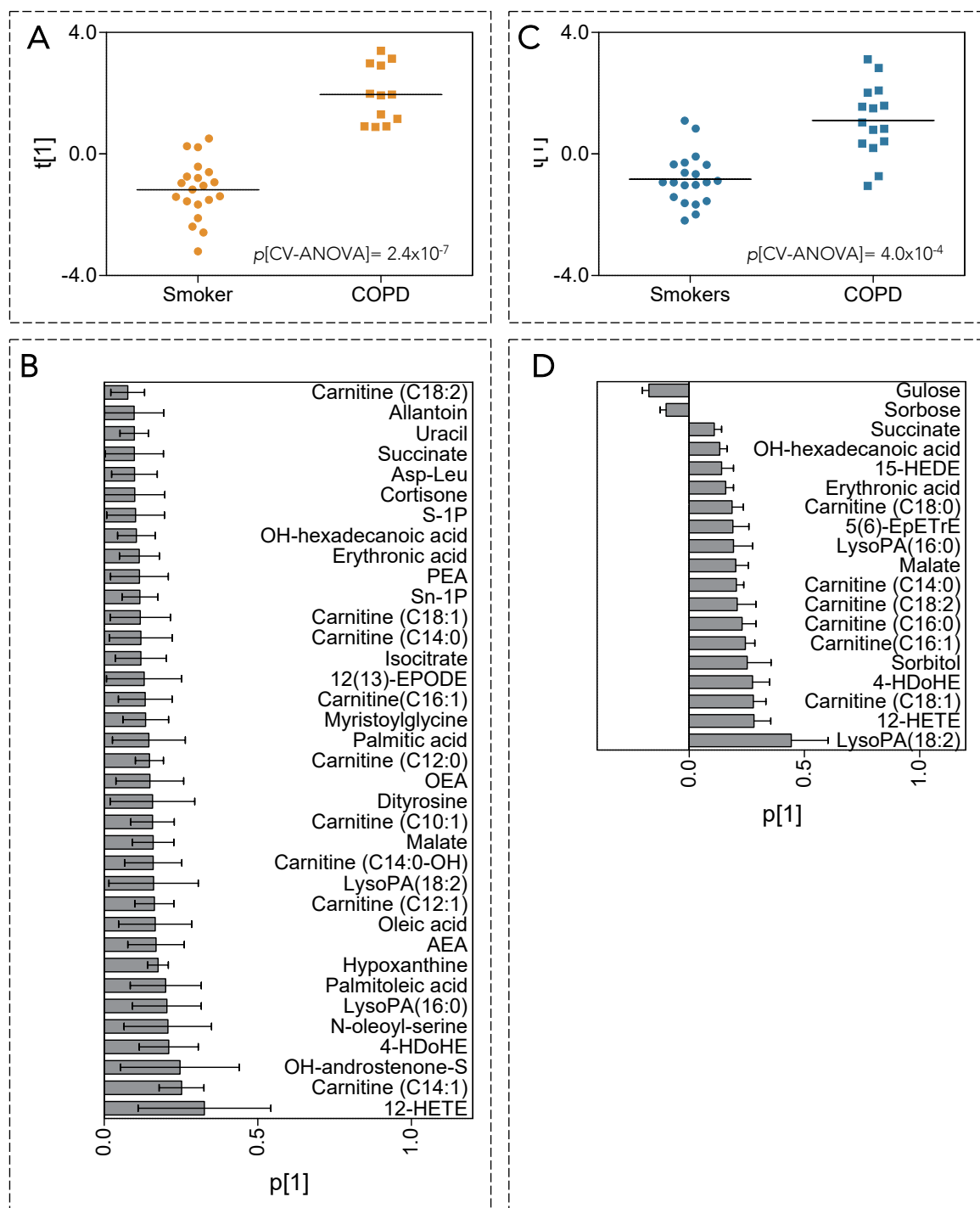
Figure 27| Optimized orthogonal projections to latent structures discriminant analysis multivariate models of confirmed serum metabolites. (A) Scores plot of female smokers *versus* females with COPD (R^2=0.73, Q^2=0.65, p[CV-ANOVA]=2.4x10[-7]). (B) Corresponding loadings of confirmed metabolites that were the most prominent for driving the separation of female smokers *versus* females with COPD. (C) Scores plot of male smokers *versus* male COPD patients (R^2=0.49, Q^2=0.38, p[CV-ANOVA]=4.0x10[-7]). (D) Loadings of confirmed metabolites that were the most prominent for driving the separation of smokers *versus* male with COPD. *Adapted from [162].*

The observed alterations in the proteome could be linked to dysregulations in specific pathways including protein processing in the ER (Female: FDR=6.5×10^{-5}), xenobiotic metabolism (Male: 2.1×10^{-1}) and oxidative phosphorylation (Female: FDR=2.5×10^{-7}, Male: FDR=5.7×10^{-2}). Pathway enrichment analysis of altered circulating metabolites stratified by gender showed the enhancement of the fatty acid (Female: FDR=1.0×10^{-2}) and sphingolipid pathways (Female: FDR=3.0×10^{-2}) in females, whereas shifts in cAMP signalling (Male: FDR=2.0×10^{-2}) and endocannabinoid (Male: FDR=2.0×10^{-2}) and tryptophan metabolism pathways (Male: FDR=5.0×10^{-2}) in males. The altered metabolic changes based upon the pathway analysis highlight a strong state of oxidative stress in COPD.
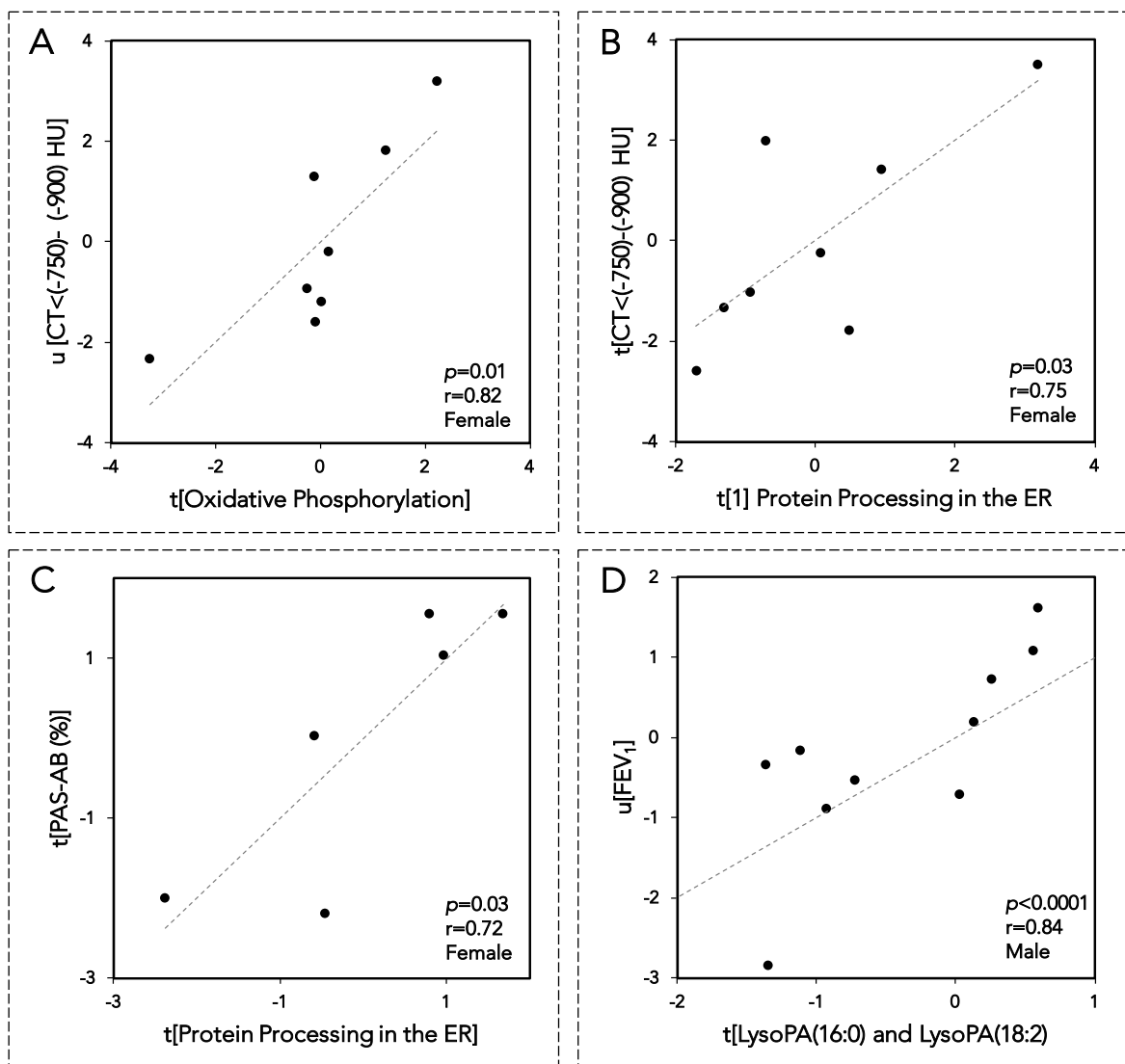
Figure 28| Correlations between the pathway level proteome and metabolome data with clinical parameters. (A) Proteins involved in the oxidative phosphorylation pathway correlated with the lung density range (-750) – (-900) HU in female COPD group (PLS inner relation: r=0.82, p=0.01). The proteins involved in Protein processing in the endoplasmic reticulum (ER) correlated both with the density of the lungs at the CT range (-750) – (-900) HU in female COPD patients (PLS inner relation: r=0.75, p=0.03) (B), and with the percentage of goblet cells in the airway epithelium, quantified by means of AB-PAS staining (C) (PLS inner relation, r=0.72, p=0.03). (D) LysoPA (16:0) and lysoPA (18:2) metabolites correlated with lung function (forced expiratory volume in 1 s (FEV1)) in male COPD patients (r=0.84, p<0.0001). *Panel D adapted from [162].*

Correlation analysis of clinical parameters with revealed pathways in the proteomic approaches and specific metabolites was subsequently performed. Proteins involved in oxidative phosphorylation could be correlated to lung density measured by computed tomography (CT (-750) – (-900) HU) (r=0.82, *p*=0.01) in females (Figure 28A). Furthermore, protein processing in the ER could be correlated to lung density measured by computed tomography in females (r=0.75, *p*=0.03, Figure 28B), as well as goblet cell density (r=0.72, *p*=0.03, Figure 28C). In the metabolic platform, lysophosphatidic acid (lysoPA) (16:0) and lysoPA (18:2) most strongly correlated with FEV1 (%) in male COPD patients (r=0.86, *p*<0.0001, Figure 28D), but not females.

Discussion and reflection

Proteomic and metabolomics investigations of early stage COPD patients in comparison to healthy smokers showed gender specific alterations in both methods primarily driven by the female population in the cohort. Selected marker proteins for classification did not overlap between the female and male comparisons, presenting very distinct gender protein profiles in the airway epithelium in early stage COPD. The metabolomics profiles showed some overlap of driving metabolites between gender. These results confirmed our previous findings of a female-associated molecular COPD sub-phenotype in several compartments, including the alveolar macrophage proteome [163], inflammatory lipid mediators in airway exudates and serum [164], the serum metabolome [162], and the chemokine/cytokine axis [165] in the same cohort.

Alterations in the epithelial proteome and serum metabolome were mainly linked to dysregulation in oxidative stress pathways. In the airway epithelium, xenobiotic metabolism was additionally altered in COPD with differential regulation in women compared to men. The lungs are constantly exposed to harmful substances and in the case of COPD, tobacco smoke that consists of a variety of toxic compounds is a major risk factor. Oxidative stress has been recognized as a predisposing factor in disease pathogenesis [166-168]. Alterations in this pathway in COPD have previously been observed in the proteome of primary epithelial cells from COPD patients [169], as well as in the alveolar macrophages [163] and oxylipins of BAL fluid (BALF) and serum [164] from the same cohort. Constant exposure to ROS and polycyclic aromatic hydrocarbons induced an imbalance between the oxidant and antioxidant, which results in severe tissue damage. Women seem to have a lower tolerance to the accumulation of ROS compared to men. This would furthermore confirm the higher susceptibility to COPD seen in women.

This work presents investigations into early stage COPD using airway epithelial cells and serum. The two sample types give an indication of on the one hand, the first line of defence

against inhaled toxicants (airway epithelial cells) and on the other hand, a potential indication of the molecular downstream processes which are closer to the clinical phenotypes and are seen in the circulating blood (serum). Overall the combination of approaches present a more thorough investigation. 2D-DIGE was able to confirm our female sub phenotype, however the approach was not sensitive enough to grasp the heterogeneity of the male population. The trend seen in all platforms used to study this cohort implies a more homogenous female population than the male subgroup. This might lead, in this context, to less strong classifications and correlations in males. This could result from the sampling and selection of patients included in the study, or alternative, a generally more heterogeneous male COPD pathology, which may result in multiple male sub-phenotypes of COPD.

Concluding Remarks

This doctoral thesis focused on biomarker and pathway discovery in sarcoidosis and COPD as well as the development of a methodology for the identification of HLA-DR bound peptides.

In order to study the HLA-DR peptidome in the human lungs and compare repertoires in heath and disease to reveal triggering antigens, a sensitive approach is needed. Samples from healthy individuals commonly contain 10×10^6 BAL cells, which presents as the limit of feasible sample amounts and the basis for the optimization performed. The developed method enables the identification of HLA-DR bound peptides from scarce clinical material [138] and potentially provides a valuable tool to investigate the repertoires of any number of immune mediated diseases with a given HLA-DR disease predisposition.

Additionally, investigation into Fc-glycans in sarcoidosis confirmed a previously established approach as suitable for the study of antibody glycosylation in serum and BALF [170]. The investigation showed that Fc-glycan abundances correlate between BALF and serum, and provide robust classification of sarcoidosis patients *versus* healthy controls. This work presents a promising angle for the discovery of non-invasive biomarkers in inflammatory lung diseases and reveals serum markers with the potential to determine the inflammatory status in the lungs.

The multi-omics investigation in early stage COPD presented in this work revealed gender specific protein and serum biomarkers in disease, which confirmed previous findings and observed gender differences in the same study cohort [162-165]. Molecular investigations of the epithelial proteome and serum metabolome revealed distinct alterations between early stage COPD patients and smokers with normal lung function, which was mainly driven by the female population in the cohort. The lack of observed alterations in males may be due to a more homogenous sub-phenotype in female COPD patients than in males. Furthermore, these distinct sets of markers could be linked to dysregulation in the maintenance of the anti-oxidative defences. This discovery hints toward oxidative stress as a major molecular mechanism in tissue damage associated with COPD.

Through the application of proteomic and metabolomic approaches and the development of an optimized methodology for immuopeptidomics, discoveries concerning biomarkers as well as insights into the underlying mechanisms were achieved. Taken together, this doctoral thesis contributes a few valuable pieces to the overall puzzle of elucidating the complex disease mechanisms of specific sub-phenotypes in chronic inflammatory lung diseases.

FUTURE PERSPECTIVES

The work presented here that investigated chronic inflammation in sarcoidosis and COPD from an omics perspective, does not terminate with this doctoral thesis. Continuation, as well as new research perspectives emerge that will contribute to the big puzzle of inflammatory mechanisms in the lungs.

In order to narrow down the question of whether antigenic peptides are involved in sarcoidosis disease onset, we finalized the sample collection of a unique sarcoidosis cohort of homozygous HLA-DRB1*15, DRB1*03 and DRB1*04 patients as well as healthy controls. This study represents a first investigation into personalized HLA-DR peptide presentation in sarcoidosis, and has great potential to reveal disease specific immunogenic peptides. The developed approach can be applied to any immune-mediated diseases that strongly associate with HLA alleles, and are likely linked to specific antigens which includes most autoimmune disorders.

The presented investigation of Fc-glycans as inflammatory markers represents a promising platform for biomarker discovery. Given the study limitations in the current study on Fc-glycans in sarcoidosis regarding sex and age matched samples, an expanded study in a larger and well matched study cohort would be necessary. Comparison with other inflammatory disease would also be of interest to confirm disease specificity *versus* alterations due to general inflammation.

The Karolinska COSMIC cohort consists of clinically well-characterized COPD patients and age matched controls (healthy never smokers and smokers with normal lung function). Since COPD represents an etiologically different chronic inflammatory lung disease than sarcoidosis, it would be interesting to apply the methodology developed in the first segment of the thesis to this cohort as well. In particular, the availability of serum and BALF could give interesting insights into the effect of smoking on glycosylation patterns and the link with inflammation. This could reveal novel markers other than CRP, which is currently the only marker used in clinical practice as a general marker for inflammation.

Work on the Karolinska COSMIC cohort in this thesis is part of a large-scale systems biology study designed to investigate molecular alterations in early stage COPD at multiple molecular levels and in multiple anatomical locations, and many platforms have been completed so far. We see similar trends between the platforms, and gender specific alterations in early stage COPD were also prominent in the specimens investigated in other anatomical compartments. The integration of all investigated platforms will be an exciting step in this project and will hopefully bring us one step closer to identifying the pathological

mechanisms in COPD. Recurrent discussions in the literature present evidence for an autoimmune component in certain sub-phenotypes of COPD [171]. Given that the proposed approach to investigate the HLA-DR immunopeptidome is tailored for investigations in the lungs, it would be highly applicable to the Karolinska COSMIC cohort. To my knowledge, no immunopeptidomic study has been previously undertaken to investigate autoimmune components in early stage COPD.

Additionally, a follow up study (COSMIC II) on the same cohort has recently started with reassessment and careful clinical characterization of patients, smokers and healthy individuals as well as sampling of blood and urine. This continuation will mainly focus on metabolomics and lipidomics to study peripheral long term effects of the initial alterations.

SAMMANFATTNING

Sarkoidos och kronisk obstruktiv lungsjukdom (KOL) är två olika kroniska lungsjukdomar, båda utan tydligt klarlagd sjukdomsorsak. Det finns ett starkt behov av diagnostiska markörer och möjlighet att identifiera patologiska förändringar i ett tidigt skede för att begränsa effekterna av sjukdomen. Ännu större är behovet av medicinsk behandling för att kunna bromsa eller helt stoppa den kroniska inflammationen och gradvisa nedbrytningen av lungvävnad.

Sarkoidos karakteriseras av akut inflammation som ger upphov till vävnadsknutor (granulom) i angripna organ och främst i lungorna. Sarkoidos anses vara en autoimmun sjukdom, det vill säga att immunförsvaret vars uppgift är att försvara kroppen mot främmande material och infektioner istället attackerar kroppens egna organ. Det har ännu inte kunnat bevisas vad som initierar den autoimmuna reaktionen i sarkoidos. Kroppens celler tar upp proteiner, (både kroppsegna och främmande) eller hela infektiösa organismer, bryter ned dem i kortare fragment, s k peptider, och presenterar dem på cellytan. Immunförsvarets celler läser av de presenterade molekylerna och avgör om det är ofarliga (kroppsegna) eller främmande (infektion, gifter). Om immuncellerna avgör att peptiderna är främmande startas en immunreaktion för att bekämpa hotet de utgör för kroppen. Den metod vi tagit fram för att identifiera peptider på cellytan möjliggör detektion av skillnader i peptidrepertoaren hos sjuka personer jämfört med friska, och kan därmed ge ledtrådar till vilka ämnen som det autoimmuna försvaret riktas mot.

Sarkoidos kan vara svår att identifiera med dagens medicinska undersökningar, och därför behövs mer specifika diagnostiska alternativ. Diagnostiken ska helst kunna utföras enkelt, till exempel i ett blodprov, och ge information om nivån av pågående inflammation. Vi mätte nivån av immunaktiverande antikroppar i lungan hos sarkoidospatienter jämfört med friska, som också de har antikroppar i blodet från tidigare sjukdomar. Vi kunde identifiera antikroppar med en modifierad sockermolekyl (glykosylering) i både lungvätska och i blod. Denna typ av antikroppar är kopplade till inflammationen vid sarkoidos. Vi kunde visa att kvoten mellan två sorters glykosylering (FA2/FA2G2) hos en specifik sorts antikroppar (IgG_4) kunde särskilja sarkoidospatienters blod från friska individer med 80% säkerhet. Detta innebär att vi har identifierat en markör med potential för diagnostik av kronisk sarkoidos.

Kronisk obstruktiv lungsjukdom är en av de vanligast förekommande lungsjukdomarna globalt, och ca 500 000 - 700 000 svenskar lider av denna sjukdom. Patienterna får vävnadsskador i luftvägarna och lungblåsorna, vilket påverkar och hämmar andningsförmågan. Rökning är en känd riskfaktor för att utveckla KOL, varvid vi riktade in vår

forskning på studier av de tidiga molekylära förändringar som uppstår i lungvävnaden vid rökning, eventuellt redan innan KOL-symptom har uppstått. Kvinnor är överrepresenterade bland KOL-patienter och uppvisar även andra kliniska symptom än män, vilket gjorde att vi valde att specifikt fokusera på könsskillnader i våra KOL-studier.

Proteiner är en stor beståndsdel av alla celler i kroppen, och utgör maskineriet som driver cellens alla funktioner och processer. Slemhinnan i luftvägarna utgörs av s k epitelceller. Dessa celler utgör en viktig del i det första försvaret mot inandad cigarettrök, och är därför av stort intresse vid studier av effekter av rökning. Vi undersökte proteiner från epitelceller från lungorna hos rökare med normal lungfunktion, samt rökare med tidiga stadier av KOL. Dessutom studerade vi metaboliter i blod, och kunde därigenom identifiera skillnader även utanför luftvägarna. Vår forskning visar att de proteiner och metaboliter som påvisas vid KOL skiljer sig åt mellan kvinnor och män. Många skyddande proteiner som uppregleras vid rökning sågs nedregleras hos kvinnliga rökare med KOL, vilket tyder på att ett nedsatt skydd hos kvinnor med KOL. Den här upptäckten kan förklara varför kvinnor och män uppvisar olika symptom vid KOL. De förändringar vi kunde se vid KOL kunde kopplas till cellernas försvar mot fria radikaler som bildas vid rökning; detta leder till något som kallas oxidativ stress. Cellernas förmåga att hantera stressen från de fria radikalerna överbelastas vilket leder till vävnadsskada. Den oxidativa stressen vi identifierat kan alltså kopplas direkt till den vävnadsskada som finns vid KOL.

I min avhandling presenteras forskning om nya metoder för att studera sarkoidos och KOL, med syftet att undersöka den molekylära bakgrunden till sjukdomarna och bidra till framtida diagnostik samt behandlingsmetoder. Vår forskning har påvisat både nya aspekter av sjukdomsbilden och identifierat nya markörer som kan ligga till grund för framtida diagnostikutveckling och läkemedelsforskning, samt bidragit med nya metoder för att studera de molekylära händelseförloppen vid inflammatoriska sjukdomar.

ZUSAMMENFASSUNG

Im Rahmen der vorliegenden Doktorarbeit wurden analytische Untersuchungen zweier chronischer Lungenerkrankungen durchgeführt: Chronisch Obstruktive Lungenerkrankung (COPD) und die Sarkoidose. Beide Erkrankungen weisen sehr unterschiedliche Krankheitsausprägungen auf, deren Auslöser bis dato nicht vollständig bekannt sind. Genetische Veranlagung, Geschlecht oder Belastungen durch die Einwirkung von Umweltfaktoren stellen eine Verbindung mit dem Ausbrechen der Krankheit dar. Die Untersuchung von molekularen Mechanismen, die für diese Krankheiten verantwortlich sind, sind Fokus aktueller Forschung und bedeutend für die Entwicklung pharmazeutischer Wirkstoffe. Für die diagnostische Früherkennung ist zudem die Bestimmung von Biomarkern ausschlaggebend, um Patienten frühmöglichst therapeutisch einzustellen und damit ihr Wohlbefinden zu verbessern.

Sarkoidose ist eine entzündliche Erkrankung, die durch die Bildung von Knötchen (Granulomen) gekennzeichnet ist. Diese können überall im Körper auftreten, jedoch sind die Lymphknoten der Lungen am häufigsten betroffen. Durch das sich bildende Narbengewebe verliert die Lunge an Elastizität und Patienten weisen deutliche Probleme bei der Atmung auf. Es gibt Anzeichen dafür, dass es sich bei der Sarkoidose um eine Autoimmunerkrankung handeln könnte. Autoimmunkrankheiten weisen eine Fehlfunktion des Immunsystems auf, welches beim gesunden Menschen die Bekämpfung von eindringenden fremden Substanzen sicherstellt. Im Falle einer Autoimmunkrankheit wird jedoch gesundes körpereigenes Gewebe bekämpft, wie im Falle von Sarkoidose das Gewebe der Lunge. Zirkulierende Zellen im Körper sind dafür verantwortlich, fremde oder abgestorbene Substanzen im Körper zu eliminieren. Dabei werden diese Substanzen aufgenommen, abgebaut und Bruchstücke (Immunpeptide) auf der Zelloberfläche dem Immunsystem präsentiert. Das Immunsystem erkennt daraufhin diese Strukturen als körpereigen oder fremd und löst im letzteren Fall eine Immunantwort aus. Bei Autoimmunerkrankungen stellen diese Strukturen potentielle Krankheitsauslöser dar, und es gibt Anhaltspunkte dafür, dass solche Strukturen für den Ausbruch von Sarkoidose verantwortlich sind. Die Identifizierung dieser Immunpeptide ist daher von großer Bedeutung bei der Erforschung krankheitsauslösender Mechanismen.

Im Rahmen dieser Arbeit wurde eine sensitive Methode entwickelt, die es ermöglicht, diese Moleküle, die das Immunsystem fehlleiten, in der Lunge einzelner Individuen zu identifizieren. Diese Methode erlaubt es, Profile von Patienten und gesunden Probanden zu erstellen, individuelle Repertoires zu vergleichen und somit zu identifizieren, welche Moleküle die Krankheit potentiell auslösen. Neben der Erforschung von sarkoidose-

auslösenden Substanzen ist die Entwicklung von Markern für die Diagnostik ebenfalls von großer Relevanz. Marker, die mit einem einfachen Bluttest die Erkrankung in der Lunge feststellen könnten, sind dabei von besonderem Interesse. Im Rahmen dieser Arbeit wurden Antikörper in Lungenfluid und Blut analysiert. Dabei wurde der Fokus auf eine spezielle Modifizierung in deren Struktur, die Glycosylierung, gelegt. Diese Modifizierung ist dafür bekannt, dass sie den Entzündungsstatus im Körper wiederspiegelt. Insbesondere das Verhältnis zweier Glykosylierungsformen von Antikörpern in Lungenfluid und Blut hat es uns ermöglicht, Sarkoidose Patienten von gesunden Probanden zu unterscheiden.

Der zweite Teil dieser Doktorarbeit konzentriert sich auf eine weitere chronische Lungenerkrankung. Die COPD ist weltweit die am weitesten verbreitete Lungenerkrankung. Patienten weisen geschädigtes Lungengewebe und damit verbundene Probleme in der Atmung auf. Rauchen stellt einen der größten Risikofaktoren für COPD dar. Das Ziel dieser Studie war es, molekulare Veränderungen durch das Rauchen von Tabak und den damit potentiell verbundenen Krankheitsausbruch zu untersuchen. Frauen weisen im Vergleich zu Männern ein höheres Risiko auf, an COPD zu erkranken. Daher wurde der Fokus speziell auf geschlechtsspezifische Unterschiede gelegt. Es wurden Proteine aus Zellen der Lungenoberflächen (Epithelzellen) und Metabolite im zirkulierenden Blut von rauchenden COPD Patienten analysiert. Dabei zeigte sich, dass Frauen und Männer sehr unterschiedliche Protein- und Metabolitprofile aufweisen, sobald Individuen Tabakrauch ausgesetzt sind und speziell bei Ausbruch der Krankheit. Dies könnte erklären, warum COPD geschlechterspezifische Symptome aufweist. Im Rahmen dieser Arbeit ist es gelungen Gruppen von Proteinen und Metaboliten zu identifizieren, die diese spezifischen molekularen Veränderungen widerspiegeln und Potential als Biomarker aufweisen. Zudem ist es gelungen, die identifizierten Profile bestimmten biologischen Mechanismen zuzuordnen, die vor allem mit oxidativem Stress in Verbindung stehen. Oxidativer Stress beschreibt ein Phänomen, das durch die Entstehung, die erhöhte Bildung bzw. das unzureichende Abfangen reaktiver Sauerstoffspezies gekennzeichnet ist. Oxidativer Stress kann zu schweren Schädigungen führen, wie im Falle der Degenerierung von Lungengewebe bei COPD-Patienten. Des Weiteren, wurde ein Fokus darauf gelegt, die Ergebnisse mit klinischen Untersuchungen in Verbindung zu bringen, um damit die Veränderung klinisch messbar zu machen. Speziell Lungenfunktionstest und Computer-Tomografie konnten einige der beschriebenen Veränderungen feststellen.

Diese Arbeit zeigt etablierte Methoden der Protein- und Metabolitanalytik auf, mit deren Hilfe chronische Lungenerkrankungen untersucht wurden. Zudem wurde eine Methode speziell für die individuelle Bestimmung von Immunpeptiden entwickelt. Diese Ergebnisse liefern Erkenntnisse sowohl über krankheitsspezifische Biomarker als auch Krankheitsmechanismen, die schlussendlich das Potential haben die Diagnostik von Lungenerkrankungen sowie die Entwicklung von Pharmazeutika zu voranzutreiben.

ACKNOWLEDGEMENTS

Many people contributed to this thesis and supported me along the way. I would like to take the opportunity to say thank you to …

All **study subjects** that contribute to research by donating samples. Without you none of this work would have been possible. I also would like to acknowledge **The Swedish Heart Lung Foundation** for the financial support I have received during my doctoral education.

My supervisors for sharing me and for all your guidance throughout my years as your Ph.D. student.

Åsa for your amazing support, for challenging me and making me aim high. I enjoyed being a part of your pulmonomics group, with its good team spirit, fantastic discussions and great science. Thanks for showing me the world of multivariate modelling and making me grow as a scientist.

Johan for being a great and caring boss. I am grateful to you for committing your holly peptide elution to me and for your belief in me that I can make it work. Thanks for everything you taught me, for all your support and understanding when I most needed it.

My co-supervisor **Roman** for giving me the great opportunity to be a part of your group. Thanks for teaching me about mass spectrometry, all the instrumentation time I got and your support in my projects.

Jimmy, my co-supervisor and companion in the peptide elution work. Thanks for your partnership in this challenging project and for everything you taught me about proteins and mass spectrometry.

Maxie, my hands-on teacher in my first year. Thanks for everything I learned from you and all the support in the last years, even from Germany!

Harald, without you I would have never made my way to Karolinska. Thank you for introducing me to the world of mass spectrometry and proteomics over 10 years ago. Danke, dass du mir als Gruppenleiter immer ein Mentor, aber vor allem ein guter Freund warst. Du fehlst uns!

My mentor **Ömer** for your friendship, all your great advice and support during these years.

All my **co-authors** for contributing to my science, for your feedback and great discussions. The **Wheelock group,** especially **Craig** and **Shama** for the great collaboration and for introducing me to metabolomics. **Carina** and **Britt-Marie** from the Proteomics Core Facility in Gothenburg. For the great work and all your support regarding mass spectrometry and

and **Abraham** and all the students we had in the last years for the good atmosphere, lots of fun lunch discussions and nice after works.

My **friends in STHLM** for making it my happy spot: **Haythem & Mustafa** for all the lovely evenings in my "second" living room and that I could always count on you. **Andreas** for fun festival times and all the philosophical discussions. **Franz** for showing me I can do anything, even run a half marathon. **Alva** for your friendship, even though I am from "Bavaria". **Cat** for sharing my addiction to music, all the concerts we went to and always having a song for me. **Ginny** for the most delicious and pretty birthday cakes and being the most fantastic proof reader. **Anna-Luisa & Marianne** for always being up for cultural events and for introducing me to rowing. **Selma & Adam** for the latest and indispensable addition to this category. **Hayley** for all the lovely times with you, **James**, **Harland** and **Archie**. **Maria** for sharing my love for fashion and design. **Sophie** for the good times in Jägergatan, and for convincing me that netball is actually fun. **Andrea** for being my great hipster neighbour. **Susie** for all the nice German lunch breaks.

My **friends** back home: **Juliane**, I am so glad I sat next to you in my first chemistry class almost 13 year ago, that we instantly became such good friends and stuck together over all that time and distance. Thanks to my CBI girls **Uli & Moni** for your great friendship and that we didn't lose each other after university, even though we all live in different countries. **Lotti** for your caring friendship and for all the great trips we made to world metropoles or to the Swedish Fjäll. Thanks to **Babsy,** for forming my career but even more, for your great friendship and all the support you gave me.

Der größte Dank gehört meiner **Familie** für ihre Liebe, die keine Entfernung kennt. **Dirk**, mein großer Bruder, der mir immer beratend zur Seite steht und mir im richtigen Moment einen Schubs gibt. **Julia** & **Lena** für all die schöne Zeit mit euch, in der man selbst den größten Stress vergessen kann. **M+P** für eure Werte, eure unendliche Unterstützung in jeder Lebenslage und dafür, dass ich immer weiß, wo zu Hause ist.

♡

1. Arkema EV, Grunewald J, Kullberg S, Eklund A, Askling J. Sarcoidosis incidence and prevalence: a nationwide register-based assessment in Sweden. Eur Respir J 2016: 48(6): 1690-1699.

2. Iannuzzi MC, Rybicki BA, Teirstein AS. Sarcoidosis. New England Journal of Medicine 2007: 357(21): 2153-2165.

3. Newman KL, Newman LS. Occupational causes of sarcoidosis. Curr Opin Allergy Clin Immunol 2012: 12(2): 145-150.

4. Mannino DM, Buist AS. Global burden of COPD: risk factors, prevalence, and future trends. Lancet 2007: 370(9589): 765-773.

5. Lopez AD, Shibuya K, Rao C, Mathers CD, Hansell AL, Held LS, Schmid V, Buist S. Chronic obstructive pulmonary disease: current burden and future projections. Eur Respir J 2006: 27(2): 397-412.

6. Lange P, Celli B, Agustí A, Boje Jensen G, Divo M, Faner R, Guerra S, Marott JL, Martinez FD, Martinez-Camblor P, Meek P, Owen CA, Petersen H, Pinto-Plata V, Schnohr P, Sood A, Soriano JB, Tesfaigzi Y, Vestbo J. Lung-Function Trajectories Leading to Chronic Obstructive Pulmonary Disease. New England Journal of Medicine 2015: 373(2): 111-122.

7. Lundback B, Lindberg A, Lindstrom M, Ronmark E, Jonsson AC, Jonsson E, Larsson LG, Andersson S, Sandstrom T, Larsson K, Obstructive Lung Disease in Northern Sweden S. Not 15 but 50% of smokers develop COPD?--Report from the Obstructive Lung Disease in Northern Sweden Studies. Respir Med 2003: 97(2): 115-122.

8. Ohar J, Fromer L, Donohue JF. Reconsidering sex-based stereotypes of COPD. Prim Care Respir J 2011: 20(4): 370-378.

9. Abbas AK, Lichtman AH, Pillai S. Basic immunology : functions and disorders of the immune system. Fourth edition. ed.

10. Parkin J, Cohen B. An overview of the immune system. Lancet 2001: 357(9270): 1777-1789.

11. Janeway C. Immunobiology : the immune system in health and disease. 6th ed. Garland Science, New York, 2005.

12. Areschoug T, Gordon S. Pattern recognition receptors and their role in innate immunity: focus on microbial protein ligands. Contrib Microbiol 2008: 15: 45-60.

13. Akira S, Takeda K, Kaisho T. Toll-like receptors: critical proteins linking innate and acquired immunity. Nat Immunol 2001: 2(8): 675-680.

14. O'Neill LA, Golenbock D, Bowie AG. The history of Toll-like receptors - redefining innate immunity. Nat Rev Immunol 2013: 13(6): 453-460.

15. Matzinger P. The danger model: a renewed sense of self. Science 2002: 296(5566): 301-305.

16. Palm NW, Medzhitov R. Pattern recognition receptors and control of adaptive immunity. Immunol Rev 2009: 227(1): 221-233.

17. Chaplin DD. Overview of the immune response. J Allergy Clin Immunol 2010: 125(2 Suppl 2): S3-23.

18. Vivier E, Tomasello E, Baratin M, Walzer T, Ugolini S. Functions of natural killer cells. Nat Immunol 2008: 9(5): 503-510.

19. Mosmann TR, Coffman RL. TH1 and TH2 cells: different patterns of lymphokine secretion lead to different functional properties. Annu Rev Immunol 1989: 7: 145-173.

20. Chen L, Flies DB. Molecular mechanisms of T cell co-stimulation and co-inhibition. Nat Rev Immunol 2013: 13(4): 227-242.

21. Segura E, Amigorena S. Cross-Presentation in Mouse and Human Dendritic Cells. Adv Immunol 2015: 127: 1-31.

22. Dustin ML. The immunological synapse. Cancer Immunol Res 2014: 2(11): 1023-1033.

23. Raphael I, Nalawade S, Eagar TN, Forsthuber TG. T cell subsets and their signature cytokines in autoimmune and inflammatory diseases. Cytokine 2015: 74(1): 5-17.

24. Neefjes J, Jongsma ML, Paul P, Bakke O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. Nat Rev Immunol 2011: 11(12): 823-836.

25. Blum JS, Wearsch PA, Cresswell P. Pathways of antigen processing. Annu Rev Immunol 2013: 31: 443-473.

26. Hillen N, Stevanovic S. Contribution of mass spectrometry-based proteomics to immunology. Expert Rev Proteomics 2006: 3(6): 653-664.

27. Rammensee HG. Chemistry of peptides associated with MHC class I and class II molecules. Curr Opin Immunol 1995: 7(1): 85-96.

28. Garboczi DN, Ghosh P, Utz U, Fan QR, Biddison WE, Wiley DC. Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. Nature 1996: 384(6605): 134-141.

29. O'Brien C, Flower DR, Feighery C. Peptide length significantly influences in vitro affinity for MHC class II molecules. Immunome Res 2008: 4: 6.

30. Goldrath AW, Bevan MJ. Selecting and maintaining a diverse T-cell repertoire. Nature 1999: 402(6759): 255-262.

31. Mosaad YM. Clinical Role of Human Leukocyte Antigen in Health and Disease. Scand J Immunol 2015: 82(4): 283-306.

32. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, Lush MJ, Povey S, Talbot CC, Jr., Wright MW, Wain HM, Trowsdale J, Ziegler A, Beck S. Gene map of the extended human MHC. Nat Rev Genet 2004: 5(12): 889-899.

33. Lefranc M-P, Lefranc G. The immunoglobulin factsbook. Academic Press, 2001.

34. Nimmerjahn F, Ravetch JV. Fc-receptors as regulators of immunity. Adv Immunol 2007: 96: 179-204.

35. Malhotra R, Wormald MR, Rudd PM, Fischer PB, Dwek RA, Sim RB. Glycosylation changes of IgG associated with rheumatoid arthritis can activate complement via the mannose-binding protein. Nat Med 1995: 1(3): 237-243.

36. Parekh RB, Dwek RA, Sutton BJ, Fernandes DL, Leung A, Stanworth D, Rademacher TW, Mizuochi T, Taniguchi T, Matsuta K, et al. Association of rheumatoid arthritis and primary osteoarthritis with changes in the glycosylation pattern of total serum IgG. Nature 1985: 316(6027): 452-457.

37. Scherer HU, van der Woude D, Ioan-Facsinay A, el Bannoudi H, Trouw LA, Wang J, Haupl T, Burmester GR, Deelder AM, Huizinga TW, Wuhrer M, Toes RE. Glycan profiling of anti-citrullinated protein antibodies isolated from human serum and synovial fluid. Arthritis Rheum 2010: 62(6): 1620-1629.

38. Bond A, Alavi A, Axford JS, Bourke BE, Bruckner FE, Kerr MA, Maxwell JD, Tweed KJ, Weldon MJ, Youinou P, Hay FC. A detailed lectin analysis of IgG glycosylation, demonstrating disease specific changes in terminal galactose and N-acetylglucosamine. J Autoimmun 1997: 10(1): 77-85.

39. Kaneko Y, Nimmerjahn F, Ravetch JV. Anti-inflammatory activity of immunoglobulin G resulting from Fc sialylation. Science 2006: 313(5787): 670-673.

40. Anthony RM, Wermeling F, Ravetch JV. Novel roles for the IgG Fc glycan. Ann N Y Acad Sci 2012: 1253: 170-180.

41. Anthony RM, Wermeling F, Karlsson MC, Ravetch JV. Identification of a receptor required for the anti-inflammatory activity of IVIG. Proc Natl Acad Sci U S A 2008: 105(50): 19571-19578.

42. de Jong SE, Selman MH, Adegnika AA, Amoah AS, van Riet E, Kruize YC, Raynes JG, Rodriguez A, Boakye D, von Mutius E, Knulst AC, Genuneit J, Cooper PJ, Hokke CH, Wuhrer M, Yazdanbakhsh M. IgG1 Fc N-glycan galactosylation as a biomarker for immune activation. Sci Rep 2016: 6: 28207.

43. Royle L, Campbell MP, Radcliffe CM, White DM, Harvey DJ, Abrahams JL, Kim YG, Henry GW, Shadick NA, Weinblatt ME, Lee DM, Rudd PM, Dwek RA. HPLC-based analysis of serum N-glycans on a 96-well plate platform with dedicated database software. Anal Biochem 2008: 376(1): 1-12.

44. Lundstrom SL, Fernandes-Cerqueira C, Ytterberg AJ, Ossipova E, Hensvold AH, Jakobsson PJ, Malmstrom V, Catrina AI, Klareskog L, Lundberg K, Zubarev RA. IgG antibodies to cyclic citrullinated peptides exhibit profiles specific in terms of IgG subclasses, Fc-glycans and a fab-Peptide sequence. PLoS One 2014: 9(11): e113924.

45. Bell E, Bird L. Autoimmunity. Nature 2005: 435(7042): 583-583.

46. Autoimmune disease. Nature biotechnology 2000: 18 Suppl: IT7-9.

47. Fridkis-Hareli M. Design of peptide immunotherapies for MHC Class-II-associated autoimmune disorders. Clin Dev Immunol 2013: 2013: 826191.

48. Fairweather D, Frisancho-Kiss S, Rose NR. Sex differences in autoimmune disease from a pathological perspective. Am J Pathol 2008: 173(3): 600-609.

49. Albert LJ, Inman RD. Molecular mimicry and autoimmunity. N Engl J Med 1999: 341(27): 2068-2074.

50. Goris A, Liston A. The immunogenetic architecture of autoimmune disease. Cold Spring Harb Perspect Biol 2012: 4(3).

51. Thorsby E, Lie BA. HLA associated genetic predisposition to autoimmune diseases: Genes involved and possible mechanisms. Transpl Immunol 2005: 14(3-4): 175-182.

52. Bogdanos DP, Smyk DS, Rigopoulou EI, Mytilinaiou MG, Heneghan MA, Selmi C, Gershwin ME. Twin studies in autoimmune disease: genetics, gender and environment. J Autoimmun 2012: 38(2-3): J156-169.

53. Leslie RD, Hawa M. Twin studies in auto-immune disease. Acta Genet Med Gemellol (Roma) 1994: 43(1-2): 71-81.

54. Kochi Y. Genetics of autoimmune diseases: perspectives from genome-wide association studies. Int Immunol 2016: 28(4): 155-161.

55. Lettre G, Rioux JD. Autoimmune diseases: insights from genome-wide association studies. Hum Mol Genet 2008: 17(R2): R116-121.

56. Miller FW. Environmental agents and autoimmune diseases. Adv Exp Med Biol 2011: 711: 61-81.

57. Klein SL, Flanagan KL. Sex differences in immune responses. Nat Rev Immunol 2016: 16(10): 626-638.

58. Klareskog L, Stolt P, Lundberg K, Kallberg H, Bengtsson C, Grunewald J, Ronnelid J, Harris HE, Ulfgren AK, Rantapaa-Dahlqvist S, Eklund A, Padyukov L, Alfredsson L. A new model for an etiology of rheumatoid arthritis: smoking may trigger HLA-DR (shared epitope)-restricted immune reactions to autoantigens modified by citrullination. Arthritis Rheum 2006: 54(1): 38-46.

59. Nielsen PR, Kragstrup TW, Deleuran BW, Benros ME. Infections as risk factor for autoimmune diseases - A nationwide study. J Autoimmun 2016: 74: 176-181.

60. Ward JPT, Ward J, Leach RM, Wiener CM, ebrary Inc. The respiratory system at a glance. At a glance. 3rd ed. Wiley-Blackwell,, Chichester U.K., 2010; p. 125 p.

61. Gibson GJ. Respiratory Medicine. Saunders, 2003.

62. Knight DA, Holgate ST. The airway epithelium: structural and functional properties in health and disease. Respirology 2003: 8(4): 432-446.

63. Crystal RG, Randell SH, Engelhardt JF, Voynow J, Sunday ME. Airway epithelial cells: current concepts and challenges. Proc Am Thorac Soc 2008: 5(7): 772-777.

64. Paulsson M. Basement membrane proteins: structure, assembly, and cellular interactions. Crit Rev Biochem Mol Biol 1992: 27(1-2): 93-127.

65. LeBleu VS, Macdonald B, Kalluri R. Structure and function of basement membranes. Exp Biol Med (Maywood) 2007: 232(9): 1121-1129.

66.   Diamond G, Legarda D, Ryan LK. The innate immune response of the respiratory epithelium. Immunol Rev 2000: 173: 27-38.

67.   Hiemstra PS. Epithelial antimicrobial peptides and proteins: their role in host defence and inflammation. Paediatr Respir Rev 2001: 2(4): 306-310.

68.   Croom E. Metabolism of xenobiotics of human environments. Prog Mol Biol Transl Sci 2012: 112: 31-88.

69.   How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General. How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General, Atlanta (GA), 2010.

70.   Shimada T, Fujii-Kuriyama Y. Metabolic activation of polycyclic aromatic hydrocarbons to carcinogens by cytochromes P450 1A1 and 1B1. Cancer Sci 2004: 95(1): 1-6.

71.   Casarett LJ, Klaassen CD, Watkins JB. Casarett and Doull's essentials of toxicology. McGraw-Hill/Medical Pub. Div., New York, 2003.

72.   Baughman RP, Teirstein AS, Judson MA, Rossman MD, Yeager H, Jr., Bresnitz EA, DePalo L, Hunninghake G, Iannuzzi MC, Johns CJ, McLennan G, Moller DR, Newman LS, Rabin DL, Rose C, Rybicki B, Weinberger SE, Terrin ML, Knatterud GL, Cherniak R, Case Control Etiologic Study of Sarcoidosis research g. Clinical characteristics of patients in a case control study of sarcoidosis. Am J Respir Crit Care Med 2001: 164(10 Pt 1): 1885-1889.

73.   Hillerdal G, Nou E, Osterman K, Schmekel B. Sarcoidosis: epidemiology and prognosis. A 15-year European study. Am Rev Respir Dis 1984: 130(1): 29-32.

74.   Grunewald J. HLA associations and Lofgren's syndrome. Expert Rev Clin Immunol 2012: 8(1): 55-62.

75.   Grunewald J, Eklund A. Lofgren's syndrome: human leukocyte antigen strongly influences the disease course. Am J Respir Crit Care Med 2009: 179(4): 307-312.

76.   Lofgren S, Lundback H. The bilateral hilar lymphoma syndrome; a study of the relation to age and sex in 212 cases. Acta Med Scand 1952: 142(4): 259-264.

77.   Grunewald J, Janson CH, Eklund A, Ohrn M, Olerup O, Persson U, Wigzell H. Restricted V alpha 2.3 gene usage by CD4+ T lymphocytes in bronchoalveolar lavage fluid from sarcoidosis patients correlates with HLA-DR3. Eur J Immunol 1992: 22(1): 129-135.

78.   Ahlgren KM, Ruckdeschel T, Eklund A, Wahlstrom J, Grunewald J. T cell receptor-Vbeta repertoires in lung and blood CD4+ and CD8+ T cells of pulmonary sarcoidosis patients. BMC Pulm Med 2014: 14: 50.

79.   Grunewald J, Kaiser Y, Ostadkarampour M, Rivera NV, Vezzi F, Lotstedt B, Olsen RA, Sylwan L, Lundin S, Kaller M, Sandalova T, Ahlgren KM, Wahlstrom J, Achour A, Ronninger M, Eklund A. T-cell receptor-HLA-DRB1 associations suggest specific antigens in pulmonary sarcoidosis. Eur Respir J 2016: 47(3): 898-909.

80. Berlin M, Fogdell-Hahn A, Olerup O, Eklund A, Grunewald J. HLA-DR predicts the prognosis in Scandinavian patients with pulmonary sarcoidosis. Am J Respir Crit Care Med 1997: 156(5): 1601-1605.

81. Rutherford RM, Brutsche MH, Kearns M, Bourke M, Stevens F, Gilmartin JJ. HLA-DR2 predicts susceptibility and disease chronicity in Irish sarcoidosis patients. Sarcoidosis Vasc Diffuse Lung Dis 2004: 21(3): 191-198.

82. Darlington P, Tallstedt L, Padyukov L, Kockum I, Cederlund K, Eklund A, Grunewald J. HLA-DRB1* alleles and symptoms associated with Heerfordt's syndrome in sarcoidosis. Eur Respir J 2011: 38(5): 1151-1157.

83. Rivera NV, Ronninger M, Shchetynsky K, Franke A, Nothen MM, Muller-Quernheim J, Schreiber S, Adrianto I, Karakaya B, van Moorsel CH, Navratilova Z, Kolek V, Rybicki BA, Iannuzzi MC, Petrek M, Grutters JC, Montgomery C, Fischer A, Eklund A, Padyukov L, Grunewald J. High-Density Genetic Mapping Identifies New Susceptibility Variants in Sarcoidosis Phenotypes and Shows Genomic-driven Phenotypic Differences. Am J Respir Crit Care Med 2016: 193(9): 1008-1022.

84. Costabel U. CD4/CD8 ratios in bronchoalveolar lavage fluid: of value for diagnosing sarcoidosis? Eur Respir J 1997: 10(12): 2699-2700.

85. Selroos O, Eklund A, Association SS. Sarkoidos. Studentlitteratur [i samarbete med] Scandinavian Sarcoidosis Association (SSA), 2005.

86. Pauwels RA, Rabe KF. Burden and clinical features of chronic obstructive pulmonary disease (COPD). Lancet 2004: 364(9434): 613-620.

87. Kurt OK, Zhang J, Pinkerton KE. Pulmonary health effects of air pollution. Curr Opin Pulm Med 2016: 22(2): 138-143.

88. To T, Zhu J, Larsen K, Simatovic J, Feldman L, Ryckman K, Gershon A, Lougheed MD, Licskai C, Chen H, Villeneuve PJ, Crighton E, Su Y, Sadatsafavi M, Williams D, Carlsten C, Canadian Respiratory Research N. Progression from Asthma to Chronic Obstructive Pulmonary Disease. Is Air Pollution a Risk Factor? Am J Respir Crit Care Med 2016: 194(4): 429-438.

89. Gan WQ, Man SF, Postma DS, Camp P, Sin DD. Female smokers beyond the perimenopausal period are at increased risk of chronic obstructive pulmonary disease: a systematic review and meta-analysis. Respir Res 2006: 7: 52.

90. Mannino DM, Homa DM, Akinbami LJ, Ford ES, Redd SC. Chronic obstructive pulmonary disease surveillance--United States, 1971-2000. MMWR Surveill Summ 2002: 51(6): 1-16.

91. Camp PG, Coxson HO, Levy RD, Pillai SG, Anderson W, Vestbo J, Kennedy SM, Silverman EK, Lomas DA, Pare PD. Sex differences in emphysema and airway disease in smokers. Chest 2009: 136(6): 1480-1488.

92. Dransfield MT, Davis JJ, Gerald LB, Bailey WC. Racial and gender differences in susceptibility to tobacco smoke among patients with chronic obstructive pulmonary disease. Respir Med 2006: 100(6): 1110-1116.

93. Grydeland TB, Dirksen A, Coxson HO, Pillai SG, Sharma S, Eide GE, Gulsvik A, Bakke PS. Quantitative computed tomography: emphysema and airway wall thickness by sex, age and smoking. Eur Respir J 2009: 34(4): 858-865.

94. Townsend EA, Miller VM, Prakash YS. Sex differences and sex steroids in lung health and disease. Endocr Rev 2012: 33(1): 1-47.

95. Triebner K, Matulonga B, Johannessen A, Suske S, Benediktsdottir B, Demoly P, Dharmage SC, Franklin KA, Garcia-Aymerich J, Gullon Blanco JA, Heinrich J, Holm M, Jarvis D, Jogi R, Lindberg E, Moratalla Rovira JM, Muniozguren Agirre N, Pin I, Probst-Hensch N, Puggini L, Raherison C, Sanchez-Ramos JL, Schlunssen V, Sunyer J, Svanes C, Hustad S, Leynaert B, Gomez Real F. Menopause Is Associated with Accelerated Lung Function Decline. Am J Respir Crit Care Med 2017: 195(8): 1058-1065.

96. Vogelmeier CF, Criner GJ, Martinez FJ, Anzueto A, Barnes PJ, Bourbeau J, Celli BR, Chen R, Decramer M, Fabbri LM, Frith P, Halpin DM, Lopez Varela MV, Nishimura M, Roche N, Rodriguez-Roisin R, Sin DD, Singh D, Stockley R, Vestbo J, Wedzicha JA, Agusti A. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report. GOLD Executive Summary. Am J Respir Crit Care Med 2017: 195(5): 557-582.

97. Vogelmeier CF, Criner GJ, Martinez FJ, Anzueto A, Barnes PJ, Bourbeau J, Celli BR, Chen R, Decramer M, Fabbri LM, Frith P, Halpin DMG, López Varela MV, Nishimura M, Roche N, Rodriguez-Roisin R, Sin DD, Singh D, Stockley R, Vestbo J, Wedzicha JA, Agusti A. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report: GOLD Executive Summary. European Respiratory Journal 2017.

98. Walhout M, Vidal M, Dekker J. Handbook of Systems Biology: Concepts and Insights. Elsevier Science, 2012.

99. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet 2015: 16(2): 85-97.

100. Friedman DB, Hoving S, Westermeier R. Isoelectric focusing and two-dimensional gel electrophoresis. Methods Enzymol 2009: 463: 515-540.

101. Beck-Sickinger A, Solodkoff C, Lottspeich F, Zorbas H, Domdey H, Eckerskorn C, Engelhardt H, Ficner R, Görg A, Haberhausen G. Bioanalytik. Spektrum Akademischer Verlag, 1998.

102. Gorg A, Drews O, Luck C, Weiland F, Weiss W. 2-DE with IPGs. Electrophoresis 2009: 30 Suppl 1: S122-132.

103. Unlu M, Morgan ME, Minden JS. Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. Electrophoresis 1997: 18(11): 2071-2077.

104. Cramer R, Westermeier R. Difference Gel Electrophoresis (DIGE): Methods and Protocols. Humana Press, 2012.

105. Sandberg A, Buschmann V, Kapusta P, Erdmann R, Wheelock AM. Use of Time-Resolved Fluorescence To Improve Sensitivity and Dynamic Range of Gel-Based Proteomics. Anal Chem 2016: 88(6): 3067-3074.
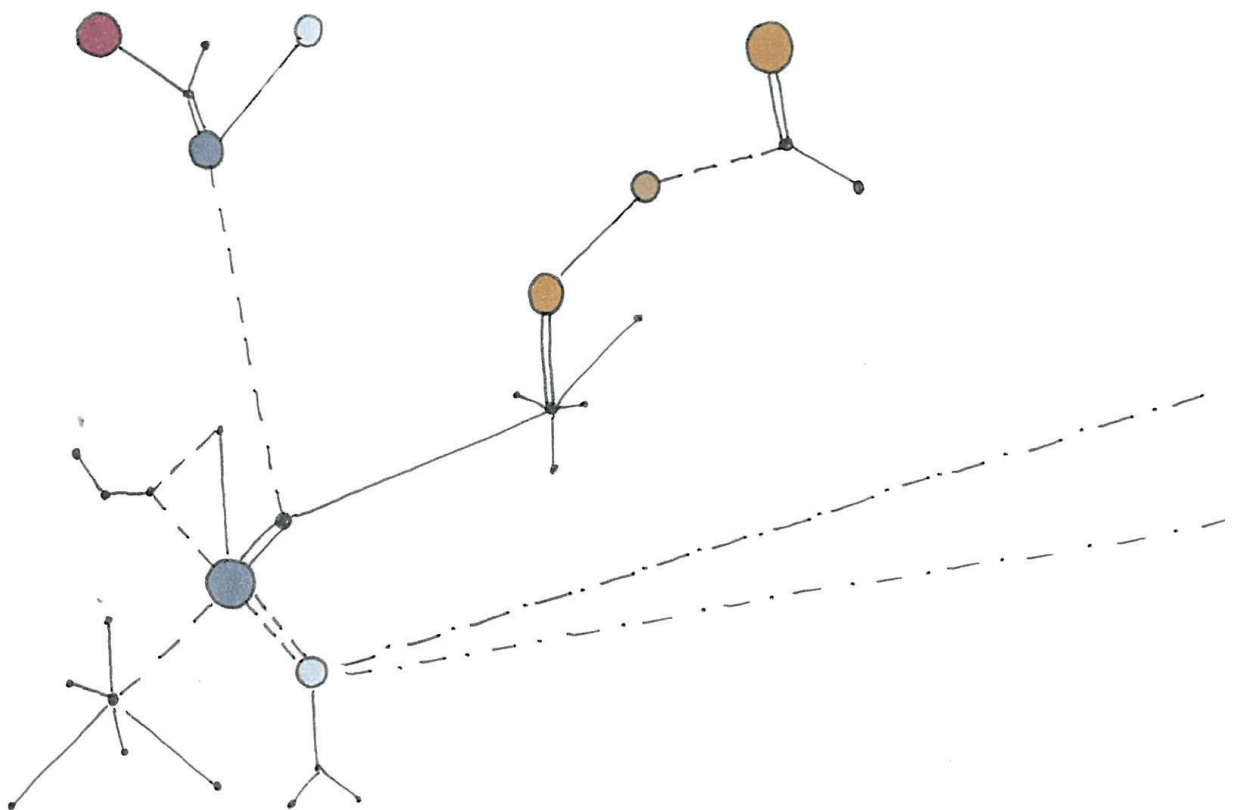
106. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003: 422(6928): 198-207.

107. Olsen JV, Ong SE, Mann M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. Mol Cell Proteomics 2004: 3(6): 608-614.

108. Wang YC, Peterson SE, Loring JF. Protein post-translational modifications and regulation of pluripotency in human stem cells. Cell Res 2014: 24(2): 143-160.

109. Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol 2012: 13(4): 263-269.

110. Buscher JM, Czernik D, Ewald JC, Sauer U, Zamboni N. Cross-platform comparison of methods for quantitative metabolomics of primary metabolism. Anal Chem 2009: 81(6): 2135-2143.

111. Yin P, Xu G. Current state-of-the-art of nontargeted metabolomics based on liquid chromatography-mass spectrometry with special emphasis in clinical applications. J Chromatogr A 2014: 1374: 1-13.

112. Alpert AJ. Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds. J Chromatogr 1990: 499: 177-196.

113. Hemstrom P, Irgum K. Hydrophilic interaction chromatography. J Sep Sci 2006: 29(12): 1784-1821.

114. Lane CS. Mass spectrometry-based proteomics in the life sciences. Cell Mol Life Sci 2005: 62(7-8): 848-869.

115. Han X, Aslanian A, Yates JR, 3rd. Mass spectrometry for proteomics. Curr Opin Chem Biol 2008: 12(5): 483-490.

116. Makarov A. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. Anal Chem 2000: 72(6): 1156-1162.

117. Bantscheff M, Lemeer S, Savitski MM, Kuster B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. Anal Bioanal Chem 2012: 404(4): 939-965.

118. Roepstorff P, Fohlman J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. Biomed Mass Spectrom 1984: 11(11): 601.

119. Ong SE, Mann M. Mass spectrometry-based proteomics turns quantitative. Nat Chem Biol 2005: 1(5): 252-262.

120. Shaw J, Rowlinson R, Nickson J, Stone T, Sweet A, Williams K, Tonge R. Evaluation of saturation labelling two-dimensional difference gel electrophoresis fluorescent dyes. Proteomics 2003: 3(7): 1181-1195.

121. Coombs KM. Quantitative proteomics of complex mixtures. Expert Rev Proteomics 2011: 8(5): 659-677.

122. Wheelock AM, Wheelock CE. Trials and tribulations of 'omics data analysis: assessing quality of SIMCA-based multivariate models using examples from pulmonary medicine. Mol Biosyst 2013: 9(11): 2589-2596.

123. Jackson JE. A user's guide to principal components. Wiley, New York, 1991.

124. Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). Journal of Chemometrics 2002: 16(3): 119-128.

125. Heyder T, Yang M, Kohler M, Merikallio H, Forsslund H, Li C, Karimi R, Eklund A, Grunewald J, Kaarteenaho R, Sihlbom C, Sköld C, Wheelock Å. Gender-enhanced alteration in airway epithelial proteome in COPD related to xenobiotic metabolism Manuscript.

126. Eriksson L, Trygg J, Wold S. CV-ANOVA for significance testing of PLS and OPLS® models. Journal of Chemometrics 2008: 22(11-12): 594-600.

127. Sandberg A, Skold CM, Grunewald J, Eklund A, Wheelock AM. Assessing recent smoking status by measuring exhaled carbon monoxide levels. PLoS One 2011: 6(12): e28864.

128. Eklund A, Blaschke E. Relationship between changed alveolar-capillary permeability and angiotensin converting enzyme activity in serum in sarcoidosis. Thorax 1986: 41(8): 629-634.

129. Karimi R, Tornling G, Grunewald J, Eklund A, Skold CM. Cell recovery in bronchoalveolar lavage fluid in smokers is dependent on cumulative smoking history. PLoS One 2012: 7(3): e34232.

130. Ranu H, Wilde M, Madden B. Pulmonary function tests. Ulster Med J 2011: 80(2): 84-90.

131. Aaron SD, Dales RE, Cardinal P. How accurate is spirometry at predicting restrictive pulmonary impairment? Chest 1999: 115(3): 869-873.

132. Swanney MP, Ruppel G, Enright PL, Pedersen OF, Crapo RO, Miller MR, Jensen RL, Falaschetti E, Schouten JP, Hankinson JL, Stocks J, Quanjer PH. Using the lower limit of normal for the FEV1/FVC ratio reduces the misclassification of airway obstruction. Thorax 2008: 63(12): 1046-1051.

133. Johns DP, Walters JA, Walters EH. Diagnosis and early detection of COPD using spirometry. J Thorac Dis 2014: 6(11): 1557-1569.

134. Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects. Mol Cell Proteomics 2002: 1(11): 845-867.

135. Issaq HJ, Veenstra TD. Proteomic and Metabolomic Approaches to Biomarker Discovery. Elsevier Science & Technology Books, Amsterdam; Boston; Elsevier, 2013.

136. Zubarev RA. The challenge of the proteome dynamic range and its implications for in-depth proteomics. Proteomics 2013: 13(5): 723-726.

137. Lampson LA, Levy R. Two populations of Ia-like molecules on a human B cell line. J Immunol 1980: 125(1): 293-299.

138. Heyder T, Kohler M, Tarasova NK, Haag S, Rutishauser D, Rivera NV, Sandin C, Mia S, Malmstrom V, Wheelock AM, Wahlstrom J, Holmdahl R, Eklund A, Zubarev RA, Grunewald J, Ytterberg AJ. Approach for Identifying Human Leukocyte Antigen (HLA)-DR Bound Peptides from Scarce Clinical Samples. Mol Cell Proteomics 2016: 15(9): 3017-3029.

139. Lundstrom SL, Yang H, Lyutvinskiy Y, Rutishauser D, Herukka SK, Soininen H, Zubarev RA. Blood plasma IgG Fc glycans are significantly altered in Alzheimer's disease and progressive mild cognitive impairment. J Alzheimers Dis 2014: 38(3): 567-579.

140. Silva E, O'Gorman M, Becker S, Auer G, Eklund A, Grunewald J, Wheelock AM. In the eye of the beholder: does the master see the SameSpots as the novice? J Proteome Res 2010: 9(3): 1522-1532.

141. Albrecht D, Kniemeyer O, Brakhage AA, Guthke R. Missing values in gel-based proteomics. Proteomics 2010: 10(6): 1202-1211.

142. Pedreschi R, Hertog ML, Carpentier SC, Lammertyn J, Robben J, Noben JP, Panis B, Swennen R, Nicolai BM. Treatment of missing values for multivariate statistical analysis of gel-based proteomics data. Proteomics 2008: 8(7): 1371-1383.

143. Lin D, Zhang J, Li J, Xu C, Deng HW, Wang YP. An integrative imputation method based on multi-omics datasets. BMC Bioinformatics 2016: 17: 247.

144. Gromski PS, Xu Y, Kotze HL, Correa E, Ellis DI, Armitage EG, Turner ML, Goodacre R. Influence of missing values substitutes on multivariate analysis of metabolomics data. Metabolites 2014: 4(2): 433-452.

145. Duan X, Young R, Straubinger RM, Page B, Cao J, Wang H, Yu H, Canty JM, Qu J. A straightforward and highly efficient precipitation/on-pellet digestion procedure coupled with a long gradient nano-LC separation and Orbitrap mass spectrometry for label-free expression profiling of the swine heart mitochondrial proteome. J Proteome Res 2009: 8(6): 2838-2850.

146. Elliott MH, Smith DS, Parker CE, Borchers C. Current trends in quantitative proteomics. J Mass Spectrom 2009: 44(12): 1637-1660.

147. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) 1995: 57(1): 289-300.

148. Storey JD. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2002: 64(3): 479-498.

149. Adamopoulou E, Tenzer S, Hillen N, Klug P, Rota IA, Tietz S, Gebhardt M, Stevanovic S, Schild H, Tolosa E, Melms A, Stoeckle C. Exploring the MHC-peptide matrix of central tolerance in the human thymus. Nat Commun 2013: 4: 2039.

150. Collado JA, Alvarez I, Ciudad MT, Espinosa G, Canals F, Pujol-Borrell R, Carrascal M, Abian J, Jaraquemada D. Composition of the HLA-DR-associated human thymus peptidome. Eur J Immunol 2013: 43(9): 2273-2282.

151. Seward RJ, Drouin EE, Steere AC, Costello CE. Peptides presented by HLA-DR molecules in synovia of patients with rheumatoid arthritis or antibiotic-refractory Lyme arthritis. Mol Cell Proteomics 2011: 10(3): M110 002477.

152. Fissolo N, Haag S, de Graaf KL, Drews O, Stevanovic S, Rammensee HG, Weissert R. Naturally presented peptides on major histocompatibility complex I and II molecules eluted from central nervous system of multiple sclerosis patients. Mol Cell Proteomics 2009: 8(9): 2090-2101.

153. Wahlstrom J, Dengjel J, Persson B, Duyar H, Rammensee HG, Stevanovic S, Eklund A, Weissert R, Grunewald J. Identification of HLA-DR-bound peptides presented by human bronchoalveolar lavage cells in sarcoidosis. J Clin Invest 2007: 117(11): 3576-3582.

154. Strug I, Calvo-Calle JM, Green KM, Cruz J, Ennis FA, Evans JE, Stern LJ. Vaccinia peptides eluted from HLA-DR1 isolated from virus-infected cells are recognized by CD4+ T cells from a vaccinated donor. J Proteome Res 2008: 7(7): 2703-2711.

155. Andreatta M, Schafer-Nielsen C, Lund O, Buus S, Nielsen M. NNAlign: a web-based prediction method allowing non-expert end-user discovery of sequence motifs in quantitative peptide data. PLoS One 2011: 6(11): e26781.

156. Rapin N, Hoof I, Lund O, Nielsen M. The MHC motif viewer: a visualization tool for MHC binding motifs. Curr Protoc Immunol 2010: Chapter 18: Unit 18 17.

157. Pucic M, Knezevic A, Vidic J, Adamczyk B, Novokmet M, Polasek O, Gornik O, Supraha-Goreta S, Wormald MR, Redzic I, Campbell H, Wright A, Hastie ND, Wilson JF, Rudan I, Wuhrer M, Rudd PM, Josic D, Lauc G. High throughput isolation and glycosylation analysis of IgG-variability and heritability of the IgG glycome in three isolated human populations. Mol Cell Proteomics 2011: 10(10): M111 010090.

158. Ercan A, Cui J, Chatterton DE, Deane KD, Hazen MM, Brintnell W, O'Donnell CI, Derber LA, Weinblatt ME, Shadick NA, Bell DA, Cairns E, Solomon DH, Holers VM, Rudd PM, Lee DM. Aberrant IgG galactosylation precedes disease onset, correlates with disease activity, and is prevalent in autoantibodies in rheumatoid arthritis. Arthritis Rheum 2010: 62(8): 2239-2248.

159. Nakagawa H, Hato M, Takegawa Y, Deguchi K, Ito H, Takahata M, Iwasaki N, Minami A, Nishimura S. Detection of altered N-glycan profiles in whole serum from rheumatoid arthritis patients. J Chromatogr B Analyt Technol Biomed Life Sci 2007: 853(1-2): 133-137.

160. Pasek M, Duk M, Podbielska M, Sokolik R, Szechinski J, Lisowska E, Krotkiewski H. Galactosylation of IgG from rheumatoid arthritis (RA) patients--changes during therapy. Glycoconj J 2006: 23(7-8): 463-471.

161. van de Geijn FE, Wuhrer M, Selman MH, Willemsen SP, de Man YA, Deelder AM, Hazes JM, Dolhain RJ. Immunoglobulin G galactosylation and sialylation are associated with pregnancy-induced improvement of rheumatoid arthritis and the postpartum flare: results from a large prospective cohort study. Arthritis Res Ther 2009: 11(6): R193.

162. Naz S, Kolmert J, Yang M, Reinke SN, Kamleh MA, Snowden S, Heyder T, Levanen B, Erle DJ, Skold CM, Wheelock AM, Wheelock CE. Metabolomics analysis identifies sex-associated metabotypes of oxidative stress and the autotaxin-lysoPA axis in COPD. Eur Respir J 2017.

163. Kohler M, Sandberg A, Kjellqvist S, Thomas A, Karimi R, Nyren S, Eklund A, Thevis M, Skold CM, Wheelock AM. Gender differences in the bronchoalveolar lavage cell proteome of patients with chronic obstructive pulmonary disease. J Allergy Clin Immunol 2013: 131(3): 743-751.

164. Balgoma D, Yang M, Sjodin M, Snowden S, Karimi R, Levanen B, Merikallio H, Kaarteenaho R, Palmberg L, Larsson K, Erle DJ, Dahlen SE, Dahlen B, Skold CM, Wheelock AM, Wheelock CE. Linoleic acid-derived lipid mediators increase in a female-dominated subphenotype of COPD. Eur Respir J 2016: 47(6): 1645-1656.

165. Forsslund H, Yang M, Mikko M, Karimi R, Nyren S, Engvall B, Grunewald J, Merikallio H, Kaarteenaho R, Wahlstrom J, Wheelock AM, Skold CM. Gender differences in the T-cell profiles of the airways in COPD patients associated with clinical phenotypes. Int J Chron Obstruct Pulmon Dis 2017: 12: 35-48.

166. Kirkham PA, Barnes PJ. Oxidative stress in COPD. Chest 2013: 144(1): 266-273.

167. Tomaki M, Sugiura H, Koarai A, Komaki Y, Akita T, Matsumoto T, Nakanishi A, Ogawa H, Hattori T, Ichinose M. Decreased expression of antioxidant enzymes and increased expression of chemokines in COPD lung. Pulm Pharmacol Ther 2007: 20(5): 596-605.

168. Rahman I. Antioxidant therapies in COPD. Int J Chron Obstruct Pulmon Dis 2006: 1(1): 15-29.

169. Hoffmann RF, Zarrintan S, Brandenburg SM, Kol A, de Bruin HG, Jafari S, Dijk F, Kalicharan D, Kelders M, Gosker HR, Ten Hacken NH, van der Want JJ, van Oosterhout AJ, Heijink IH. Prolonged cigarette smoke exposure alters mitochondrial structure and function in airway epithelial cells. Respir Res 2013: 14: 97.

170. Lundstrom SL, Zhang B, Rutishauser D, Aarsland D, Zubarev RA. SpotLight Proteomics: uncovering the hidden blood proteome improves diagnostic power of proteomics. Sci Rep 2017: 7: 41929.

171. Cosio MG, Saetta M, Agusti A. Immunologic aspects of chronic obstructive pulmonary disease. N Engl J Med 2009: 360(23): 2445-2454.

Karolinska
Institutet