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

SYSTEMS BIOLOGY APPROACHES TO INVESTIGATE MECHANISMS OF OBSTRUCTIVE LUNG DISEASES

Marika Ström

Stockholm 2022

All previously published papers were reproduced with permission from the publisher.
Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2022

ISBN 978-91-8016-739-0
Cover illustration by Johanna Ström using the website https://openai.com/

Systems biology approaches to investigate mechanisms of obstructive lung diseases

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Marika Ström

The thesis will be defended in public at Petrén, Nobels väg 12B, Karolinska Institutet, Solna Campus, Stockholm, Sweden, on Friday October 14, 2022 at 9:00

Principal Supervisor:
Docent Åsa Wheelock
Karolinska Institutet
Department of Medicine Solna
Center for Molecular Medicine
Respiratory Medicine Unit

Karolinska University Hospital
Department of Respiratory Medicine
and Allergy

Co-supervisor(s):
Professor/överläkare Magnus Sköld
Karolinska Institutet
Department of Medicine Solna
Center for Molecular Medicine
Respiratory Medicine Unit

Karolinska University Hospital
Department of Respiratory Medicine
and Allergy

Professor Susanne Gabrielsson
Karolinska Institutet
Department of Medicine Solna
Division of Immunology and Allergy

Ph.D./överläkare Eva Berggren-Broström
Karolinska Institutet
Södersjukhuset
Department of Clinical Science and Education

Sach's Children's and Youth Hospital
Department of Pediatrics

Opponent:
Lector Malin Linder Nording
Umeå University
Department of Chemistry

Examination Board:
Docent Mikael Adner
Karolinska Institutet
Institute of Environmental Medicine
Centre for Allergy Research

Professor Peter Nilsson
KTH Royal Institute of Technology
SciLifeLab
Department of Protein Science
Division of Affinity Proteomics

Docent Ning Xu Landén
Karolinska Institutet
Department of Medicine Solna
Dermatology and Venereology Division

Karolinska Institutet
Stockholm node
Ming Wai Lau Centre for Reparative Medicine

To my mother Britt-Marie

POPULAR SCIENCE SUMMARY OF THE THESIS

Asthma and chronic obstructive pulmonary disease (COPD) are obstructive lung diseases that cause shortness of breath. They affect a large portion of the population, with around 300 million affected and 400 000 deaths from asthma, and around 400 million affected and 3 million deaths from COPD worldwide. The most common cause of COPD is smoking, and about half of those who have smoked for a long time have developed COPD by the age of 75. In COPD, both the upper and lower parts of the airways and the blood vessels may be affected, which leads to impaired oxygen uptake and shortness of breath. Increased mucus formation may also cause coughing, as well as recurrent respiratory infections.

Another risk factor for developing obstructive lung disease is premature birth, since the lungs have not had time to develop properly, and this can lead to difficulty breathing. To facilitate oxygen uptake, the newborn may be given extra oxygen, but since oxygen is reactive this may also damage the lungs further. This damage may persist into adulthood and cause COPD-like disorders.

In asthma, the immune system reacts excessively to foreign substances in the air. The bronchial tubes contract and this causes difficulties breathing. Asthma attacks may come suddenly, with periods in-between with little to no symptoms, unlike COPD symptoms which progresses more slowly and generally is more persistent.

Neither COPD nor asthma has any curative drug, and the symptoms are often difficult to treat. The wide diversity in symptoms and responses to treatment in different individuals suggests that the diseases may be divided into several different diseases. To identify different subgroups of the diseases, we have analyzed different biomolecules, primarily in bronchoalveolar lavage fluid obtained by rinsing the airways and alveoli in the lungs, but also in blood and urine.

The amounts of the different substances vary and are called variables. To compare the different groups of patients, one variable at a time was studied and all variables together in a multivariate analysis method called orthogonal projections to latent structures (OPLS). To ensure that the models created were significantly better than those occurring by chance, we compared our models with models created after randomizing the group affiliation of the samples. We made this comparison for models both before selecting the most interesting variables and after producing new models using only the most interesting variables. We also included the choice of variables in the randomization itself, and thus obtained a robustness test for the variable choice. A tool for doing this, called roplspvs, was developed in the statistical programming language R. By applying this tool as well as the software SIMCA, which has a more user-friendly graphical user interface, to several clinical cohorts, we found several alterations related to asthma and COPD.

In the Karolinska COSMIC cohort, COPD patients who are current-smokers and ex-smokers have been compared with healthy controls who have never smoked, and smokers with normal lung function to investigate gender differences in smoking-induced COPD. MicroRNAs are a type of important regulators for regulating protein expression in the cell. Here, we investigated the levels of different microRNAs in small extracellular vesicles, liquid particles enclosed by lipid membranes and secreted from cells, in bronchoalveolar lavage. Among the microRNAs that were altered, we investigated which genes they regulate. We also studied which mechanisms these genes regulate. The mechanisms linked to COPD and smoking were mainly cell growth and cell death, with p53-related mechanisms most altered, while the microRNAs that were weakly correlated to only COPD in men were linked to degradation by autophagy and proteolysis.

In the LUNAPRE cohort, we study obstructive lung disease related to premature birth. Young adults who had been born very prematurely, and received oxygen therapy as infants due to respiratory complications (known as bronchopulmonary dysplasia, BPD) were compared with prematurely born individuals who did not need oxygen therapy in infancy, as well as healthy controls and individuals with mild asthma. It has previously been found that the BPD group have poor lung function, with comparable to COPD patients. Here, a comparison was made how the number of different immune cells differed between these groups. We found that certain types of immune cells were altered, with an increase in cytotoxic T-cells and decrease in helper T-cells in those who had BPD as newborns.

U-BIOPRED is a pan-European cohort for the study of severe asthma. As for the other studies in this thesis, the focus in U-BIOPRED was on subgrouping severe asthma into groups more relevant for the mechanisms of the disease by investigating a large number of biomolecules, and use the detailed data to build so called “handprints” of the disease subgroups. Two projects from the cohort are included in this thesis: 1) Asthmatics had been grouped by using molecules found in their blood, which resulted in eight or even 16 different groups. These groups were compared, and differences were identified using common healthcare investigations including questionnaires. 2) The metabolites in urine were measured using mass spectrometry and the effect of eating oral corticosteroids was studied. A group of metabolites, carnitines, were found to be most altered in severe asthma and was not due to eating oral corticosteroids.

In summary, a tool for multivariate analysis with careful significance testing of the models has been developed and used to compare different groups of asthmatics and COPD patients. Changes in the amount of miRNA, immune cells, and metabolites in samples, and changes in clinical picture have been identified between groups of patients. This is a small part of all the studies performed on these individuals. Together, this aims to identify the mechanisms for different variants of COPD and asthma to be able to diagnose and find new therapies for all subgroups of COPD and asthma, with the ultimate goal of finding cures.

POPULÄRVETENSKAPLIG SAMMANFATTNING

Astma och kronisk obstruktiv lungsjukdom (KOL) är båda obstruktiva lungsjukdomar som gör det svårt att andas. De drabbar en stor del av världens befolkning med årligen cirka 300 miljoner drabbade och 400 000 dödsfall i astma, och cirka 400 miljoner drabbade och 3 miljoner dödsfall i KOL. Den vanligaste orsaken till KOL är rökning, och bland dem som rökt länge har ungefär hälften utvecklat KOL vid 75 års ålder. Vid KOL kan både övre och nedre delarna av luftvägarna samt även blodkärlen påverkas, vilket leder till nedsatt syreupptagningsförmåga och andnöd. Ökad slembildning kan också ge hosta samt upprepade luftvägsinfektioner.

En annan riskfaktor för att utveckla KOL är för tidig födsel. Vid för tidig födsel har inte lungorna hunnit utvecklas till fullo, och det nyfödda barnet kan då ha svårt att andas. För att underlätta syreupptaget kan den nyfödda få extra syre, men eftersom syre är reaktivt kan detta också skada lungorna ytterligare. Skadorna på lungorna kan kvarstå till vuxen ålder och ge KOL-liknande besvär.

Vid astma reagerar immunsystemet överdrivet på främmande ämnen i luften. Luftrören drar då ihop sig och det kan bli svårt att andas. Detta kan komma i plötsliga astma-attacker, medans individen kan ha få till inga besvär emellan episoder, till skillnad från KOL-symptomen som är mer långsamt tilltagande och kroniska.

Varken KOL eller astma har något botande läkemedel och symptomen kan ofta vara svårbehandlade. Den stora variationen i symptom och svar på behandling hos olika individer tyder på att sjukdomarna var för sig skulle kunna delas upp på flera olika sjukdomar. För att identifiera olika subgrupper av sjukdomarna har vi analyserat olika substanser i framför allt bronkoalveolär sköljvätska som erhålls genom sköljning av alveolerna i lungorna, men även blod och urin.

Mängderna av de olika substanserna kan variera och kallas för variabler. För att jämföra de olika grupperna av patienter tittade vi både på en variabel i taget och på alla variabler tillsammans i så kallad multivariat analys. Metoden vi använde kallas ortogonala projektioner till latenta strukturer, OPLS. För att försäkra oss om att modellerna vi bildade var signifikant bättre än de som bildas av en slump jämförde vi våra modeller med modeller som vi skapat efter att ha blandat om grupptillhörigheten hos proverna. Denna jämförelse gjordes både innan vi valt ut de intressantaste variablerna och efter att vi gjort nya modeller på bara de intressantaste variablerna. Vi inkluderade även valet av variabler i själva randomiseringen och fick på så vis ett robusthetstest av variabelvalet. Ett verktyg för att göra detta utvecklades i statistikprogrammet R och heter roplspvs. Verktyget gjordes så pass lättanvänt att även ovana användare av R kan använda det. Med hjälp av detta verktyg samt mjukvaran SIMCA, som har ett grafiskt användarsnitt för att konstruera OPLS-modeller, fann vi åtskilliga förändringar relaterade till astma och KOL.

U-BIOPRED är en paneuropeisk kohort för studier av svår astma. Liksom de andra studierna i denna avhandling låg fokus i U-BIOPRED på att subgruppera svår astma i grupper med hjälp av ett stort antal biomolekyler, för att bygga så kallade "handavtryck" av sjukdomsundergrupperna. Två projekt från kohorten ingår i denna avhandling: 1) Astmatiker hade grupperats med hjälp av molekyler som hittades i deras blod, vilket resulterade i åtta eller till och med 16 olika grupper. Dessa grupper jämfördes och skillnader hittades i data från vanligt förekommande undersökningar i vården. 2) Metaboliterna i urinen mättes med hjälp av masspektrometri och effekten av att äta orala kortikosteroider studerades. En grupp metaboliter, karnitiner, visade sig vara lägre vid svår astma och det berodde inte på att man ätit orala kortikosteroider.

Rökande och icke-rökande KOL-patienter från kohorten COSMIC jämfördes med friska som aldrig rökt samt rökare utan KOL. Eftersom mikroRNA är viktiga reglerare för att styra gen-uttryck undersökte vi dem. Här tittade vi på mängden av olika mikroRNA i vesiklar (membranförsedda partiklar) anrikade på små extracellulära vesiklar som utsöndras från celler. Bland de mikroRNA som var förändrade tittade vi på vilka gener dessa reglerar och vilka mekanismer dessa gener i sin tur reglerar. Mekanismerna som var kopplade till KOL och rökning var främst celltillväxt och celldöd där p53-relaterade mekanismer var mest förändrade, medan de mikroRNA som var svagt korrelerade till enbart KOL hos män var kopplade till nedbrytning genom autofagi och proteolys.

För att studera KOL relaterat till för tidig födsel undersöktes unga vuxna som varit för tidigt födda och fått syrgas jämfört med för tidigt födda utan behov av syrgas. Dessa unga vuxna ingår i kohorten LUNAPRE. Man hade tidigare sett att de hade dålig lungfunktion jämförbar med KOL-patienter. Här jämfördes hur mängden av olika immunceller skiljde sig mellan dessa grupper. Det visade sig att det fanns mer cytotoxiska T-celler och mindre hjälpar-T-celler hos dem som haft syrgas länge som nyfödda. I avhandlingen visas att denna skillnad mest kunde härledas till de kvinnliga för tidigt födda.

Sammanfattningsvis har ett verktyg för multivariat analys med noggrann testning av signifikans hos modellerna utvecklats och använts för att jämföra olika grupper av astmatiker och KOL-patienter. Förändringar av mängden mikroRNA, immunceller och metaboliter i prover samt förändringar i klinisk bild har identifierats mellan grupper av patienter. Detta är en liten del av de studier som gjorts på dessa individer där även andra variabler studeras i andra vävnader i kroppen. Tillsammans siktar detta på att finna mekanismen för olika varianter av KOL och astma för att kunna diagnostisera och hitta nya terapier för alla subgrupper av KOL och astma, med mål att slutligen finna bot.

ABSTRACT

Both asthma and chronic obstructive pulmonary disease (COPD) are obstructive lung diseases with a large impact on global health, causing 400 000 and 3 million deaths respectively each year. The numbers may be underestimated, since COPD often contributes to death without being registered as the cause of death. There is currently no therapeutic cure for either of these diseases, only symptomatic relief. One problem is that the available therapeutics do not always work, since the diseases present a range of clinical phenotypes with possibly different endotypes, so-called umbrella diseases.

The aim of this thesis was to study asthma, COPD caused by smoking, and obstructive lung disease related to preterm birth using systems biology approaches. This includes studying several analytical platforms in a range of compartments in order to subphenotype the patients into subgroups and elucidate the related mechanisms. Identifying these endotypes increases the possibility of finding effective therapeutics for all patient groups.

Obstructive lung diseases are often studied by collecting bronchoalveolar lavage (BAL) and epithelial cells from the lungs during bronchoscopy. This procedure is invasive, and costly which is why the cohorts studied are often small. Subgrouping results in even smaller sample sizes, which decreases the power of statistical analysis.

Using multivariate analysis is a means of increasing power by taking all variables into account. A workflow for performing the multivariate method orthogonal projections to latent structures discriminant analysis (OPLS-DA) to compare groups one by one in small sample sizes was developed using the programming language R, and was formatted into an R package entitled roplspvs. The roplspvs package performs OPLS modeling using the package ropls in R, including variable selection to extract the variables driving the separation the most. As OPLS models are prone to overfitting, the significance of the models was investigated thoroughly using permutations. Using roplspvs on small sample sizes, it was shown that permutations performed before variable selection (termed “sans v.s.”) and permutations including the variable selection step (termed “over v.s.”) are better suited to estimate the level of model statistics achieved by random than permutations post variable selection. An example of running the package was shown using a publicly available metabolomics dataset.

The roplspvs packages, along with the commercially available software SIMCA and univariate statistics, were then applied to investigate alterations between groups in a range of projects, including three clinical cohorts of asthma, COPD and BPD, as well as a project investigating the degradation of proteins in the processing of blood samples prior to biobanking the evaluate protein stability.

COPD in smokers and ex-smokers was studied by investigating the miRNA content of small extracellular vesicles (EVs) using OPLS modeling as well as univariate analysis, with the finding that COPD gave highly altered miRNA content of small EVs compared to healthy

subjects. After stratifying the analysis by gender, potential alterations compared to smokers were identified in males with significant p[CV-ANOVA]=0.05 and p[permutations over variable selection]=0.12, but permutations sans variable selection were highly insignificant. The alterations were connected to potentially affected pathways through pathway analysis of genes regulated by the altered miRNA. Pathway affected by COPD and smoking were mainly connected to cell- growth and death with the p53-pathway mostly altered, while the less pronounced miRNA alterations related to COPD alone was connected to degradation through autophagy and proteolysis.

Premature birth has been connected to lung obstruction in adults who developed bronchopulmonary disease (BPD) during the neonatal period. To characterize T-cells in adults with a history of BPD, FACS analysis was performed on BAL cells. Univariate analysis showed increased levels of CD8+ T-cells, and decreased levels of CD4+ T-cells in subjects with BPD. Applying OPLS and stratifying the analysis by gender, it was indicated that the alterations were mostly driven by females.

Asthmatic subjects were subphenotyped into clusters using four platforms from blood and urine into phenotypic groups, which were studied using OPLS models to compare the groups. Clinical features were extracted that separated a large portion of the groups.

Finally, the metabolome of urine was used to separate asthmatics into severe and mild asthma, stratifying the analysis by oral corticosteroids (OCS). It was found that carnitines, which were the strongest drivers for separating the groups, were not affected by OCS use. Using roplspvs, it was shown that the levels of carnitines were strongly affected by gender, with higher levels in males than in females.

In conclusion, it was shown that OPLS models can be used to investigate cohorts consisting of small sample sizes, and that permutation procedures including variable selection efficiently test the significance of the models. Subgroups of COPD and asthmatic subjects were compared, showing alterations in miRNA levels, metabolome, and lymphocyte composition, as well as in clinical data connected to potential endotypes with separate disease mechanisms. Stratifications by gender supported earlier findings that gender has a strong effect on obstructive lung diseases. Together with further analysis on the cohorts in this study using other platforms, this is a step towards finding candidates for diagnostics and therapeutics.

LIST OF SCIENTIFIC PAPERS

I. **Permutation analysis prior to variable selection greatly enhances robustness of OPLS analysis in small cohorts**
Ström M, Wheelock ÅM
Manuscript

II. **Alterations in extracellular vesicle miRNA cargo correlates with lung function and small airways disease**
Ström M *, Eldh M*, Brundin B*, Bhakta NR, Chuan-xing Li, Yang M, Pollack JL, Heyder T, Karimi R, Nyrén S, Erle DJ, Gabrielsson S, Sköld CM, Schekman RW, Wheelock ÅM
Manuscript

III. **Increased cytotoxic T-cells in the airways of adults with former bronchopulmonary dysplasia**
Um-Bergström P, Pourbazargan M, Brundin B, **Ström M**, Ezerskyte M, Gao J, Berggren Broström E, Melén E, Wheelock ÅM, Lindén A, Sköld CM Eur Respir J. 2022, Online ahead of print. PMID: 35210327

IV. **Asthma patient stratification through integration of multiomics and clinical data: U-BIOPRED adult blood-urine handprints of clinical utility**
De Meulder B*, Li CX*, **Ström M**, Kermain NZ, Yen RTC, Lefaudeux D, Bigler J, Loza M, Burg D, Schofield J, Brandsma J, Kolmert J, Skipp P, Postle A, Maitland-van der Zee AH, Bakke PS, Caruso M, Chanez P, Fowler SJ, Geiser T, Howarth P, Horváth I, Krug N, Behndig A, Singer F, Musial J, Shaw DE, Dahlén B, Rowe A, Baribaud F, Montuschi P, Sterk PJ, Chung KF, Roberts G, Djukanovic R, Dahlèn SE, Adcock IM, Wheelock CE, Auffray C*, Wheelock ÅM* and the U-BIOPRED study group
Manuscript

V. **Urinary metabotype of severe asthma evidences decreased carnitine metabolism independent of oral corticosteroid treatment in the U-BIOPRED study**
Reinke SN, Naz S, Chaleckis R, Gallart-Ayala H, Kolmert J, Kermani NZ, Tiotiu A, Broadhurst DI, Lundqvist A, Olsson H, **Ström M**, Wheelock ÅM, Gómez C, Ericsson M, Sousa AR, Riley JH, Bates S, Scholfield J, Loza M, Baribaud F, Bakke PS, Caruso M, Chanez P, Fowler SJ, Geiser T, Howarth P, Horváth I, Krug N, Montuschi P, Behndig A, Singer F, Musial J, Shaw DE, Dahlén B, Hu S, Lasky-Su J, Sterk PJ, Chung KF, Djukanovic R, Dahlén SE, Adcock IM, Wheelock CE; U-BIOPRED Study Group.. Eur Respir J. 2022 Jun 30;59(6):2101733. doi: 10.1183/13993003.01733-2021. PMID: 34824054; PMCID: PMC9245194.

VI. **Proteome stability markers identified in plasma samples using tandem mass tags**
Ström M, Olsson BM, Tybring G, Sihlbom C*, Wheelock ÅM*
Manuscript

SCIENTIFIC PAPERS NOT INCLUDED IN THE THESIS

I. **Widespread episodic thiamine deficiency in Northern Hemisphere wildlife**
Balk , Hägerroth PÅ, Gustavsson H, Sigg L, Åkerman G, Ruiz Muñoz Y, Honeyfield DC, Tjärnlund U, Oliveira K, Ström K, McCormick SD, Karlsson S, **Ström M**, van Manen M, Berg AL, Halldórsson HP, Strömquist J, Collier TK, Börjeson H, Mörner T & Hansson T
Sci Rep 2016, 6, 38821

CONTENTS

LIST OF ABBREVIATIONS

ACQ	Asthma Control Questionnaire
AGS	altered gene sets
AM	Alveolar macrophages
AQLQ	Asthma Quality of Life Questionnaire
BAL	Broncho alveolar lavage
BALF	Broncho alveolar lavage fluid
BEC	Epithelial cells
BPD	Bronchopulmonary dysplasia
CD4	Cluster of differentiation 4
CV	Cross validation
COPD	Chronic obstructive pulmonary disease
CPAP	Continuous positive airway pressure
E+P	estrogen and progestin
E-only	Estrogen only
ECM	extracellular matrix
ELF	epithelial lining fluid
ESCRT	Endosomal sorting complex required for transport
ESS	Epworth Sleepiness Scale
EV	extracellular vesicles
sEV	Small extracellular vesicles earlier called exosomes
FeNO	fraction of exhaled nitric oxide level
FEV1	Forced expiratory volume in one second
FDR	False discovery rate
FVC	Forced vital capacity
GINA	Global Initiative for Asthma
GM_CSF	granulocyte-macrophage colony-stimulating factor
GOLD	Global initiative for chronic obstructive lung disease
GSEA	gene set enrichment analysis
HADS	Hospital Anxiety and Depression Scale
HBEC	bronchial epithelial cells

HLA	Human leukocyte antigen
IPA	Ingenuity Pathway Analysis
IRDS	Infant respiratory distress syndrome
IL	Interleukin
LC	Liquid chromatography
LLD	Lower limit of detection
LLN	Lower limit of normal
LLQ	Lower limit of quantification
miRNA	microRNA
MHC	Major histocompatibility complex
MMA	Mild to moderate asthmatics
MMP	matrix metallopeptidases
MS	Mass spectrometry
MVA	Multivariate analysis
MVB	multivesicular bodies
NEC	necrotizing enterocolitis
NF-κB	nuclear factor-κB
NIPAL	Nonlinear Iterative Partial Least Squares
OPLS	Orthogonal projections to latent structures
ORA	over-representation analysis
PC-CVA	principal components–canonical variate analysis
PCR	Polymerase Chain Reaction
Perm.	Permutations
P(corr)	
PH	pulmonary hypertension
PLS	Partial least squares
PCA	Principle component analysis
Q2	Q2(cum), cumulative fraction of variationcd
Q-TOF	Quadrupole time-of-flight
RASP	Refractory asthma stratification programme
ROC	Receiver-operator-characterstics

ROS	Reactive oxygen species
RSD	Residual standard deviation
R2	R2Y(cum), cumulative fraction of variance of all Y´s explained by the model
RV	Residual volume
SANS	Severe asthmatics non smokers
SA	Severe asthmatics
SD	Standard deviation
SNF	Similarity Network Fusion
SUS	Shared and unique structures
Th1	T helper cells class 1
Th2	T helper cells class 2
TMM	Trimmed mean of M-values
TMT	tandem mass tags
TNF-α	tumor necrosis factor α
UV	Unit variance
VEGF	vascular endothelial growth factor
VIP	Variable Importance in the Projection
VC	vital capacity
v.s.	Variable selection
YBX1	Y-box protein 1

1 INTRODUCTION

Many diseases are so-called umbrella diseases, meaning that similar symptoms and may not always be caused by a range of different pathological mechanisms. The mechanism for developing the disease may thus differ between subjects. Subgroups related to mechanism are generally referred to as endotypes, whereas subgroups related to observable characteristics – such as symptoms or biomarkers unrelated to mechanism – are referred to as phenotypes. In the search for endotypes, subjects are characterized by their phenotypes. Subgrouping subjects with the same phenotype and performing pathway enrichment analysis is a means of finding different endotypes of a disease. Characterizing the different endotypes is important to be able to find specific diagnostics and therapeutics that are effective for each endotype. Systems biology is the method of studying the overall picture by integrating information about phenotypes.

Both asthma and COPD are examples of umbrella diseases. They both have an increasing disease burden. In 2010, the number of COPD cases was estimated at 384 million globally [1], and this was the third most common cause of death, with 3 million COPD-related deaths [2]. In 2019, the number had increased to 392 million people [3]. Since the most common risk factors – smoking, air pollutants, and premature birth – are not decreasing worldwide, the disease burden is not anticipated to decrease. Even if the prevalence of smoking is decreasing, the population is growing, resulting in an increasing number of smokers [4]. There is currently no cure for the persistent airway obstruction caused by COPD, and only symptom relief through, e.g., bronchodilators and inhaled corticosteroids are available.

Asthma is an endemic disease that causes morbidity in 300 million people worldwide, and caused 400 000 deaths during 2019 [5]. Like COPD, asthma has no cure, and the treatment includes symptom relief and control of inflammation. In addition to bronchodilators and inhaled corticosteroids used to treat COPD, the treatment for certain subgroups of asthma patients includes biological anti-inflammatory agents. Typical features of asthma include episodic narrowing of the airways due to inflammation, hyperreactivity, and increase in smooth muscle mass, as well as excessive mucus production which may result in wheezing, shortness of breath, and coughing. However, these airway modulations and symptoms exist in numerous variations, representing a large spectrum of asthma phenotypes [6].

To study the local inflammation in the airways, resident immune cells, airway exudates and immunomodulatory biomolecules can be collected from the site of inflammation through bronchoalveolar lavage using bronchoscopy. Since this is an invasive procedure, the number of research subjects is usually scarce, and sub-phenotyping of this heterogeneous group results in even smaller sample sizes.

Using univariate statistical methods sometimes gives a statistical power that is too poor to be able to separate different subgroups when the sample sizes are too small. One way to increase power is to use multivariate methods, taking a wide range of variables into account. One method that is especially suitable for separating groups and extracting biomarkers that drive the separation is orthogonal projection to latent structures (OPLS). OPLS is a supervised method that has been developed from PLS, where the first principal component describe the variation between the groups, and the next principal components describes the variation within the groups [7]. Even though the methods perform equivalently the OPLS models are more easily interpreted as the first principal component represents the predictive part separating the groups. A common tool used for OPLS modeling is SIMCA, which has a graphical interface. Modeling in SIMCA is time consuming, and it is easy to introduce errors during the tedious modeling, especially when many groups are compared. Using command line in R results in automated and reproducible modeling. It also makes it easy to apply the analysis to new data, which not only saves time, but also avoids human errors. OPLS models are easily overfitted, and perfect separation can always be fitted if there are too many variables and too few subjects. Using too many principal components in an OPLS model may also result in an overfitted model. Therefore, it is important not only to interpret the analysis visually, but also to utilize the appropriate model statistics and to test the significance of the models. One method for significance testing uses permutations, which involves performing models on randomized data to test whether models are better than a model created at random.

Here, we present an R package that performs OPLS-DA modeling, comparing groups of subjects one by one and including variable selection. The development of this tool represents the core of this PhD dissertation, and features include significance testing using permutation tests both before and after variable selection, as well as including variable selection in the permutation procedure. In order to include the variable selection in the permutation test, variables are selected from both the unpermutated dataset and the permutated dataset. This allows for significance testing of the model itself as well as the variable selection process.

OPLS-DA models and univariate analysis have been performed comparing subgroups of asthmatics and COPD subjects, using a wide range of platforms including miRNA cargo of extracellular vesicles (EVs), levels of lymphocytes, metabolites, and proteins, as well as clinical data. The script has also been used to test the performance of OPLS models on small sample sizes, and shows that permutations both before and including variable selection are useful tools for establishing a baseline significance level in order to avoid erroneously interpreting overfitted models as being significant.

2 LITERATURE REVIEW

2.1 DATA ANALYSIS

2.1.1 Univariate analysis

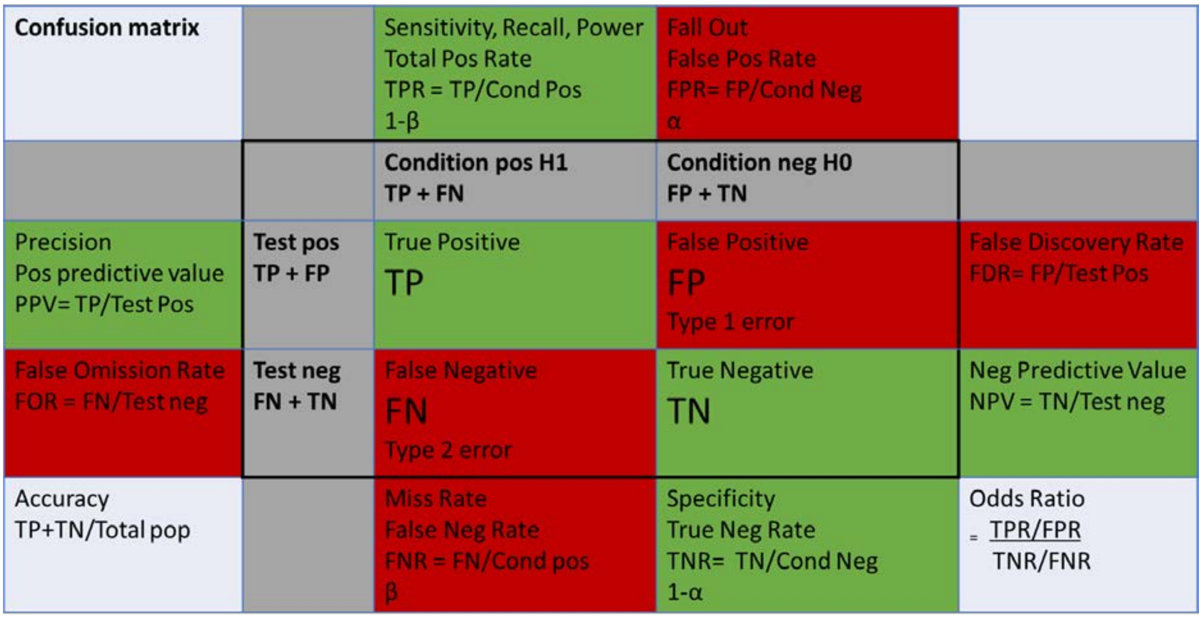
Univariate statistical testing refers to the process of comparing the means of two populations or samples. There are a range of different methods available which can be roughly divided into parametric tests, where some assumptions are made regarding the structure and distribution of the data, and non-parametric tests, that do not apply any strict rules on the data distribution. In univariate statistics, each variable is tested independently from the others in the data set. This may cause an inflated false positive rate in large data sets.

2.1.2 Multiple testing correction

A particular challenge in the analysis of large datasets resulting from omics screening studies is that the datasets consist of many analytes but relatively few observations, rendering a “short and wide” data matrix. Under these circumstances, traditional univariate statistical methods may result in many false positives. One strategy to circumvent the high false positive rate is to do multiple testing corrections, where the p-value is adjusted to account for the number of tests (i.e., independent variables). Methods for p-value correction include the Bonferroni adjustment and the more moderate Benjamini-Hochberg procedure. With either approach, the result is an increase in false negatives, i.e., a loss of statistical power, as the false positives decrease.

2.1.3 False findings

There is an inherent risk of false findings (Figure 1) in research [8], and it has been proposed that these differ between sciences with more positive results in psychology and psychiatry and fewer positive result in space science compared to biological studies [9]. The likelihood of finding false positives, i.e., making a type 1 error, in a study using established methodology is a major issue in all kinds of fields, ranging from modeling [10] and micro arrays [11] to clinical trials [12]. The reasons for the problem have been summarized by Ioannidis as depending on many factors, including how many groups are involved in trying to find statistically significant findings and how many studies has been performed on the same subject, as well as publication biases due to it being more rewarding to publish positive findings than negative findings. Other factors that result in more false findings are hot fields, greater flexibility in study design, and small sample sizes. Finally, the smaller the effect sizes, the greater the risk of the finding being false [13]. In the field of lung research, bronchoscopy is often used to collect samples from the lung. The invasive procedure of bronchoscopy often results in small cohorts being used. As a result, when all the parameters are fixed, decreasing alfa lowers the number of false findings.



Confusion matrix		Sensitivity, Recall, Power Total Pos Rate TPR = TP/Cond Pos 1-β	Fall Out False Pos Rate FPR= FP/Cond Neg α	
		Condition pos H1 **TP + FN**	**Condition neg H0** **FP + TN**	
Precision Pos predictive value PPV= TP/Test Pos	**Test pos** **TP + FP**	True Positive TP	False Positive FP Type 1 error	False Discovery Rate FDR= FP/Test Pos
False Omission Rate FOR = FN/Test neg	**Test neg** **FN + TN**	False Negative FN Type 2 error	True Negative TN	Neg Predictive Value NPV = TN/Test neg
Accuracy TP+TN/Total pop		Miss Rate False Neg Rate FNR = FN/Cond pos β	Specificity True Neg Rate TNR= TN/Cond Neg 1-α	Odds Ratio = TPR/FPR TNR/FNR

Figure 1. *Confusion matrix describe outcomes when testing a condition.*

2.1.4 Power

Statistical power is defined as the ability to identify true positives in statistical analysis (Figure 1). The problem with false findings must be balanced against the risk of making a type 2 error of not finding true positive findings, decreasing power of analysis, as decreasing alfa will also decrease the power. Balancing alfa should be done based on the specific research question, as a type 1 error is sometimes worse than a type 2 error and vice versa. The traditionally selected alfa of 5% and power of 80% may not always be satisfactory, as for example having a low power in new fields may miss out on the opportunity for further investigations. Factors that decrease both type 1 and type 2 errors are decreasing variance, increasing sample size, and increasing the effect size. Calculating the required sample size to achieve a desired power is strongly recommended, and is often requested in applications [14]. Failing to do so may result in an underpowered study, which may lead to a temptation to draw conclusions from too low a significance. Sample size is calculated from the power, the effect size, and the significance level using webtools or R.

2.1.5 Effect size

As mentioned above the effect size is needed for estimating required sample size is usually the most difficult to estimate in advance [14].

To know if an alteration is important not only is the significance of the alteration needed but also the effect size.

It is not only necessary to now if an alteration is significant but also the size of the alteration. Effect size is a standardized size of an alteration calculated by difference between means divided by the standard deviation

Effect size for parametric test Cohen's d may be calculated using the cohens_d function and non-parametric Wilcoxon effect size using the wilcox_effsize function, both included in R package rstatix.

2.1.6 Multivariate statistical modelling

One way to avoid the problem of low power is to use multivariate analysis (MVA) [15]. In multivariate statistical analysis, all variables in a dataset are incorporated in one (or a limited number of) models, thereby reducing the number of statistical hypothesis tests performed compared to univariate statistics. The covariance of the data is taken into account in MVA, resulting in both fewer false negatives and fewer false positives.

In multivariate methods the relationship between two matrixes, X and Y are described. Y contains the explained variables, also called dependent variables or responses, while X contains the explanatory variables, also called independent variables or predictors. In multivariate statistical modeling, the X matrix is projected onto latent variables called principal components. Multivariate analysis can be divided into the two categories, supervised and unsupervised. A supervised uses a definition of the group belonging, often times assigned in the Y matrix, while in an unsupervised method the group belonging is not defined in the model and is thereby data-driven. The most commonly used method principal component analysis (PCA) is unsupervised. It can be used to see the large structure in the dataset such as whether groups of subjects cluster, to identify batch effects and to find outliers.

2.1.7 PLS and OPLS models

Projection to latent structures by means of partial least squares (PLS) [16] provides a linear regression that explains the data block Y using the data block X. The variables in X are projected into a scores vector. By adding principal components to the model, additional scores vectors are also added to the model. The loadings explain how each variable contributes to modeling Y in proportion to the other variables.

A development of PLS is orthogonal projections to latent structures (OPLS) [7], which often is used as a supervised method. The first predictive principal component is rotated compared to a classical PLS, and oriented to separate the groups as efficiently as possible, whereas the orthogonal components represent the subject variation within the group (e.g., technical noise or confounders). The predictive power of OPLS is the same as for PLS, but OPLS is easier to interpret [17] as the most important group separating power is described by the predictive component. In OPLS-DA [18] (discriminant analysis), the Y-matrix is binary and defines the groups as 1 or 0 (e.g., patients or healthy individuals). OPLS-DA describes models that efficiently separate groups. It deserves mention that OPLS can also be used in an unsupervised exploratory fashion to correlate two data blocks, such as in O-2-PLS [19]. Further expansions to allow correlation of multiple data blocks have also been developed, as in the O-n-PLS algorithm [20].

An advantage of OPLS models is that they can model variables that are noisy and collinear, which makes them suitable for omics data, particularly from human cohorts where lifestyle factors and genetic diversity can contribute to large intra-group variance [21]. In OPLS, it is important for the group sizes to be as balanced as possible; otherwise, the larger group will be penalized. The mean of the groups will be shifted toward the larger group, and thereby, misclassifications of the larger group may occur [22].

OPLS is frequently used in drug design [23] and omics data analysis [15], especially in the field of metabolomics.

A common tool for performing OPLS modeling is the software SIMCA [24], which has a graphical interface and was developed by Umetrics, a company that is now incorporated into Sartorius. Tools for OPLS have also been developed in R, and a kernel-based OPLS package in R [25].

2.1.8 Validation of the models

R^2Y, often simply referred to as R^2, describes how much of the variance in the data is explained by the model, and is calculated by subtracting 1 from the sum of squares. Q^2Y, often simply referred to as Q^2, is the goodness of prediction determined by cross-validation. Cross-validation is a tool frequently used in small cohorts that do not allow for the preferred strategy of dividing the full data set into a training set and test set. In cross-validation, models are iteratively created by removing a portion of the data, then classifying the removed data using the remaining data. Our group has previously described the minimum requirements of model statistics that should be presented for OPLS models, recommending that R^2 and Q^2 be reported [15]. These are now the most common statistics to present when reporting results in metabolomics data [26].

Another important means of quality control (QC) of Q^2 is by permutating the dataset. E.g., in SIMCA, the default is to split the data by order of the data matrix, into seven-fold cross-validation. This praxis may affect the calculated Q^2 based on the order of the groups in the data set, especially when the sample sizes are small [27]. The confidence of Q^2 may be determined by repeated cross-validation using multiple iterations of cross-validation sets.

2.1.9 Overfitting OPLS models

In univariate models, it is important to compensate for multivariable selection. This is done by adjusting p-values for multiple hypothesis testing bias (see section s2.1.2). It is important also in multivariate modeling to be aware of the effect of selecting variables from among many variables. Models with high R^2 values can be created by applying PLS models to randomized data, yielding seemingly perfect separation [28]. It is well known that increasing the number of orthogonals results in a model that separates the groups in a score plot, at least by visual inspection. This can easily be tried, resulting in increased R^2. In this context, it is essential to evaluate the effect on other model statistics. As the R^2 is inflated, representing an overfitting of the model to the specific data set at hand, the Q^2 value, representing the predictive power of the model when applied to new data, decreases when too many orthogonals are added. It is therefore important to keep the difference between R^2 and Q^2 low. It is less well known that too many variables and too few samples can also produce a perfect separation according to a score plot [22] . The reason for this overfit is that the PLS and OPLS algorithm finds structures in the dataset and uses the variables that correlate the best with Y. An example is shown in Figure 2, where the model is created using 8 random subjects in each group and 1,016 variables. It has therefore been suggested that cross-validated score plots should be used, which is less prone to show the overfit.[29]

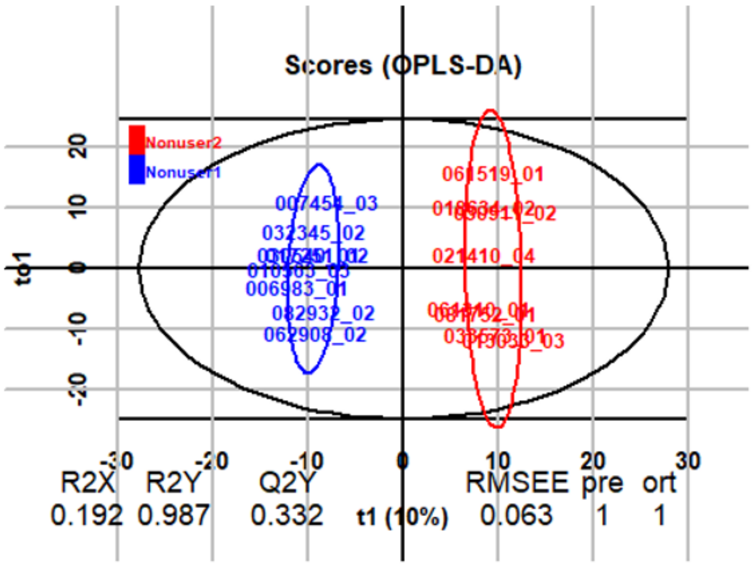


Figure 2. *A model using many variables (k=1016) and few subjects (n=16) comparing groups created by randomly selected subjects from a group of women aged 61-65.*

2.1.10 Significance of model

How the significance of models is tested in the literature varies to a large extent. Often, no variable selection is performed, as is the case for the multivariate models in paper V, which avoids overfitting caused by variable selection. Significance is sometimes tested by permutations that are performed before variable selection [30], but significance is often only tested for accuracy [31], and at other times it is unclear whether the model statistics presented refer to testing performed pre- or post variable selection.

The risk of performing permutation test solemnly post-variable selection, is that models that truly are based on random variability in the data set can be interpreted as true positives. To avoid this dilemma, and to assure that the selected variables best describe the separation, cross-validation or permutations including the variable selection should be used, an area that is explored in this thesis.

There are a number of methods for significance test that may be used on omics data, which usually consists of a much larger number of variables than the number of samples analyzed.

Such methods include but are not limited to permutation tests, decision trees, bootstrapping, and CV-ANOVA. CV-ANOVA [32] is a two-way analysis of variance of prediction results using cross-validation [33].

A strength of permutation tests is that it may be used as a test of significance without any assumption about the distribution of the population, i.e., in a non-parametric fashion [34]. Permutation tests are widely used to test the significance of PLS and OPLS models [35]. During the permutation procedure, the group labels of the subjects (Y-matrix) are randomized. By comparing R^2 and Q^2 of the un-permutated model with the permutated, the best model produced at random can be established for the data set, thereby setting a threshold for what model statistics may be considered significant. [36]

Permutation test may be used to determine the p-value of R^2 and Q^2. Permutation tests are included in both the ropls package and the SIMCA software.

2.1.11 Feature selection

In order to increase the interpretability and remove noise from the analysis, nonsignificant variables may be removed from the analysis in a process called feature selection or variable selection. Feature selection is performed to remove irrelevant and redundant variables.

A common way to select variables is by means of variable importance on projection (VIP) [37]. VIP measures the relative influence of each X-variable on the model, but simply represents a ranking of the variables centered around 1.0. If all variables influence the model equally, their VIP would each be 1. A common cutoff is 1. However, the fact that VIP is relative limits its utility in variables selection. We therefore often use scaled loadings (P[corr)]. Using the partial and semi-partial correlation coefficients for each variable removes the effect of all other variables, and makes its values robust between models. The cutoff |P(corr)|>0.4 correspond to an approximate p-value of 0.05 for the cohort size used in LUNAPRE and COSMIC [38].

OPLS models are very well suited for extracting features to include in pathway analysis [39]. For this purpose, it is useful to include as many variables as possible that contribute significantly to the model. When extracting features for diagnostic or therapeutic purposes, it is useful to extract as few features as possible while still creating a significant model.

2.1.12 Significance of variable selection

It is essential to assure that OPLS models are not overfitted, especially when small sample sizes are used. Visual inspection of whether groups separate also in a non-supervised method, e.g., PCA, is a common way to test the reliability of an OPLS model [40]. However, in large scale omics data sets from human cohorts, where the groups sizes may be limited, the number of confounders sizeable, and the number of variables large, a unsupervised model is unlikely to provide a satisfactory group separation. Shrinkage, where fitted data performs worse on new data than on the data used to create the model, was proposed to be accounted for in the

Lasso and Ridge regression for linear regression. In order to account for shrinkage due to variable selection, different methods have been proposed to monitor this shrinkage. One way is to apply bootstrap over the variable selection [41], performing repeated variable selection in each bootstrap to estimate a p-value for the selected variables. Other options are to perform cross-validation on the variable selection process [42] [43] or a procedure for variable selection by double cross-validating the feature selection procedure [44]. Finally, the permutations over variable selection procedure has been suggested by Lindgren et al. [45], suggesting that the increase in Q^2 during variable selection in models created on permutated data is a measure of the overestimation (overfitting) of the model due to variable selection. Permutations over variable selection has been used for PLS and random forest in the R package MUVR [46].

2.2 THE HEALTHY LUNG

2.2.1 Lung physiology

The lung is the organ in which oxygen and carbon dioxide is exchanged between the air and the blood. The right lung consists of three lobes: the upper, lower, and middle lobes. The left lung consists of two lobes – the upper and lower lobes – with the middle lobe being replaced by part of the upper lung, which is called the lingual and has a similar function to the middle lobe of the right lung. Air is breathed in by contracting the diaphragm. This creates lower pressure in the thoracic cavity which is transferred to the lungs via the pleura, enabling the air to flow into the lungs. The maximum amount of air that can be breathed out in a relaxed way including TV, IRV, and ERV is called the vital capacity (VC), and the amount that can be

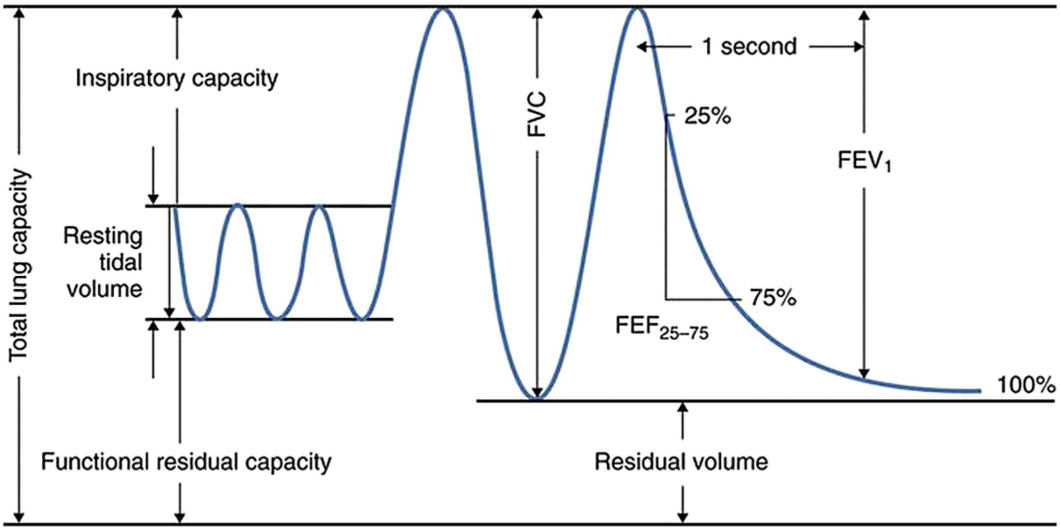


Figure 3. *Spirogram showing someone breathing normally and proceeding to inhale as much as possible two times and exhaling as much as possible two times. FVC Force vital capacity Lung volumes and FEV$_1$ forced expiratory volume at one second.*

Reprinted from Semin Fetal Neonatal Med 19(2), Gibson, A. M. and L. W. Doyle. "Respiratory outcomes for the tiniest or most immature infants.". 105-111, Copyright (2014) with permission from Elsevier stressed out is called forced vital capacity (FVC). A common measure for establishing lung disease is to measure the amount of air that can be exhaled in one second in a forced manner, called forced expiratory volume (FEV1), and then compare this with FVC (Figure 3).The air remaining in the lungs after forced exhalation is called the residual volume (RV). The dead space is the volume of air that does not participate in gas exchange. This consists of alveolar dead space as well as the conducting zone, i.e., the nose, trachea, and bronchi, and is about 30% of the TV.

During inhalation, the air passes through the nasal cavity where it is humidified and filtered, via the trachea that divides into the right and left main bronchi, which further divides into the secondary bronchi leading to each lobe. The branching continues and forms a tree that on average divides 23 times [47] via proximal and terminal bronchioles belonging to the conducting zone of the lung and the respiratory bronchioles, which belong to the respiratory zone. Finally, the alveolar ducts, connect to the alveolar sacs containing multiple alveoli where the gas exchange take place. The trachea and bronchi are surrounded by hyaline cartilage and smooth muscles to prevent them from collapsing on expiration. The bronchioles are only supported by smooth muscles.

The airway epithelium from the trachea through the proximal bronchioles primarily [48] consists of columnar ciliated cells and goblet cells. Goblet cells produce mucins, an important constituent of the epithelial lining fluid (ELF). Together, the ELF and the beating of the cilia form the mucociliary escalator, which clears particles, microorganisms, and other pollutants from the airway. In the distal and terminal airways, the goblet cells are increasingly replaced by the secretory Clara cells, which are nonciliated, protein-secreting cells with metabolic and immunological activity. Basal cells are more common in the upper airways, and are the primary stem cells of the lung. The epithelial cells are connected to the basement membrane, which is part of the extracellular matrix (ECM). The cells that produce ECM are fibroblasts, and they are also major players in injury repair. Infiltrating immune cells, primarily monocytes, macrophages, and dendritic cells, are also present in the epithelia. Mesenchymal stromal cells (MSCs) are progenitor cells that are resident in the lungs.

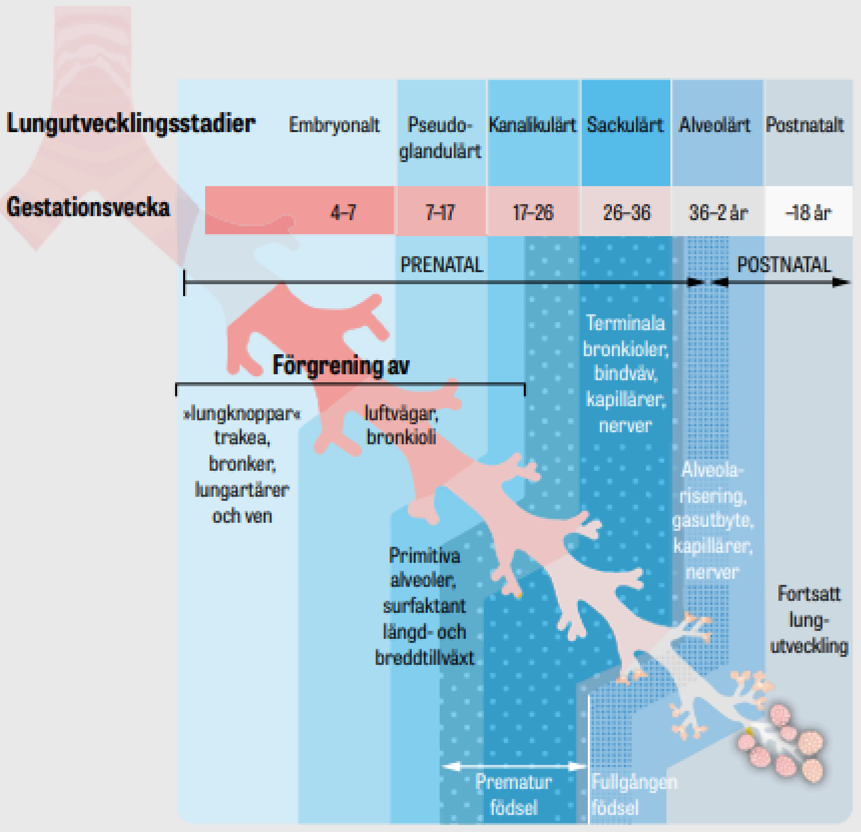
The alveolar epithelium is 95% covered by type 1 cells, a flat thin cell type that – together with the endothelial cells – enables gas exchange [49]. The remaining 5% of the alveolar epithelium are covered by type 2 cells, a cuboidal cell type that secretes surfactant and epithelial lining fluid [50]. Type 2 cells are also the main progenitor cells in the alveolar epithelium, and may differentiate into type 1 cells which cannot divide by themselves. Alveolar macrophages also reside in the alveolar space. It has been proposed that there are two subpopulations of macrophages, M1 an M2 [51] with M1 promoting Th1 respons and M2 promoting Th2 respons. This classification is an oversimplification when it comes to resident alveolar macrophages though, and investigations with single-cell analyses indicates a broad repertoire of macrophage phenotypes with significant plasticity [52].

The lung is perfused by two separate blood circulation systems [53]. The pulmonary circulation goes from the right ventricle of the heart to the alveolar capillaries, and returns oxygenated blood to the left atrium of the heart. The bronchial circulation brings oxygenated blood via arteries to the airway structure of the lung.

2.2.2 Lung development

The development of the human lung is divided into five phases [47] during gestation (Figure 4), but the lung continues to develop until the age of 20–25 years. During the embryonic stage (0–7 weeks), the lung bud forms from the foregut, further developing into the right and left main bronchi, with further divisions resulting in 18 major lobules and segmental bronchi by the end of the period. Simultaneously, the pulmonary arteries and veins start to develop. The pseudo glandular stage (7–17 weeks) is so called because the primitive alveoli looks like glandular tissue. During this phase, further branching of the respiratory tree take place. The cuboidal cells, which will later develop into type 2 alveolar cells, and smooth muscles form. During the canalicular phase (17–27 weeks), bronchioles, alveolar ducts, and primitive alveoli are formed. Alveolar type 1 and 2 cells are formed, and by week 24 alveolar type 2

FIGUR 1. Schematisk skiss över lungans utveckling in utero och efter partus



▶ Bilden visar dels de embryonala utvecklingsstadierna i relation till gestationsvecka, dels vilka delar av lungutveckling som sker vid respektive stadium. Tidsintervallet för fullgången graviditet (vecka 37–42) är också indikerat.
Illustration: Fuad Bahram/Typoform.

Figure 4. *Embryonic development stages with gestational weak showing which parts of the lung are developed during each stadium. Fullterm birth (weak 37-42) is indicated. Reprinted from Läkartidningen. 2022;119:21214, Stern R, Um-Bergström P, Sköld M, ”Lungkomplikationer hos vuxna ibland kopplade till tidig födsel" with permission from Läkartidningen and Fuad Bahram/ Typoform*

cells start producing surfactant. During the saccular phase (28–36 weeks), enlargement of acinar tubules forms saccules, resulting in thinner type 1 alveolar and endothelial cells leading to better gas transfer [54].

2.2.3 Surfactant

Surfactant is produced primarily in the alveoli by alveolar type 2 cells, and to some extent by club cells in the distal airways. It consists of dipalmitoylphosphatidylcholine and other phospholipids, together with surfactant proteins. It lowers the surface tension between the cell membranes of the airway and alveolar epithelia, thereby preventing collapse of the airway and alveolar structure on exhalation. In 1959, it was realized that the lack of surfactant in early born children could cause disease. In 1972, it was shown that prematurely born rabbit fetuses could inhale more air after the administration of surfactant from adult rabbits. This was also done by Fujiwara and Maeta in 1980, when the first preterm infants with RDS were treated with bovine surfactant [55].

2.3 PRETERM BIRTH

The prevalence of preterm birth has increased for various reasons. Worldwide, 11.1% of all births were preterm from 1990 through to 2010, ranging from 5% to 18% in different countries [56]. Preterm birth is generally divided into two categories: spontaneous early onset and induced early onset. The difference in prevalence is due to many factors, including increased use of in vitro fertilization, resulting in multiple pregnancies with a higher risk of premature delivery [57]. Other risk factors are intrauterine growth restriction, uterine over distension, previous preterm birth, low body-mass index in the mother, and pre-eclampsia or eclampsia and other maternal medical disorders, such as infections [58].

2.3.1 Definitions of premature birth

According to the 2012 WHO report “Born Too Soon: The Global Action Report on Preterm Birth” [36], the time of delivery is defined in three categories:

- Preterm birth until gestational week 36 and 6 days
- Term birth between gestational week 37 and 0 days and 41 week and 6 days
- Post-term from gestational week 42 and 0 days

In this definition, preterm birth is divided into three categories:

- Extremely preterm birth until gestational week 27 and 6 days
- Very preterm birth between gestational week 28 and 0 days and week 31 and 6 days
- Late preterm birth between gestational week 32 and 0 days and week 36 and 6 days

2.3.2 IRDS

The main cause of infant respiratory distress syndrome (IRDS), also called hyaline membrane disease, is lack of surfactant in the lungs of preterm-born babies, leading to collapse of the airways and alveoli during exhalation [59]. IRDS is diagnosed by an increased rate of breathing, retractions during breathing, and chest x-rays showing decreased lung volumes and a reticulogranular pattern. It is defined as an increasing need for oxygen during the first day of life [60].

The progression of IRDS differs between patients, even in children of the same age. In a study of 1340 infants born between weeks 23 and 27, it was shown that the infants recover to different degrees during their first two weeks of life, with 20% showing mild disease with consistently low fraction of inspired oxygen, 38% showing an initial recovery and reoccurring need for supplemental oxygen, and 43% showing severe symptoms with early and consistently high fraction of inspired oxygen [61].

2.3.3 BPD

Bronchopulmonary dysplasia (BPD) is a chronic lung disease that may develop from IRDS, with more severe IRDS sufferers being more prone to develop BPD [61]. The definition of BPD is that supplementary oxygen is needed for at least 28 days after delivery. The degree of BPD is determined at week 36 after conception. Mild BPD is defined as no need for oxygen at this time. Moderate BPD means less than 30% oxygen is needed and severe BPD means more than 30% is needed.

The importance of early life events in the development of obstructive lung disease was recently highlighted by Lange et al. [62], who propose four distinct trajectories (TR) of decline in lung function over life based on observations from three large population-based studies (Figure 5). As opposed to the classical COPD phenotype resulting from an accelerated decline in lung function due to, e.g., smoking (upper dashed curve in Figure 5), two of these trajectories (lower curves in Figure 5) start out with a lower-than-normal peak lung function in the mid-twenties. This reduced maximal plateau is a result of derangements occurring due to early life events, e.g., frequent infections or premature birth. Even at a normal rate of lung function, many of these individuals will get a diagnosis of COPD early in life (lower dashed curve Figure 5), and lifestyle choices such as smoking, environmental or occupational exposure, etc. will have a larger impact on these individuals. This group may represent completely different pathways leading to COPD, where early life events rather than smoking trigger the accelerated decline in lung function leading to COPD later in life (lower line in Figure 5). Premature birth, and the need for oxygen treatment during the neonatal period (BPD) represent risk factors for developing early onset obstructive lung disease (lower dashed line Figure 5), as opposed to the “classical” smoking-induced sub-phenotypes of COPD, with onset at age 45–70 (upper dashed line Figure 5). Thanks to tremendous improvements in neonatal care, prematurely born infants now represent >1% of all newborns (4). Our knowledge and understanding of the underlying molecular mechanisms of COPD in this rapidly growing group, as well as in other never-smokers’ COPD phenotypes, is very limited.

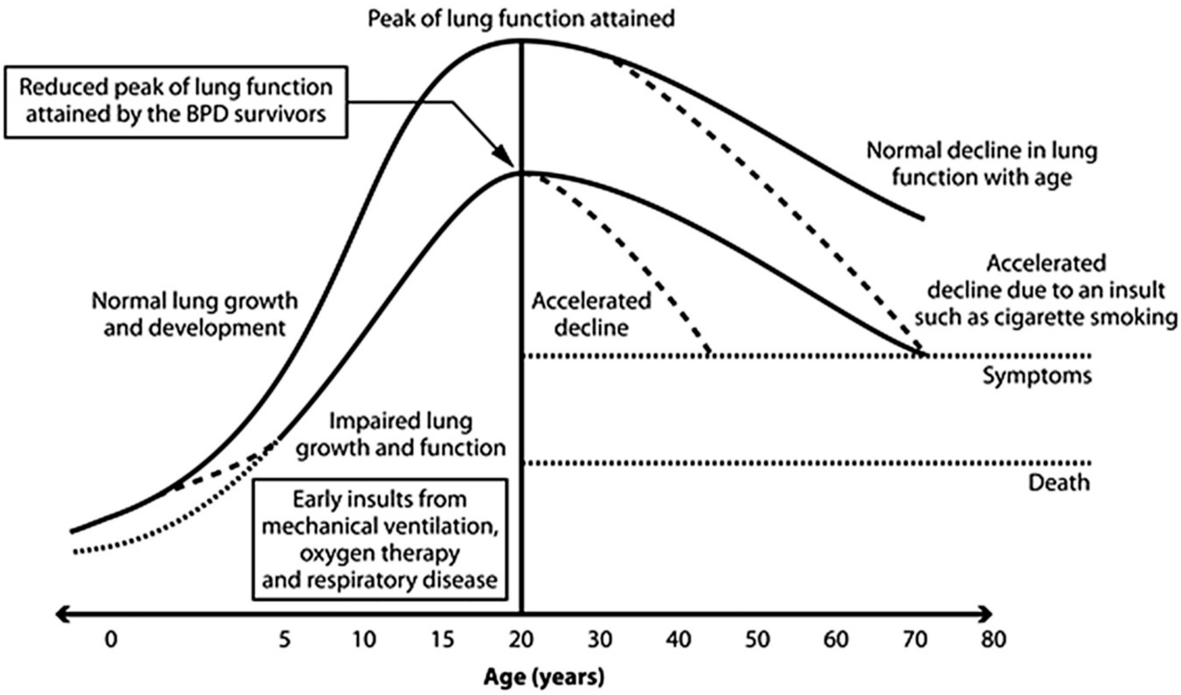


Figure 5. *Lung function decline over life. Two cohorts described in this thesis focus on different origins of COPD. In the LUNAPRE cohort we investigate preterm birth- and BPD-related phenotypes (Reduced peak of lung function), and in the Karolinska COSMIC cohort we investigated gender differences in smoking-related COPD (Accelerated decline due to an insult such as cigarette smoking). Reprinted from Semin Fetal Neonatal Med 19(2), Gibson, A. M. and L. W. Doyle. "Respiratory outcomes for the tiniest or most immature infants.". 105-111, Copyright (2014) with permission from Elsevier*

2.3.4 Factors influencing the development of IRDS and BPD

There are many factors besides gestational age and weight at birth that may affect the different degrees of symptoms of IRDS, and its subsequent development into BPD. As mentioned, administering surfactant decreases the risk of severe symptoms. Surfactant is administered into the central airways. Administering surfactant using continuous positive airway pressure (CPAP) instead of forced ventilation has enhanced the treatment of prematurely born children, resulting in a large increase in survivals [63]. An important contributor to the enhanced survival of prematurely born children is the administration of corticosteroids to pregnant women at risk of premature delivery before 34 weeks of gestation, to accelerate fetal lung maturation [64].

Not only are the lungs poorly developed in prematurely born children, so too are the gastrointestinal tract and the immune system [65]. Inflammation has been suggested as one risk factor for developing BPD. It has been shown that antenatal diagnosis of chorioamnionitis increases the risk of higher FiO_2 [66]. Probiotics show protective properties against necrotizing enterocolitis (NEC), which is a serious complication of premature birth. Combinations of probiotic strains have shown synergetic effects compared to administrating one strain alone [67]. There are indications that BPD may also be influenced by the microbiome. In a study of three-day-old prematurely born children, there was reduced diversity in the microbiome of the airways in those developing BPD [68].

2.4 CHRONIC OBSTRUCTIVE PULMONARY DISEASE (COPD)

2.4.1 Features, risk factors, and prevalence of COPD

Chronic obstructive pulmonary disease (COPD) is a disease with progressive deterioration of lung function. The main symptoms are shortness of breath and productive cough. The main risk factor for COPD is smoking, but – as described above – low birth weight and premature birth are also risk factors [69].

COPD is an umbrella diagnosis representing many different molecular phenotypes. Clinically, COPD is often subdivided into chronic bronchitis, which is more common among females, and emphysema, which is more prevalent in males [70]. Many patients have a mixture of the two phenotypes, and chronic bronchitis often precedes emphysema development. Emphysema also affects the airway structure, as the destruction of the alveolar septa may cause a collapse of the small airways, with air trapping and hyperinflation as a result. Females have also been shown to have a higher incidence of the disease [71, 72], and a faster decline in lung function especially after menopause. The decline in lung function is often expressed in exacerbations. In a Swedish study, it was shown that females had exacerbations more frequently than males [72].

The increased mortality in smokers due to COPD and other smoking-related diseases results in a loss of life expectancy of 13.2 years for males and 14.5 years for females [73]. The result is that COPD is now the fourth most common cause of death worldwide [74].

As mentioned, there are risk factors other than smoking, and in a worldwide assessment 28% of COPD patients (according to GOLD criteria) had never smoked. When considering the lower limit of normal (LLN) criteria [75], 23% of the COPD patients had never smoked [76]. In Sweden, the corresponding figures according to GOLD criteria were 14% for males and 27% among women. Overall, the prevalence for COPD was 7% among non-smokers and 24% among smokers [77].

2.4.2 Diagnosis of COPD

Diagnosis of COPD relies heavily on spirometry, and is generally defined by forced expiratory volume during one second divided by forced vital capacity (FEV_1/FVC) being below 0.70, i.e., the patient is unable to exhale 70% of their lung volume in one second. Traditionally, the severity of COPD is determined by % of predicted FEV_1.

GOLD 1 (mild): $FEV_1 \geq 80\%$

GOLD 2 (moderate): $50\% \leq FEV_1 < 80\%$

GOLD 3 (severe): $30\% \geq FEV_1 < 50\%$

GOLD 4 (very severe): FEV_1 < 30%

However, the latest update to the Global Initiative for Chronic Obstructive Lung Disease (GOLD) in 2017 [78] also emphasizes the importance of symptoms and exacerbations in staging of disease. In addition to the stage (A–D) determined by spirometry, the symptoms are also graded by breathlessness using the Modified British Medical Research Council (mMRC) Questionnaire and number of exacerbations. By using computed tomography (CT) scans, a more detailed diagnosis may be determined, including degree of disease by extent of bronchitis, emphysema, and air trapping [79].

2.4.3 Therapy for COPD

There is currently no efficacious treatment to cure or halt the progression of COPD. The pharmacological treatments available for COPD all reduce symptoms, but have no proven effect on the long-term prognosis. The pharmaceuticals used are grouped into bronchodilators, which relax the smooth muscles in the bronchi, and corticosteroids, which show anti-inflammatory effects [80].

There are two kinds of bronchodilators: β2 agonists and anticholinergics [81]. β2 agonists act by blocking the β2 adrenergic receptor (also called the β2 adrenoreceptor) from binding to epinephrine, resulting in the relaxation of smooth muscles. The long-acting β2 agonists (LABAs) include salmeterol, formoterol, indacaterol and tiotropium. There are also short-acting β2 agonists, such as salbutamol and terbutaline, which are more often used as “rescue treatments” during acute shortness of breath. Anticholinergics act by blocking cholinergic nerves by blocking acetylcholine receptors, resulting in the relaxation of smooth muscles. There are two types of acetylcholine receptors that bind to acetylcholine and muscarine (called muscarinic acetylcholine receptors) and to nicotine (called nicotinic acetylcholine receptors). An example of a long-acting anticholinergic is tiotropium, and long-acting muscarinic antagonists (LAMA) include aclidinium and umeclidium bromide, while ipratropium is short-acting.

Other maintenance treatments used in COPD are inhaled corticosteroids (ICS), e.g., budesonide. ICS have many functions, including controlling inflammation by depressing migration of leukocytes and fibroblasts and by controlling protein synthesis. They act by recruiting histone deacetylase-2 (HDAC2), which inhibits the acetylation of histones via histone acetyltransferase (HAT) and its activation of inflammatory proteins. They also bind directly to glucocorticoid response elements (GRE) in promotor regions of DNA, activating anti-inflammatory proteins [82].

Comparing different therapies in a meta-analysis, a combination of LABA and corticosteroids gave the greatest improvement in quality of life and lung function [81].

In severe COPD, supplemental oxygen treatment and eventually lung transplantation may be necessary.

2.4.4 Immunology of COPD

COPD involves both local and systemic inflammatory response. Locally in the lung, the inflammatory response is characterized by increased levels of macrophages and dendritic cells, which – together with epithelial cells – release chemokines and may attract neutrophils, monocytes, and T and B lymphocytes [83, 84]. This inflammation is further aggravated during exacerbations, which are initiated by bacterial or viral infections. Once triggered, the inflammatory cascade leading to tissue damage may continue even after smoking cessation [85].

The increased manifestation of inflammation in COPD compared to the inflammation caused by smoking alone is due to many factors [86]. A known genetic susceptibility is deficiency of α1-antitrypsin, an inhibitor of neutrophil elastase. A range of mutations in this gene has been shown to lead to the early onset of emphysema, particularly in smokers. Main players affecting the inflammation are oxidative stress and the imbalance of proteinases and antiproteinases. Another factor is age, which affects the ability to repair tissue due to smoking. The frequent activation of the immune system due to infections also increases the inflammation. This is due to the cilia being defective in COPD, the increased mucus production, bronchiectasis, emphysema space formed where bacteria can grow, and the reduced ability of the immune system to clear the infection.

Both the innate and adaptive immune systems are involved in COPD, and their role has been nicely summarized by Peter J. Barnes [83].

The first cells to react against an intruding agent are the innate immune system, consisting of macrophages, monocytes, neutrophils, dendritic cells, natural killer cells, and mast cells, all contributing to the inflammation of COPD.

Macrophages play a major role in COPD, and are the main cells found in bronchoalveolar lavage (BAL). Their numbers are increased in BAL. Macrophages release a number of inflammatory mediators, including CCL2 and CXCL1 that recruit monocytes, which upon entry into the bronchoalveolar lumen may differentiate into alveolar macrophages. CCL2 also attracts dendritic cells and memory T-cells. Macrophages produce tumor necrosis factor α (TNF-α), inducing NF-κB, which is activated in macrophages from patients with COPD [87] and regulates most of the inflammatory proteins in COPD. TGFβ is also secreted by macrophages, and is involved in fibrosis in bronchioles acting on fibroblasts. Macrophages release matrix metallopeptidases (MMP-2, MMP-9, and MMP-12), which are involved in the degradation of extracellular matrix and emphysema formation.

Finally, macrophages recruit neutrophils by releasing leukotriene B4 (LTB_4), CXCL1, CXCL5, and CXCL8. Increased numbers of neutrophils are found in the lumen and correlate to disease severity. Neutrophils secrete myeloperoxidase (MPO), lipocalin, serine proteases, cathepsin G, proteinase-3, MMP-8, and MMP-9. Neutrophils also contribute to mucus secretion stimulated by neutrophil elastase.

Epithelial cells cover the airways down to the proximal bronchioles. They are key players in COPD, but have not been studied as thoroughly as macrophages due to the more invasive methods needed to collect them. They take part in the inflammatory response by secreting many of the chemokines that are also secreted by macrophages, including TNF- α, interleukin-1β, IL-6, and CXCL-8. They also secrete granulocyte-macrophage colony-stimulating factor (GM_CSF), which stimulates stem cells to produce neutrophils and monocytes.

The adaptive immune system also takes part in the immunology of COPD. Cytotoxic T-cells (T_c) are increased more in COPD patients than helper T-cells (T_h), and the numbers correlate to the severity of the disease. CXCL9, CXCL10, and CXCL11 released by macrophages recruit T-cells to the site of inflammation by binding to CXCR3. T_c induces apoptosis in infected cells by releasing perforine, granxyme B, and TNF-α.

2.4.5 Oxidative stress

Reactive oxygen species (ROS) and reactive nitrogen species (RNS) are oxygen/nitrogen free radicals. Their function and pathways in health and disease are summarized by Valko et al. [88]. They are present in homeostasis as superoxide anion radical (O2•−), hydroxyl radical (•OH), peroxyl radical (ROO•), or nitric oxide (NO•), and are produced in the mitochondria. On exposure to cigarette smoke or other environmental toxicants, there may be overproduction of ROS and RNS, resulting in harmful oxidative stress and nitrosative stress. Peroxisomes convert ROS into hydrogen peroxide (H_2O_2), and if they are damaged the H_2O_2 leaks out to the cytosol and contributes to oxidative stress.

In oxidative stress, ROS may harm cell structures, nucleic acids, lipids, and proteins. Sulfhydryl-containing proteins are particularly susceptible to oxidation, forming mixed disulfides. However, many other amino acids may form carbonyl residues, which may react further to form advanced glycation end-products (AGEs). Our group has observed increased oxidative stress primarily in female COPD patients, as shown by upregulation of proteins involved in oxidative phosphorylation [89] and increased levels of many metabolites associated with fatty acid β-oxidation pathway [90]. One reason for the larger impact on females has been suggested to be larger downregulation of antioxidants in females [91].

2.5 ASTHMA

More than 300 million people are affected by asthma worldwide causing more than 400000 deaths related to asthma. Typical symptoms of asthma are breathing difficulties that are sometimes acute in so called asthma attacks, chest tightness and coughing.

Asthma is strongly sex dependent with more (65%) boys being affected at low ages (<13 years) and more women (65%) being affected in adulthood [92]. This indicate that asthma is sex hormone dependent. There is also an effect on asthma during menstruation, pregnancy and menopause, further supporting a role of sex hormones in asthma [93]. The mechanisms for this are not well understood.

Asthma is diagnosed by spirometry setting asthma diagnose if FEV_1 improves more than 12% and at least 200 ml post-bronchodilator [94], and the peak expiratory flow rate increases 20% by salbutamol corticosteroids or prednisone. Provocation challenge using Methacholine to assess airway hyperreactivity is also used for diagnosing asthma, if FEV_1 drops 20% or more.

Even if many asthma symptoms overlap for individuals with asthma, there is a dramatic variation in severity and response to treatment. Therefore, asthma has increasingly been described as a disease with many endotypes [6]. Several different subphenotyping efforts has proposed. The classification of asthma into allergic and non-allergic asthma, origininally called extrinsic and intrinsic asthma, was first proposed in 1947 [95]. Allergic asthma is usually developed early in life and non-allergic asthma later in life. Around 37-50% of asthma patients are affected by allergic asthma characterized by increased numbers of eosinophils [96, 97]. Non-allergic asthma is often developed later in life and affects 50-63% of asthmatics [98].

A common classification in recent years is Th2 high and Th2 low or non Th2 asthma, with Th2 high overlapping with allergic asthma while non Th2 overlap with non-allergic asthma [97, 99]. Th2 high asthma affects 80% of children with asthma and 60% of adults, and is characterized by the Th2 associated cytokines IL-4, IL-5 and IL-13 as well as type 2 innate lymphoid cells. IL-5 stimulates eosinophils [100] while IL-4 and IL-13 regulate immunoglobulin E (IgE) and activation of mast cells [101]. Corticosteroid treatment is the most common treatment to reduce inflammation in asthma and can be taken orally (OCS) or inhaled (ICS). In many patients treated with OCS the Th2 inflammation is not suppressed and 40% remain uncontrolled [102]. For Th2 low patients the CS treatment is not always effective. Even with the emerging biologics, with antibody treatments of anti-IL-4, anti-IL5 and anti-IL13 [103], and tools for stratification of patients with severe asthma is important.

Another subphenotyping that has been proposed is TAC1 (high eosinophilia), TAC2 (high neutrophilia), and TAC3 (moderate levels of eosinophilia) [104]. Yet another sub phenotyping has been proposed by Refractory Asthma Stratification Programme (RASP) to stratify severe asthma patients, that do not respond to CS treatment and are therefore using high doses, into three groups depending on adherence to treatment named RASP: The first group being no-adherent, the second group having impaired CS responsiveness even when using high doses of CS, with persistent high Th inflammation, and a third group who are non-responsive to CS but show low Th2 inflammation [105].

Despite the stratifications performed the mechanisms are not fully understood, especially not for severe asthma. Therefore, efforts are continued to stratify asthmatics further with aim to find endotypes with defined mechanisms to be able to develop therapeutics.

2.6 EXTRACELLULAR VESICLES

There are three groups of extracellular vesicles (EVs); microvesicles, apoptotic bodies and exosomes.

Exosomes are nano sized EVs with a size of 30–100 nm and are usually referred to as small extracellular vesicles (sEVs). Unlike micro vesicles that bud directly from the cell membrane, sEVs are formed in multivesicular bodies (MVB), also termed late endosomes, and are then actively secreted into the extracellular space [106]. They are found in all body fluids including blood [107, 108], saliva [109], urine [110], amniotic liquid, seminal fluid, breast milk [111], BALF [112-114], and CSF. They are released from many different cell types, including stem cells, tumor cells [115], epithelial cells [116], BAL cells [113], B-cells [112, 117], t-cells [118], and dendritic cells [119, 120]. SEVs have been shown to carry a cargo of different molecules including proteins, lipids, mRNA, and miRNA, and can transfer this content from one cell to another, enabling communication between cells [121].

There are many indications that both the formation of sEVs and the uptake of sEVs by cells are active processes, enabling communication with distant cells through sEVs. A gene TSAP6 (downstream gene of p53) seems to be necessary for both p53 and sEV production [122]. Endosomal sorting complex required for transport (ESCRT) is responsible for sorting proteins into the intraluminal vesicles which are the precursors of sEVs [123]. Rab, a family of membrane proteins, regulates sEV secretion with Rab 7 sorting them into lysosomes and degradation while Rab 11 sorts them into the extracellular space [124, 125]. Selective fusion of vesicles with the correct acceptor membrane is mediated by the SNARE proteins. The v-SNARE on the sEV has to match the t-SNARE on the recipient cell before they fuse [126].

SEVs consist of a lipid bilayer which contains many types of molecules. Both surface and cargo molecules may vary depending on disease and origin [127]. However, some molecules appear to be generically enriched in sEVs. Most sEVs have the tetraspanins; CD63, CD81, CD9, and heat shock protein (Hsp70), and may thus be used as crude EV markers. EV content also varies depending on diagnosis, including lung cancer [128], autoimmune disease [129], cystic fibrosis [130], and sarcoidosis [127] and asthma [113]. Also, the number of sEVs secreted may be altered in diseased patients compared to healthy individuals, as shown for sarcoidosis [127].

SEVs from BAL cells have been shown to contain both MHC class I and II molecules together with co-stimulatory molecules [114, 117]. This shows that sEVs may both inhibit [131] and stimulate the immune system, as has also been shown in clinical studies [132-134]. SEVs from B lymphocytes induce T-cell proliferation [112, 117] and may thus be involved in inflammatory and allergic immune responses. SEVs from macrophages and DCs contain functional enzymes for leukotriene biosynthesis which are proinflammatory lipid mediators with a role in asthma and inflammation [135]. BALF sEVs have also shown significant upregulation in sarcoidosis of inflammation-associated proteins, such as leukotriene A4 hydrolase, together with vitamin D-binding protein [136]. Epithelial cell sEVs induce proliferation and chemotaxis of undifferentiated macrophages [137].

SEVs may also contain mRNA and miRNA, which are interesting for their ability to regulate gene expression in their target cells [138]. They will be discussed further in the next section.

The effect of sEVs on the target cells may be either pathological or beneficial. In a study of pulmonary hypertension (PH) in mice, sEVs that were isolated from sick animals and injected into healthy mice induced PH changes, while mesenchymal stem cells (MSC-EXOs) prevented and reversed PH in the same animal model [139, 140]. In another study sEVs from COPD patients were given to mice and the neutrophil elastase carried by the sEVs caused destruction of extracellular matrix in the mice [141].

2.7 MIRNA

Non-coding RNA (ncRNA) is RNA that is not translated into a protein. The first ncRNA, transfer RNA (tRNA), was characterized in 1965. Since then, a whole range of ncRNAs have been identified, including ribosomal RNA (rRNA) and small RNAs. Small RNAs are less than 200 nucleotides and include small interfering RNA (siRNA), microRNA (miRNA), Piwi-interacting RNA (piRNA), small nucleolar RNA (snoRNAs), small rDNA-derived RNA (srRNA), small nuclear RNA (U-RNA), and tRNA-derived small RNA (tsRNA) [142]. They are involved in RNA silencing. One way of silencing is RNA interference (RNAi), which is the ability of RNA to inhibit gene expression by degrading or inhibiting mRNA – a discovery for which Andrew Fire and Craig C. Mello received the Nobel Prize in Physiology or Medicine in 2006. RNA that may be involved in RNA interference are small interfering RNA (siRNA) and microRNA (miRNA), both of which are around 20–25 nucleotides in length. SiRNA is formed from exogenous double stranded RNA (dsRNA), while miRNA is formed from endogenous pre-miRNA (around 70 nucleotides in length). Pre-miRNA is cleaved by dicer-TRBP complex into miRNA duplex and further formation of mature miRNA via the RNA-induced silencing complex RISC complex and Ago2 [143]. Highly matched sequences and tight binding of the miRNA to the target mRNA strand results in degradation of the mRNA, while an imperfect match results in the repression of translation [144]. MiRNA are pluripotent, with each miRNA binding in average to more than 500 target sites on mRNA, and each mRNA may be inhibited by many different miRNAs [145].

2.7.1 MiRNA cargo in EV in the lung

MiRNA may be found extracellularly, with reportedly more than 90% bound to Ago2 in plasma [146]. Some of it may also be found bound to high-density lipoproteins (HDL) and some to apoptotic bodies, but most interest has been paid to the miRNA found in or on EVs, including sEVs. The sEVs have been suggested to protect the miRNA from degradation by RNases, but extracellular miRNA has also been found to be surprisingly stable compared to RNA, both when bound to Ago2 [146] and when found in sEVs [147]. MiRNAs are believed to be selectively packed into sEVs through an intricate molecular machinery found to be regulated by the YBX1 protein [148, 149], among others. The secretion and uptake of the sEVs by the recipient cells is tightly regulated. Different mechanisms for this selective extracellular export of miRNA have been suggested to involve differences in sequence [150] or carriers of miRNA [151]. The difference in miRNA content between cells of origin and with respect to diagnosis make them interesting as biomarkers of disease.

In the lungs, sEVs have been found in BAL fluid [114] and pleural fluid [152]. Our group has previously shown that the miRNA content of sEV-enriched EVs from BAL fluid differed in mild intermittent asthma compared to healthy subjects. Significant differences were detected for 24 miRNAs including the let-7 and miRNA-200 families, many of them implemented in the IL-4 and IL-13 cytokine pathways [113]. Fujita et al. have also shown that in vitro stimulating human bronchial epithelial cells (HBEC) with cigarette smoke extracts upregulated cellular and EVs miRNA-210 and downregulated let-7. The same group also showed that the EVs induce myofibroblast differentiation in primary lung fibroblasts (LFs), and that miRNA-210 downregulates autophagy via ATG7 which is an essential component of autophagy [153]. Alterations in the EV contents of miRNA have been found in sputum from COPD patients, and several miRNAs were associated with FEV_1, a measure of disease severity. Differences in EV miRNA content have also been observed in lungs from deceased newborns with BPD compared to normal preterm or term infants [154].

2.8 OMICS ANALYSIS

To analyze the phenotype, a large range of platforms can be analyzed including genes, transcripts, proteins, metabolites, immune cells, and microbiota, as well as data from questionnaires and other clinical data. Several methods can be used to quantify these analytes. Gene entities that can be analyzed include DNA methylation and histone modification. Analyzing transcripts includes messenger RNAs (mRNAs), long noncoding RNAs (lncRNAs), or small RNAs such as microRNAs (miRNAs) and small interfering RNA (siRNA). Proteins can be analyzed using mass spectrometry or antibodies. A range of different metabolites can be analyzed including lipids, sugars, and cytokines. These analytes can also be analyzed in a large range of compartments including different organs, tissues, blood, immune cells, or extracellular vesicles (EVs). In recent years, the sensitivity of analysis has advanced to the degree that analyzed compartments may also consist of single cells rather than bulk.

2.9 PATHWAY ENRICHMENT ANALYSIS

A molecular pathway describes a chain of molecular reactions and how these reactions are regulated. All genes connected to the molecules involved in a pathway are targeted in a pathway analysis. Extensive efforts to construct large databases that consolidate the known scientific literature into consensus pathways, such as KEGG [155] and Reactome [156], have been an essential step towards facilitating pathway enrichment analysis.

Pathway analysis is a means of investigating how genes affect pathways in the body. The altered pathways can explain the mechanism of a disease. There are different ways to perform pathway analysis including over-representation analysis (ORA) and functional class scoring.

The best-known method for functional class scoring is gene set enrichment analysis (GSEA) [157].

ORA tests whether a functional gene set (FGS) is overrepresented compared to what is expected with random chance, taking into account which genes are generally altered. These genes are gathered in altered gene sets (AGS), which can be entered as background in ORA. The probability that the pathways are enriched is expressed using the statistical test Fisher's exact test, resulting in a p-value for the overrepresentation.

GSEA does not use an AGS. Instead, GSEA uses the genes in the analysis under investigation as reference. It compares every single gene to a statistic calculated from the genes in the FGS. As the genes are sorted from the most to the least sorted, it uses the genes in the middle of the list and assumes those to be stable. These comparisons are then aggregated into a statistic for each pathway, and the significance for each pathway is assessed.

There are many tools for performing ORA and GSEA, including Ingenuity Pathway Analysis (IPA) from Qiagen and open resources such as Cytoscape. Tools such as miEAA [158] provide both ORA and GSEA tailored for miRNA analysis. Diana, mirPath, and miRWalk also providing tools, and there are many pathway analysis tools in R including miRLAB, miRNApath, and pathfinder.

2.10 NETWORK ANALYSIS

Network analysis in omics research is a means of analyzing the connections between different biomolecules and subjects and thereby creating a network to get a comprehensive view of a given biological process. Usually, a network is based on clustering of variables.

Similarity Network Fusion (SNF) on the other hand is based on clustering of subjects into groups [159]. It can be used to combine several different omics platforms. SNF is a non-linear method to integrate data in an unsupervised method. Similarity network fusion has been used for multi-omics integration of COPD patients in the COSMIC cohort. It has been shown that using multi-omics data it was possible to reduce the group size from 30 to 6 and still keep 95% power and 95% accuracy using seven omics platforms [160].

Another way to combine multi-omics platforms is to create hierarchical models. Models are first created, for example using OPLS, for each omics dataset. The resulting scores from each model are used to create new models, so called a hierarchical model. Further expansions to allow correlation of multiple data blocks have also been developed, as in the O-n-PLS algorithm [20]. [161]

3 RESEARCH AIMS

The aim of this thesis was to investigate mechanistic differences between sub-groups of subjects in the umbrella diagnoses of asthma and chronic obstructive pulmonary disease (COPD). The cohorts studied were collected with the aim of applying systems biology approaches to study the lungs using a large number of platforms.

Paper I. The aim of this project was to develop a method for generating pairwise OPLS-DA models with variable selection reproducibly in an automated fashion, including tools for discriminating models being true positives from models at risk of being a produced from random variability (i.e., false positives).

Paper II. The aim of this project was to analyze alterations in the miRNA cargo of sEV-enriched EVs isolated from bronchoalveolar lavage fluid from patients with COPD from the Karolinska COSMIC cohort, as well as how molecular pathways are regulated by the affected miRNAs. The long-term goal of the overall project is to better understand the mechanisms of COPD related to smoking and gender.

Paper III. The specific aim in this study was to investigate the composition of T-cell subsets in the bronchoalveolar lumen from subjects from the LUNAPRE cohort, which is designed to investigte the mechanisms of early onset obstructive lung disease related to preterm birth and bronchopulmonary dysplasia.

Paper IV. The aim of this study was to perform data-driven clustering of asthma patients from the U-BIOPRED cohort based on multi-omics data from blood and urine in order to identify molecular sub-phenotypes. The overall goal was to identify the biomolecules that drove the clustering, as well as to identify clinical characteristics between the subgroups.

Paper V. The aim of this project was to investigate the urinary metabolome from individuals with severe-to-mild asthma from the U-BIOPRED cohort. A secondary goal was to determine how the metabolome was affected by the use of oral corticosteroids.

Paper VI. The aim was to study the stability and degradation of the proteome of EDTA-plasma samples due to delays in centrifugation and separation of plasma during biobanking. The overall goal was to identify biomarkers of degradation to be able to determine the delay in processing that occurred in samples that withdrawn from biorepositories.

4 MATERIALS AND METHODS

4.1 COHORTS

4.1.1 The MTBLS136 data set (studied in paper I)

The dataset analyzed in paper I, MTBLS136, came from the Cancer Prevention Study II Nutrition Cohort, which was designed as a prospective study of cancer with 39 376 subjects. MTBLS136 consists of postmenopausal women, and was designed to study differences in metabolite content in serum associated with hormone treatment. The dataset is publicly available at Metabolights https://www.ebi.ac.uk/metabolights/MTBLS136/files. The women were aged 50 to 74. The metabolomics data set includes consists of 782 cancer cases and 782 controls, with 17 being both cases and controls as they were diagnosed with cancer, giving a total of 1547. After excluding those that did not meet the inclusion criteria, 1336 women – 667 did not receive any hormone therapy (nonusers), 332 receiving estrogen therapy only (E-only users), and 337 receiveing both estrogen and progestin treatment (E+P users) – were included in this evaluation.

4.1.1.1 Ethical approval

The dataset is a publicly available, deidentified dataset at Metabolights CPS-II Nutrition Cohort that has been approved by the Emory University Institutional Review Board (IRB00045780).

4.1.2 The LUNAPRE cohort (studied in paper III)

The full LUNg obstruction in Adulthood of PREmaturely born (LUNAPRE) cohort (http://clinicaltrials.gov/ct2/show/NCT02923648) has been described in detail in 2019 [162]. It consists of 26 preterm-born young adults with a history of bronchopulmonary dysplasia (BPD), and has three control groups: one group with 23 preterm-born young adults without BPD, one group with 23 asthmatics, and one group with 24 healthy controls. Of these, 22 in each group and 24 of the healthy controls underwent bronchoscopy, collecting bronchoalveolar lavage and airway epithelial brushings. Their ages ranged from 18.2 to 23.8, with the two preterm groups being around one year younger on average. They were all non-smokers and without respiratory tract infections at least three months prior to investigation.

4.1.2.1 Ethical approval

The study has been approved by the Swedish Ethical Review Authority (ref: 201211872-31/4).

4.1.3 The KAROLINSKA COSMIC cohort (studied in paper II)

The Karolinska COSMIC (Clinical & Systems Medicine Investigations of Smoking-related COPD; ClinicalTrials.gov NCT02627872) cohort was collected at the Karolinska University Hospital, and was designed to study gender differences in protein expression in cells from the airways in patients with mild -to-moderate COPD compared to healthy subjects and smokers with normal lung function. The cohort consists of 120 subjects in three groups: 40 subjects with COPD, 40 smokers without COPD, and 40 healthy never-smokers, with the same numbers of males and females. The inclusion criteria were that those with COPD and smokers with normal lung function had been smoking for more than ten pack-years and had smoked more than ten cigarettes per day for the past six months. COPD ex-smokers, on the other hand, had quit smoking more than two years before inclusion in the study. To avoid acute effect of smoking, the COPD patients and the healthy smokers were asked not to smoke for a period of eight hours before the bronchoscopy, which was confirmed by measuring exhaled CO [163]. Exclusion criteria included history of allergies or asthma, and use of corticosteroids within the past 3 months, or exacerbations during a period of three months before inclusion in the study allowed. The subjects were examined thoroughly using chest radiography, computed tomography, spirometry, and questionnaires. Only COPD patients with mild to moderate disease (GOLD stages I–II/A-B) were included in the study. Details of the full cohort are described by Karimi et al. [164].

BAL fluid collected by bronchoscopy in the COSMIC cohort was used for paper II. To avoid sampling of the proximal airways instead of distal, subjects with a recovery of BAL <35% were excluded [165]. In addition, five samples with less than 85% macrophages were excluded. Finally, samples with clear experimental problems were removed from further analysis. This resulted in 19 smoking and six ex-smoking COPD subjects, 21 smokers with normal lung function, and 20 healthy never-smokers, all aged 44–66. Details of the subgroup of subjects from the cohort included in this study are presented in Table 1 in paper II.

4.1.3.1 Ethical approval

The study was approved by the Stockholm Regional Ethical Board (ref. 2006/959-31/1, 2006/959-31/1, 2007-743-32, 2007/748-31/3, 2008/600-32, 2009/1358-32, 2010/1064-32, 2011/1322-32).

4.1.4 The adult U-BIOPRED cohort (studied in papers IV and V)

The U-BIOPRED (Unbiased BIOmarkers for the Prediction of REspiratory Disease outcomes) cohort is a multicenter collaboration originally led by Professor Sterk in Amsterdam University, now headed by Sven-Eric Dahlén at the Karolinska Institute and Ian Adcock at Imperial College London. The cohort was collected to integrate physiological and clinical variables, as well as in depth multi-omics characterizations from collected from blood, urine, induced sputum, bronchial biopsies, airway epithelial brushings, and exhaled air, in total generating 18 omics data sets, in order to be able to sub-phenotype asthmatics, with the final goal of identifying therapeutics that have an effect on different endotypes of asthma. The cohort consists of 110 smoking/ex-smoking severe asthmatics, 311 non-smoking severe asthmatics, 88 mild to moderate asthmatics, and 101 healthy controls, including both males and females.

The inclusion criterion for the ex-smokers with severe asthma was that they had not smoked for 12 months and had a smoking history of at least five pack-years. Non-smoking severe and mild asthmatics and healthy controls had not smoked for 12 months and had less than five pack-years of smoking history. Severe asthmatics had uncontrolled symptoms according to the Global Initiative for Asthma (GINA) and/or more than two exacerbations per year despite a high dose of inhaled corticosteroids (⩾1000 µg fluticasone propionate or equivalent per day). Mild to moderate asthmatics had controlled asthma according to GINA and used <500 µg fluticasone propionate or equivalent per day.

Thorough investigation included bronchoscopy and telemonitoring sessions, spirometry, measuring fraction of exhaled nitric oxide level (FeNO), and testing allergic status. Induced sputum was performed, collecting supernatants and cell pellets to obtain eosinophil and neutrophil counts. Exhaled breath was obtained to measure metabolites. Blood and urine were collected for lipidomic, proteomic, and transcriptomic analysis. Some subjects underwent genetic analysis, plethysmographic measurements, and high-resolution lung computed tomography. The subjects were reinvited 12–18 months after baseline visit to assess longitudinal data.

The subjects were assessed thoroughly using a large number of questionnaires, including the Asthma Control Questionnaire (ACQ5), the Asthma Quality of Life Questionnaire (AQLQ), the Hospital Anxiety and Depression Scale (HADS), the Sino-Nasal Outcomes Test (SNOT20), the Epworth Sleepiness Scale (ESS), and the Medication Adherence Report Scale (MARS).

Details of the cohort are described by Shaw et al. [166].

4.1.4.1 Ethical approval

2011/1254-31/3.

4.1.5 Biobank project

The cohort studied in paper VI consisted of 16 healthy controls, with an even gender distribution. Whole blood samples were drawn into EDTA-plasma tubes. To study the stability of proteins in the samples, they were stored for one, three, eight, 24, and 36 hours at 4°C and 22°C before centrifuging and separating plasma.

Details of the cohort are described by Shen et al. [167].

4.1.5.1 Ethical approval

2011/341-31/3, addition 2013/703-32.

4.2 SAMPLING THE LUNG

4.2.1 Bronchoscopy

In both paper II and paper III, bronchoscopy was used to collect samples. Bronchoscopy is a procedure to collect samples from the lung to facilitate the study of inflammation in the lung at the site of injury. During bronchoscopy, a flexible fiber optic tube is inserted through the nose to the middle-lobe bronchus. The distal airways are rinsed using phosphate-buffered saline (PBS) and the recollected bronchoalveolar lavage (BAL) fluid contains BAL cells which are immune cells including mostly macrophages, but also lymphocytes, neutrophils, eosinophils, and mast cells. The cell differentials vary between diagnosis, and may thus be used as diagnostics for lung diseases.

The BAL fluid also contains EVs, including sEVs. As these can both inhibit and stimulate immune responses and their content varies depending on diagnosis, they are also interesting to study.

Epithelial cells (BEC) can be collected by taking brushings during bronchoscopy. They are part of the immune system and release chemokines to attract immune cells to the lung.

Lung tissue can also be collected using biopsies to study lung tissue, which can be examined through microscopy. Using biopsies in the COSMIC study showed alterations of Mucin (MUC) macromolecules in basal cells and goblet cells and of epithelial growth factor receptor (EGFR) in basal cells and ciliated cells [168].

4.3 PAPER I

4.3.1 OPLS modeling using SIMCA

SIMCA is a tool for multivariate analysis produced by Umetrics (now part of Sartorius). It uses a graphical user interface with user friendly workflow which ensures quick startup in modeling. However, when performing modeling more frequently this becomes tedious due to the manual repeated procedures. One of the multivariate methods used in SIMCA is OPLS discussed in section 2.1.8. SIMCA was used to produce OPLS-DA models for paper VI, paper II, and at the beginning of paper IV.

When creating OPLS models in SIMCA the maximum missing values can be set which in this thesis is set to 25%. PCAs to analyze for detecting outliers and also to verify that the normalization is appropriate.

4.3.2 Initiating the roplspvs project

When comparing clinical differences between cluster 16 groups in SIMCA, some models that had been deemed significant using SIMCA CV- ANOVA and permutations post variable selection even if models pre variable selection were deemed insignificant using permutations sans variable selection. It was therefore suspected that the models could have random variance. To investigate whether the models were produced from random variance, they were permutated over variable selection in SIMCA. Since the procedure of permutations over variable selection is very time-consuming in SIMCA, only five permutations were performed for five comparisons. This showed that four of the comparisons had permutated models with higher R^2 than the unpermutated model, and three had Q^2 that was higher indicating that the models might be produced from random data. To be able to create more permutations, a script in R that performed permutations over variable selection was created. This resulted in the package and workflow named roplspvs.

4.3.3 The roplspvs workflow

The roplspvs consists of a workflow that runs the roplspvs function which renders rmarkdown files to create html files. One html file is created for each comparison consisting of score plots and p(corr) plots that describe the predictive component separating the groups. In addition, figures are produced describing the optimization of the model and the significance of the models. Parameters are added to configure files that are separated into basic required settings and advanced settings containing the default settings. The identified variables containing potential biomarkers are returned as text files.

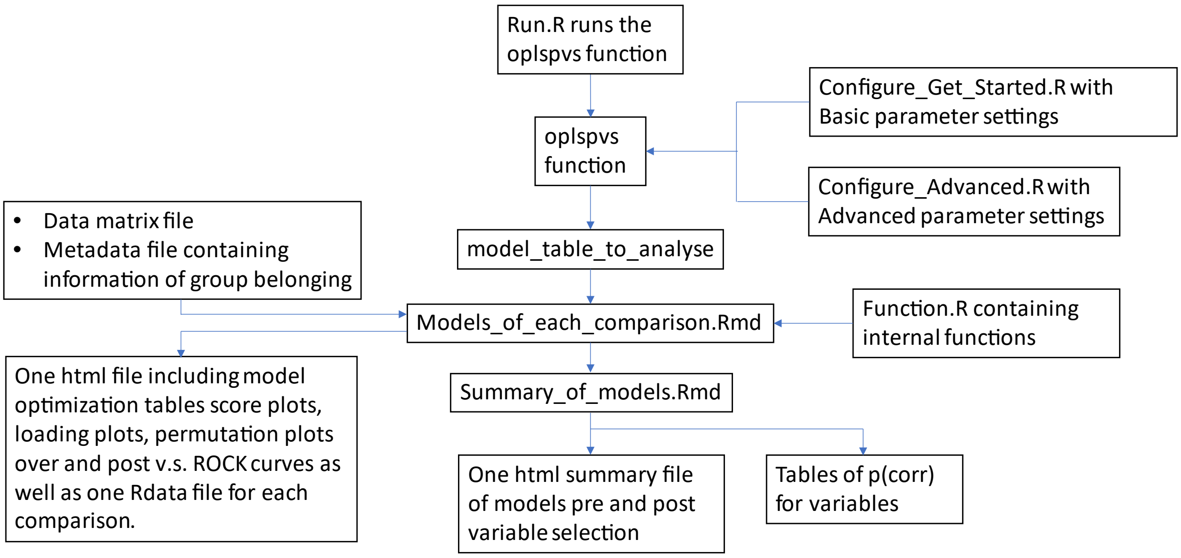

Run.R runs the oplspvs function
oplspvs function
Configure_Get_Started.R with Basic parameter settings
Configure_Advanced.R with Advanced parameter settings
• Data matrix file
• Metadata file containing information of group belonging
model_table_to_analyse
Models_of_each_comparison.Rmd
Function.R containing internal functions
One html file including model optimization tables score plots, loading plots, permutation plots over and post v.s. ROCK curves as well as one Rdata file for each comparison.
Summary_of_models.Rmd
One html summary file of models pre and post variable selection
Tables of p(corr) for variables

Figure 6. *Workflow of roplspvs uses rmarkdown to produce one html file for each comparison, one summary file summarizing all model statistics and one text file containing tables of selected variables for all model strategies models of all comparisons.*

4.3.4 Pre-process

The dataset was pretreated according to the schedule in Figure 7. LLD is the lowest amount that can be detected during an analysis. According to FDA, LLD should be determined by measuring blanks and adding three times standard deviation, which should be at least three times the blank. LLQ is the amount that may be quantified, and should be at least ten times the blank. If selecting to replace 0 or NA with LLD, roplspvs calculates LLQ as the lowest value in the dataset after removing 0 and NA. LLD is calculated by dividing LLQ by three. This includes optionally replacing 0 with NA or lower limit of detection (LLD) and replacing

NA with LLD. A userset value for LLD may be chosen. Data filtering is performed, allowing for a userset missing value tolerance for each group in each comparison.

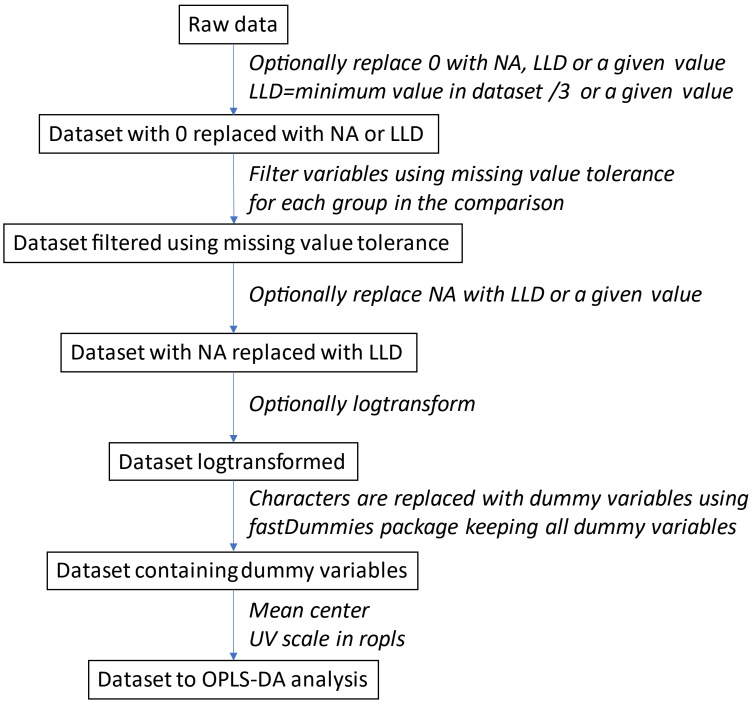

Raw data
Optionally replace 0 with NA, LLD or a given value
LLD=minimum value in dataset /3 or a given value
Dataset with 0 replaced with NA or LLD
Filter variables using missing value tolerance
for each group in the comparison
Dataset filtered using missing value tolerance
Optionally replace NA with LLD or a given value
Dataset with NA replaced with LLD
Optionally logtransform
Dataset logtransformed
Characters are replaced with dummy variables using
fastDummies package keeping all dummy variables
Dataset containing dummy variables
Mean center
UV scale in ropls
Dataset to OPLS-DA analysis

Figure 7. *Preprocessing steps in roplspvs including 0 value and missing data (NA) handling, filtering of NAs, transformation, mean centering and unit variance (UV) scaling.*

4.3.1 OPLS modeling using roplspvs

PCA and OPLS modeling is performed using the R package ropls by Thevenot [169]. Outliers can be controlled for in the PCA plots included in the html. Roplspvs applies the ropls default setting using NIPAL (Nonlinear Iterative Partial Least Squares) for missing data if NAs are not filtered away or replaced by LLD. NIPAL is a method developed by Wold in 66 [170] to impute a value for missing data. It builds on PCA and PLS, using an iterative procedure until it converges to a value for the missing data. Mean centering subtracts the average of all subjects from each data point. Unit variance (UV) scaling involves dividing each data point by the standard deviation (SD) of all subjects. The UV scaling allows variables with low abundance to contribute to the model equally with high abundance variables. If applying UV, it is therefore important to filter out variables lower than LLD to make sure models are not found that build on data from the blank.

The roplspvs package creates models even if the Q2 is negative enabling the models to be created continuously without interruptions.

4.3.2 Variable selection

The features that contribute the most to the model are extracted using p(corr) or optionally also using VIP. P(corr) of a variable is the Pearson correlation between the raw data and the scores of the model, i.e., a measure of how well a variable correlates with the model. The cutoff for p(corr) is set by the user either to a value or to correspond to a p-value for the Pearson correlation. VIP is a relative measure of how much the variable contributes to the model. If VIP is larger than one, the variable contributes more than average to the model.

If no model is created due to too high p(corr) value resulting in no variables being selected, roplspvs returns NA.

4.3.3 Five model strategies

The five different model strategies in the package have different goals (Figure 8). The first modeling strategy enables the user to select the number of orthogonals and p(corr) cutoff. This allows to use the package to test the significance using permutations over variable selection on models optimized using other tools roplspvs. It also allows the user to investigate a model that has not been optimized. When model parameters are not optimized, this removes the bias in the permutation over variable selection of having the tested model using optimized p(corr) cutoff and the permutated models being less optimized by using the same p(corr) cutoff.

Choosing a p(corr) cutoff corresponding to a p(pearson) cutoff in modeling strategy 2, includes all variables that are significantly correlated to the model. Having a larger selection of variables is useful for pathway analysis, as finding more genes related to a pathway increases the power to identify an altered pathway.

For other purposes, it is often desirable to extract the variable that builds a model that is most predictive. This is achieved in modeling strategy 3. Optimizing the predictability by maximizing Q^2 is a means of finding the variables that are most likely to be able to discriminate the groups in a new dataset. This is valuable in the search for biomarkers for diagnosis and therapeutics. At the same time as Q^2 is kept high, the difference between R^2 and Q^2 is kept low to avoid selecting an overfitted model. Adding too many orthogonals results in R^2 increasing without Q^2 increasing, resulting in overfitting of the model. Here, we aim – if possible – to keep R^2–Q^2 below 0.2 and p[R^2 and Q^2 permutated post variable selection] is kept low. A detailed description of how this is performed is shown in Figure 9.

For some datasets, modeling strategy 3 still results in a large number of variables, and modeling strategy 4 can be used to reduce the number of variables even further. The variable pcorr_diff describes the minimum increase in p(corr) cutoff that can allow a 1% decrease in Q^2 using the formula

Δpcorr / (ΔQ2/Q2*100)< pcorr_diff

This results in a model and fewer variables but a slightly lower Q^2 than modeling strategy 3.

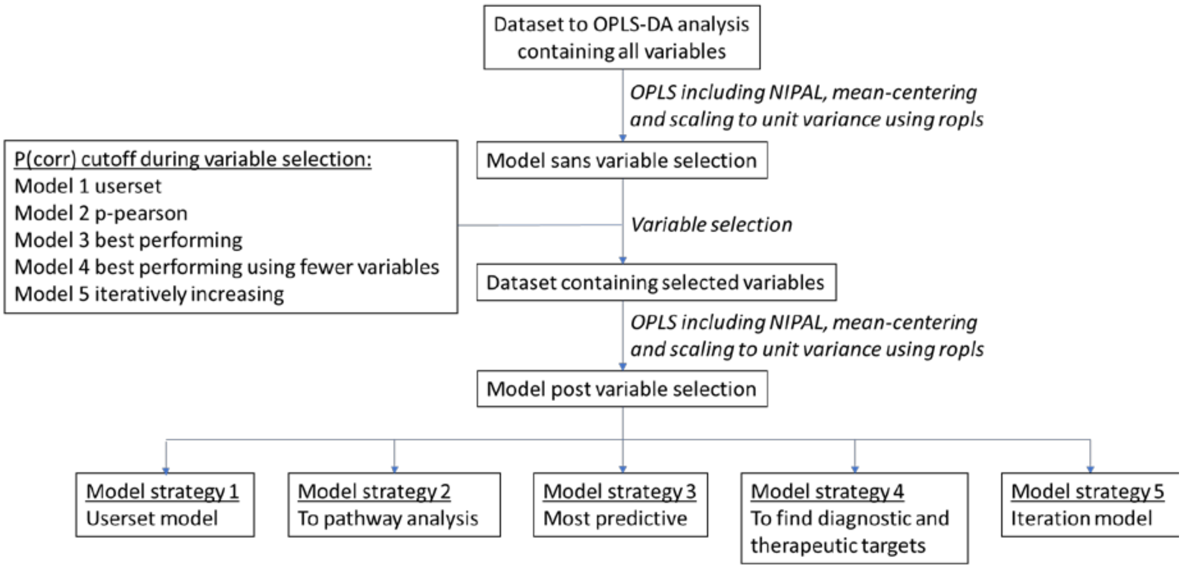

Dataset to OPLS-DA analysis containing all variables
OPLS including NIPAL, mean-centering and scaling to unit variance using ropls
Model sans variable selection
P(corr) cutoff during variable selection:
Model 1 userset
Model 2 p-pearson
Model 3 best performing
Model 4 best performing using fewer variables
Model 5 iteratively increasing
Variable selection
Dataset containing selected variables
OPLS including NIPAL, mean-centering and scaling to unit variance using ropls
Model post variable selection
Model strategy 1
Userset model
Model strategy 2
To pathway analysis
Model strategy 3
Most predictive
Model strategy 4
To find diagnostic and therapeutic targets
Model strategy 5
Iteration model

Figure 8. *Overview of the five modeling strategies in roplspvs. The p(corr) cutoffs used in each modeling strategy are shown.*

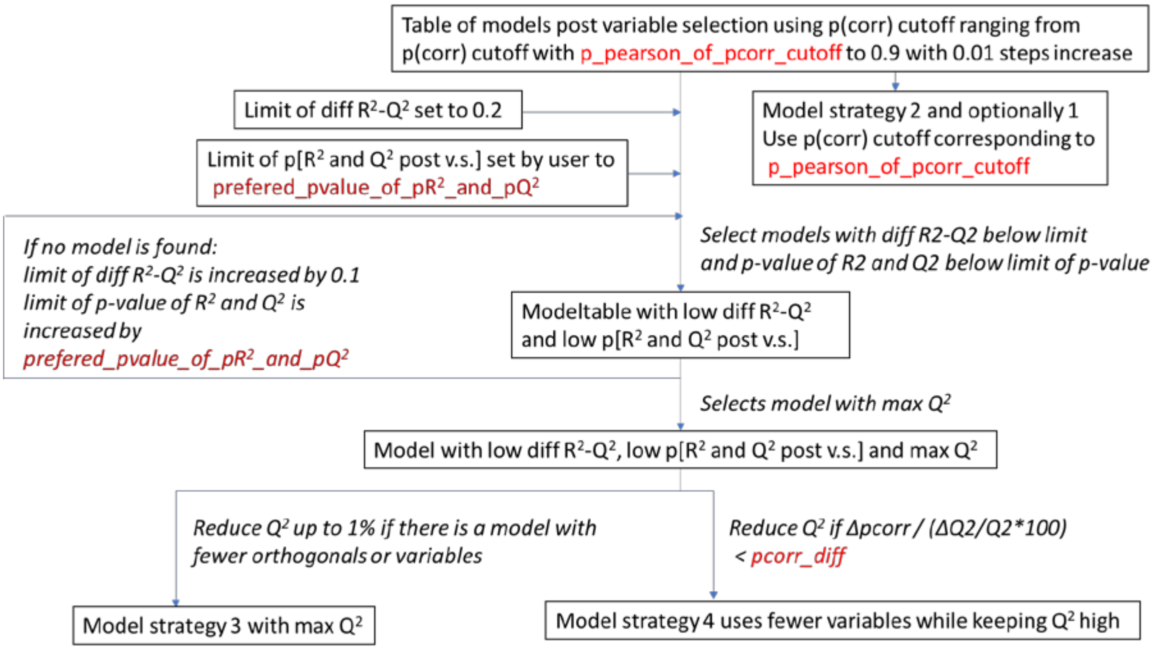


Table of models post variable selection using p(corr) cutoff ranging from p(corr) cutoff with p_pearson_of_pcorr_cutoff to 0.9 with 0.01 steps increase
Limit of diff R²-Q² set to 0.2
Limit of p[R² and Q² post v.s.] set by user to prefered_pvalue_of_pR²_and_pQ²
Model strategy 2 and optionally 1
Use p(corr) cutoff corresponding to p_pearson_of_pcorr_cutoff
If no model is found:
limit of diff R²-Q² is increased by 0.1
limit of p-value of R² and Q² is increased by prefered_pvalue_of_pR²_and_pQ²
Select models with diff R2-Q2 below limit and p-value of R2 and Q2 below limit of p-value
Modeltable with low diff R²-Q² and low p[R² and Q² post v.s.]
Selects model with max Q²
Model with low diff R²-Q², low p[R² and Q² post v.s.] and max Q²
Reduce Q² up to 1% if there is a model with fewer orthogonals or variables
Reduce Q² if Δpcorr / (ΔQ2/Q2*100) < pcorr_diff
Model strategy 3 with max Q²
Model strategy 4 uses fewer variables while keeping Q² high

Figure 9. *Details for choosing p(corr) cutoff for modeling strategy 3 and 4 in roplspvs with userset parameters set in red.*

4.3.4 Number of orthogonals

In early versions of the ropspvs script, the number of orthogonals were optimized using a similar procedure to when choosing p(corr) cutoff. Q^2 post variable selection was optimized, keeping the difference between R^2 and Q^2 as well as p[perm post v.s.] low. Using these settings in the small sample investigation gave the same results as using the ropls default settings for the number of orthogonals, indicating that the overfit observed in the small sample size study was not due to overfit of the model pre variable selection. However, to avoid overfitting of model pre variable selection, the default settings of the orthogonals were applied in the package. Orthogonals optimized post v.s. were used in models in paper III (the LUNAPRE project) and in paper V (the U-BIOPRED carnitines project). Orthogonals optimized post v.s. were used for modeling focused subjects and default ropls setting were used for modeling all subjects in paper IV (the U-BIOPRED handprint project).

4.3.5 Permutations

Model validation is performed using permutation sans, post, and over variable selection (Figure 10).

The permutation procedure is initiated by randomizing the group labels of the subjects. This creates a new randomized dataset. Fitting a model using this randomized dataset produces a model created from random variation in the actual dataset to be tested. The number of permutations is set by the user. Finally, the statistics of the permutated models are compared to the statistics of the model to be validated. The percentage of a model statistic of the permutated models that is better than the model statistic of the model to be validated represents a non-parametric method for determining the p-value of that model statistic. This gives the p-value of R^2 and Q^2 permutated without variable selection termed p[R^2 perm. sans v.s.] and p[Q^2 perm. sans v.s.].

The permutations can also be performed by randomizing the group labels of the dataset containing the variables after variable selection, creating permutated models post variable selection. Comparing these permutated models to the model to be validated results in p-values of R^2 and Q^2 permutated post variable selection, p[R^2 perm. post v.s.] and p[Q^2 perm. post v.s.].

Using the permutated datasets sans variable selection and performing variable selection on each permutation results in different variables being selected for each permutation. Fitting models to these permutated selected datasets enables us to find permutated models over variable selection. By comparing those permutated models over variable selection with the model to be validated post variable selection, a p-value over variable selection can be established. This verifies that the variable selection procedure results in a model that is significantly better than a model post variable selection produced from random variance in a dataset.

Sometimes the randomized labels happen to be the correct settings for the labels by chance. If this happens more often than the significance level, the model to be validated is deemed insignificant also when otherwise deemed significant. The smaller the sample size, the greater the risk of having the correct label setting at random. In some tools the randomly correct labeled permutations are removed, but here the permutations are performed without removing the correct labels. To control for having the correct labels at random, the permutation plot or table can be studied. The permutation plot shows the correlation between the permutated R^2 and Q^2, and the correlation coefficient between permutated subject labels and unpermutated subject labels. The correct labels have a correlation between permutated and unpermutated subject labels that equals 1. It should also be verified that the correlation in the plot is positive, with higher R^2 and Q^2 for models that have stronger correlation between permutated and unpermutated subjects. Permutation plots over variable selection are added by roplspvs in addition to the ropls permutation plots of sans and post variable selection. An example of a permutation plot over variable selection is shown in Figure 7 paper 1.

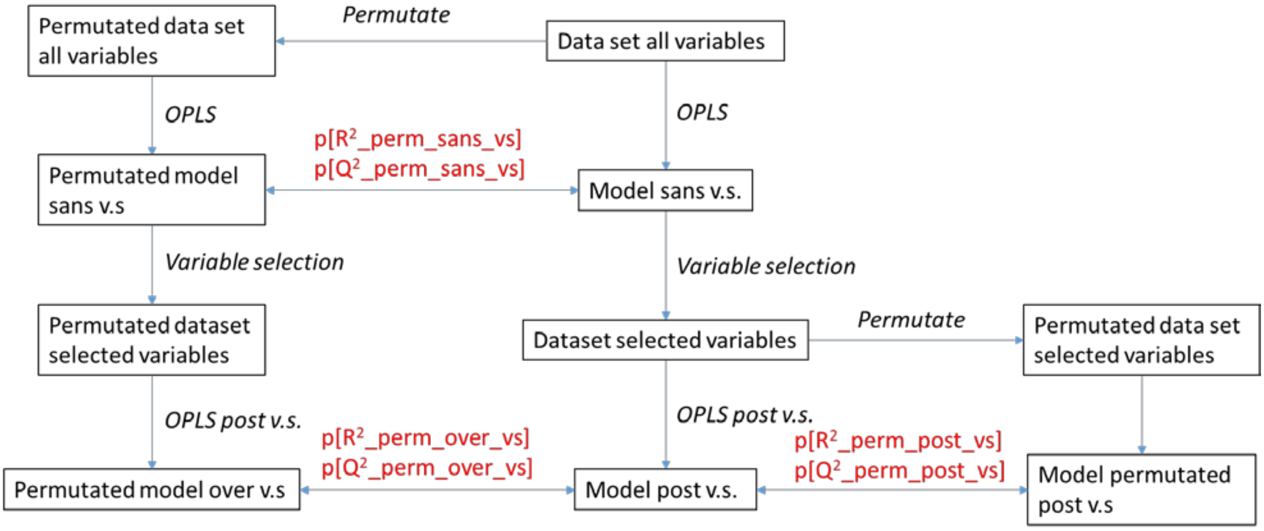


Figure 10. *Workflow for estimating significance level by permutations sans, post and over variable selection using roplspvs.*

4.3.6 Establishing an adjusted Q^2

As suggested by Lindgren et al. [45], the overfit may be established for a specific dataset by the difference between the average of Q^2 of the permuted models over variable selection and the average of Q^2 of the permutated models sans variable selection. This difference establishes the average increase in Q^2 during variable selection in the specific dataset. This difference in Q^2 may contain an actual enhancement of the model by removing unrelated variance, and may also contain a measure of the overfit of the model.

By removing the calculated overfit from the Q^2 post variable selection, we create an adjusted Q^2 which should represent a prediction without overfit. If this adjusted Q^2 is negative, the model should be rejected and considered insignificant. A limit for how low it can be and still represent a model remains to be evaluated.

4.3.7 SUS plots

To compare two models from roplspvs, correlation plots of p(corr) for each variable in each model are plotted in a so-called shared and unique structure (SUS) plot [171]. These are easy to create for models pre variable selection where all variables are present in both models. Models post variable selection are trickier to compare, as the variables selected in the models differ. Therefore, to compare models post variable selection, the variables from both models were used to fit a new model. For the visualization, p(corr) is selected firstly from the original model and secondly from the newly fitted model. The variables are colored for being shared or unique to each model. The names in the plot can be substituted for new names. Alternatively, the length and part of the name can be userset.

The SUS plot gives an overview of whether the same variables drive the models or whether the variables are unique for each model. The variables that cluster along the positive diagonal contribute similarly to the models, and variables that cluster along the negative diagonal contribute inversely. Variables along the axes contribute uniquely to one model.

4.3.8 Qualitative variables

To be able to analyze the clinical data in paper IV, roplspvs was developed to analyze variables with qualitative data. All variables containing characters are transformed into dummy variables using the package fastDummies. This is set to “remove selected columns”, which is the original column containing the character data the dummy variables are created from, but leaves all dummy variables by setting the variable “remove first dummy” to false. Keeping all dummy variable results in some redundant information, but makes it easier to view how all settings effect the model.

4.4 PAPER II

4.4.1 Isolation of EVs

For paper II, the sEVs were isolated using serial ultracentrifugation starting at 3000 g for 20 minutes to remove cell debris, followed by 10000 g for 30 minutes, 4°C using a rotor Ti45 from Beckman Coulter to remove large to medium vesicles, and finally pelleting the EVs that are enriched in sEVs using 140 000 g for 2 hours at 4°C with the same rotor. This fraction was used for the microarray analysis.

For next generation sequencing analysis of the miRNA, the sEVs were purified further using a sucrose gradient.

4.4.2 Charaterization of EVs using flow cytometry

In paper II isolated BALF-EVs were adsorbed onto 4 μm Aldehyde/Sulfate latex beads (Molecular Probes, Paisley, UK), coupled with mouse anti-human HLA-DR (T-1361 anti-human HLA II, Clone HKB1, BMA Biomedicals, Augst, Switzerland) as previously described (36). The EV-bead complexes were stained with FITC–conjugated antibodies or isotype controls (BioLegend, San Diego, CA, USA).

4.4.3 RNA extraction

For the microarray analysis in paper II, RNA extraction was performed using NucleoSpin® miRNA kit. RNA yield, size distribution of EVs, and cellular RNA integrity were accessed by gel electrophoresis using total RNA 6000 Pico LabChips and processed on Agilent 2100 Bioanalyzer.

4.4.4 Microarray

The miRNA in paper II were studied using microarray. The miRCURY 101 LNA microRNA power labeling kit was used to label the small RNA with Cy3-CTP. These were bybridized to one-color Agilent custom UCSF miRNA with 894 miRNAs for the BAL cells and the bronchoepithelial cells and 1212 miRNA for the EVs. The raw signals were extracted with Feature Extraction software after removing samples with scratches. The median feature pixel intensity was quantile normalized before univariate and multivariate analysis.

4.4.5 Pathway analysis

To investigate which pathways were affected by the miRNA, those miRNA that were drivers for the OPLS models, were included in pathway analysis. An overrepresentation analysis (ORA) was performed. The list of miRNA included in each model was submitted to the online tool miRWalk [172] to search for target genes. Only genes found in miRTarBase were selected, as they are experimentally validated. Pathway analysis was performed, inputting the gene list into the online application KOBAS 2.0 [173]. The pathways identified were filtered for KEGG pathways. Finally, the R package KEGGREST was used to remove the disease- and drug-related pathways from the list of pathways.

4.4.6 Next generation sequencing on RNAse treated samples

Treating EVs with RNase is a means of investigating whether miRNA is located inside the EVs or on the surface. To investigate if the miRNA analyzed using microarrays was located within the extracellular vesicle samples, one pool of samples from healthy individuals and one pool of samples from smokers were prepared. The EVs were enriched for sEVs by differential centrifugation and by sucrose gradient. Half of the samples were treated with RNase as this would result in degradation if the miRNA was located on the surface.

The miRNA extraction was performed using the NucleoSpin miRNA kit. An miRNA library was prepared using the Illumina TruSeq adapter and inhouse set of indexed primers for multiplexing. Pippin Prep with size selection was used to collect segments of 134-170 bp. An adaptor length of 125 nucleotides corresponds to small RNA of 9-45 nucleotides being captured. The Agilent Fragment Analyzer was used for quality control.

The UPPMAX cluster was employed to set up the workflow using Cutadapt-FastQC with Trim Galore to perform the trimming to around 22 nucleotides as that is the usual length of miRNA. The adapters were set to be autodetected. FastQC was used and showed that the sequence length was 15-23 bp. To choose the alignment tool, the results from Bowtie1 and Bowtie2 were compared. This showed that some additional miRNA were detected using Bowtie2. As Bowtie1 is recommended for fewer than 50 base pairs, and it was suspected that gaps were allowed using Bowtie2, Bowtie1 was used for alignment to mature and hairpin miRNA base pairs. The sequenced miRNAs were aligned to hairpin and mature miRNAs, resulting in a coverage of 45-514 thousand counts and an alignment rate of 0.4-3.6%.

The low alignment rate could be because the concentration of miRNA is very low, with only one miRNA copy out of 100 EVs [174]. The reason that miRNA were still present after RNase treatment could also be that proteins present on the surface of EVs might protect the RNA against degradation by binding to the miRNA. Therefore, actively bound proteins might still be present on the surface even though miRNA from dead cells that passively stick to the surface is degraded [175]. Because no protease was added to the samples, there is a risk that the miRNA was not degraded by the RNase even if it was present on the surface. The reason for the low alignment could also be that RNase is difficult to remove from the sample and hence risks degrading the internal miRNA. Another aspect to consider is that RNase is present in some bodily fluids [176], and this content could differ with diagnosis [177].

The data analysis was performed using EdgeR in R. The trimmed mean of M-values method (TMM) [178] was used for normalization as it avoids false positive rates better than other normalization methods do [179].

4.5 PAPER V

4.5.1 Mass spectrometry

For paper V, the metabolites were analyzed with liquid chromatography–high resolution mass spectrometry (LC-HRMS) using a previously described method [180]. This method applied hydrophilic interaction liquid chromatography (HILIC) on a SeQuant ZIC-HILIC (Merck, Darmstadt, Germany) column, enabling the detection of hydrophilic metabolites. In addition, reversed phase (RP) was applied on a 1290 Infinity II ultrahigh performance liquid chromatography (UHPLC) system using the column Zorbax Eclipse Plus C18, RRHD to separate nonpolar metabolites. The LC systems were coupled to an Agilent 6550 and a 6490 iFunnel quadrupole time-of-flight (Q-TOF) mass spectrometer equipped with electrospray ionization source (Agilent Technologies), using both positive and negative modes for hydrophilic metabolites and positive mode for the hydrophobic metabolites. A pooled sample for each day was prepared to use as an internal standard.

4.6 PAPER VI

4.6.1 Mass spectrometry

For paper VI, the peptides derived from trypsin digestion of the proteins were analyzed with liquid chromatography using an analytical column from Thermo Fisher Scientific NanoViper Acclaim Pepmap C18 particles 3 µm (350x0.075 mm I.D.). The samples were 10-plex TMT with ten samples in each set. The advantage of using multiplexed isobaric labeling is that it minimizes variation and saves time (Figure 11).

One of the samples was a reference sample, containing a pool with equal amounts of all 16 study subjects. This was used to account for batch effects by expressing the levels as ratios against the internal standard. Before the MS analysis, the samples were trypsin digested into tryptic peptides. Trypsin cleaves the peptides at arginine and lysine at the C-termini of the peptide, resulting in peptides from intact proteins. Semi-tryptic peptides are also formed which are only digested by trypsin at one end, and may be formed from degradation products formed during storage of the samples. To be able to analyze ten samples at a time, the peptides were labeled using isobaric labeling called Tandem mass tags (TMT) (Figure 11). This contains a reporter tag and a balancing tag with a total constant weight so that the labeled peptides will still weigh the same and emerge as one peak after mixing the ten samples.

The eight fractions of samples were analyzed using an Orbitrap Fusion Tribrid mass spectrometer. Electrospray ionization source (Agilent Technologies) in positive mode was used. For mass spectrometry 2, the peptides were fragmented by collision induced dissociation (CID).

The experimental setup for the MS analysis was that each individual was included in the same set, except for half of the 36-hour samples which were run in a separate set. To increase the number of peptides detected, a second injection was performed.

Quantification and identification were performed using Proteome Discoverer, and a database search was conducted with the Mascot search engine, using the Homo Sapiens Swissprot database. The data was exported and analyzed at protein level, at peptide level, and using both tryptic and semi-trypic peptides.

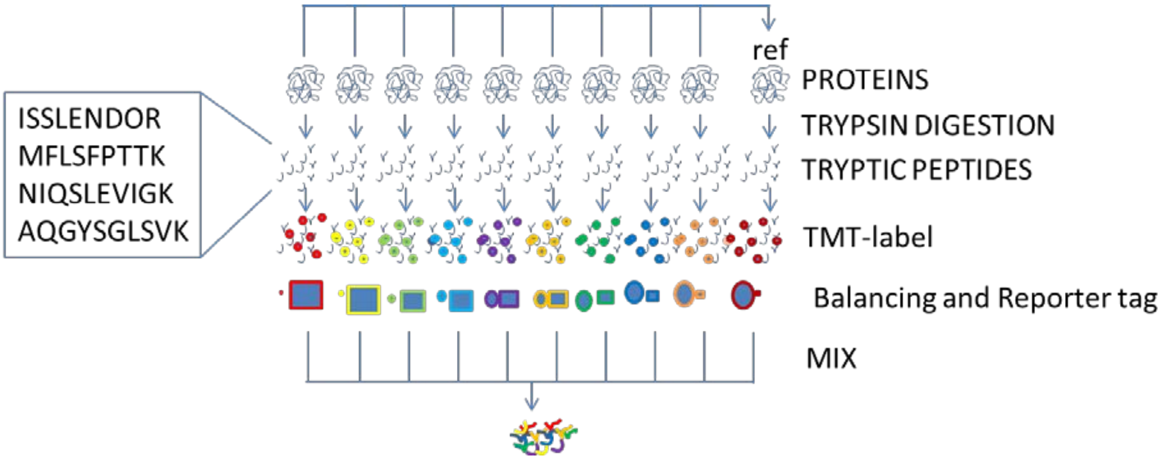


Figure 11. *Workflow of mass spectrometry using tandem mass tags (TMT) multiplexing. The proteins are trypsin digested at arginine and lysine producing tryptic peptides which are TMT-labeled including balancing and reporter tag before analyzing the mixture of samples. The procedure decreases the variation between samples connected to different runs.*

5 RESULTS AND DISCUSSION

5.1 PAPER I

5.1.1 Results

5.1.1.1 The roplspvs package and workflow

Oplspvs is an R package that creates OPLS-DA models using the Bioconductor package ropls, and it also contains a workflow for running the package, as well as performing variable selection, model optimization, significance testing, creating figures, creating tables of biomarkers, and creating summary tables of models. Variable selection is performed using p(corr) cutoff, with the option of also utilizing VIP. P(corr) cutoff is defined by the user to optimize the predictivity of the model, as determined by cross-validation.

The significance of the models is tested using permutations including the variable selection procedure, resulting in p-values for R^2 and Q^2 (p[R^2 and Q^2 perm over v.s.]) that include the variable selection procedure as well as permutation pre (a.k.a. sans) and post variable selection. The package workflow for running the package is user-friendly and flexible, with parameter settings in two files: one for basic parameter settings and one for advanced settings where default values may be changed.

The OPLS workflow in roplspvs provides five different modelling strategies to perform pairwise group comparisons, including optional stratification by user-provided metadata. The first modeling strategy produces models with user-defined p(corr) cutoff and maximum number of orthogonals. This enables the user to test a model created in another software, or to explore another model than the one selected automatically in the script. The second modelling strategy includes all variables significantly correlated to the model. This selection of variables is optimized for pathway mapping. The third modelling strategy includes the variables that create the most predictive model for identifying biomarkers. The fourth modelling strategy provides the most stringent selection of predictive variables, and can be of use e.g. for selection of therapeutic or diagnostic candidates. For comparison with iterative modeling, a fifth modelling strategy using an iteratively increasing p(corr) cutoff is available. Finally, a modelling strategy 0 is also available, which provide a model without variable selection. After the models have been created, p(corr) of the variables in the models can be compared using SUS plots. Comparing models post variable selection usually involves comparing lists of variables that do not overlap completely. Therefore, to be able to compare the variables in the SUS plot, a new model including the variables from both models are created. For the SUS plot the p(corr) from the original model is used firstly and secondly by the p(corr) from the new model. Shared and unique variables of original models are indicated using different colors.

5.1.1.2 *Example of an roplspvs application*

The performance of the script was demonstrated using previously published data from the public repository Metabolights (https://www.ebi.ac.uk/metabolights/MTBLS136/files).

MTBLS136 is a metabolomics dataset comparing progestin (P) and estrogen (E) users with nonusers. We show significant models of metabolites driving separation between progestin and estrogen users compared to nonusers, stratifying the analysis by different age groups.

The models comparing E+P and E-only versus nonusers were all significant using permutations pre, post, and over variable selection.

Comparing E+P versus E-only gave significant models for all age groups except 56–60 and 76–80 using permutations pre variable selection. These models also had negative Q^2 pre variable selection, indicating models without any predictivity. Proceeding to variable selection, the models that were created post variable selection were significant for all age

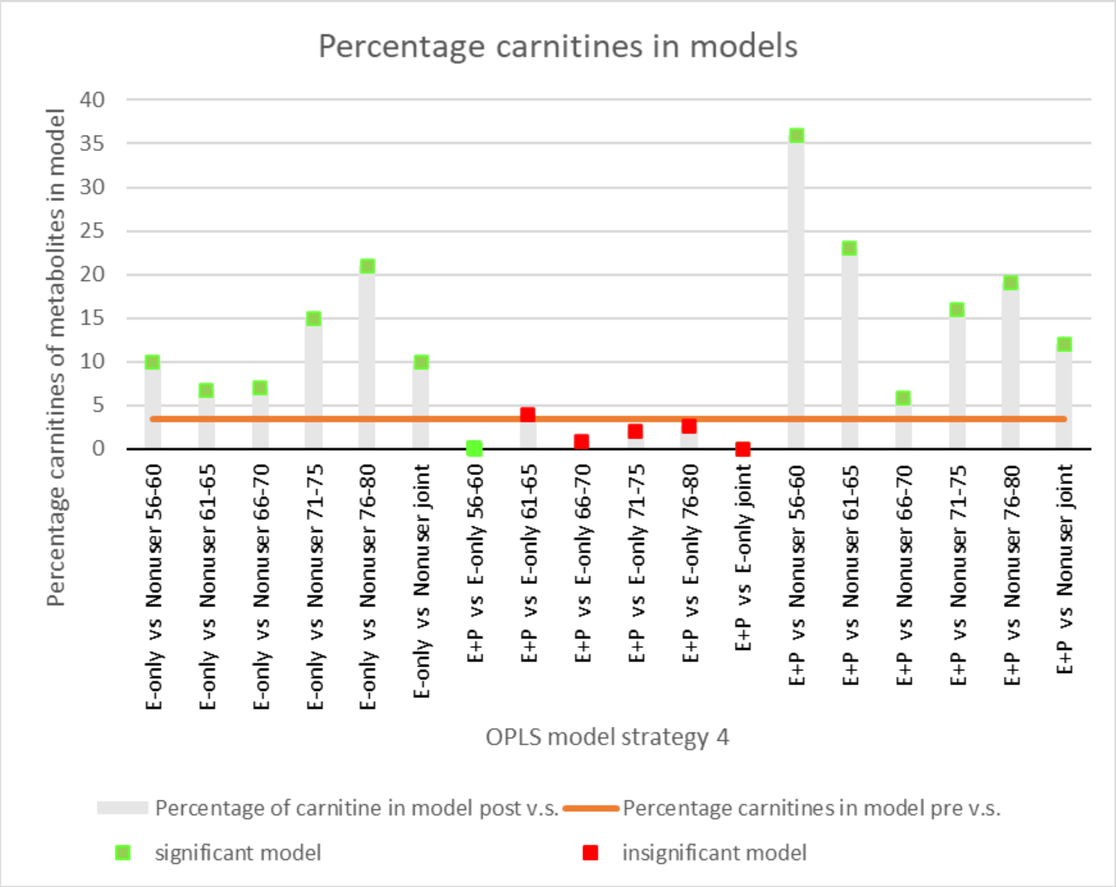


Figure 12*. Percentage carnitines of metabolites driving models between hormone users and nonusers as well as between hormone users of combined hormons and single hormone users. Carnitines were more frequent drivers in models comparing hormone users with nonusers than would be expected by random contribution. Green squares: Significant model p(Q2 permutated over variable selection)≤0.05 and positive adjusted Q2; Red squares: Insignificant models p(Q2 permutated over variable selection)>0.05 or negative adjusted Q2.*

groups except age group 76–80 using permutations over variable selection and regression over permutations. However, the adjusted Q^2 indicated that the models were not predictive, as all models had negative adjusted Q^2 except for the model 56–60 years old. On investigating which variables contributed to the models, it was found that carnitines were more frequent among selected variables in all age groups models compared with the frequency of carnitines among all metabolites (Figure 12). This is in line with what Stevens et al. showed when analyzing the same dataset, i.e. that acyl carnitines were lowered in post menopausal users.

5.1.1.3 Effect of small sample sizes on OPLS model statistics

Using the same dataset, the effect of sample size on the model statistics was investigated. A strong model (R2=0.57 and Q2=0.53) comparing E+P versus nonusers age group 61–65 was chosen representing a strong model, and a model comparing E+P versus E-only users was chosen as a weak model (R2=0.14 and Q2=0.10). Comparing nonusers age group 61–65 to each other represented a random model. A subset of samples were drawn from each group, creating even sample sizes ranging from four to 80 subjects. Using roplspvs, OPLS models were created for all sample sizes. The procedure was repeated 12 times, and the Q^2 for each model using modeling strategy 3 was plotted against sample sizes.

Using decreasing sample sizes resulted in increasing R^2 and Q^2 of models post variable selection to the extent that it approached 1 at the smallest sample sizes for all three comparisons; random model, weak model and strong model. When using permutations post variable selection, all models were significant including comparing nonusers with nonusers, i.e., in the artificial models produced as a negative reference (Figures 14A-C).

In these models pre variable selection R^2 also approached 1.0 (Figure 13), whereas Q^2 only increased slightly in the model comparing nonusers with nonusers, was not affected in the weak model, and decreased in the strong model (Figures 14J-L). Using permutations sans variable selection discriminated well between the strong model and the model nonuser versus nonusers. It was only at the very small sample sizes that the strong models were insignificant.

Using permutations over variable selection also discriminated well between the strong model and the model produced from nonusers versus nonusers. However, a slightly higher fraction of models than expected when considering significance level were deemed significant (Figures 14D-F). When also considering the regression over permutations, the number of models deemed significant on random data correlated well with the significance level (Figures 14G-I).

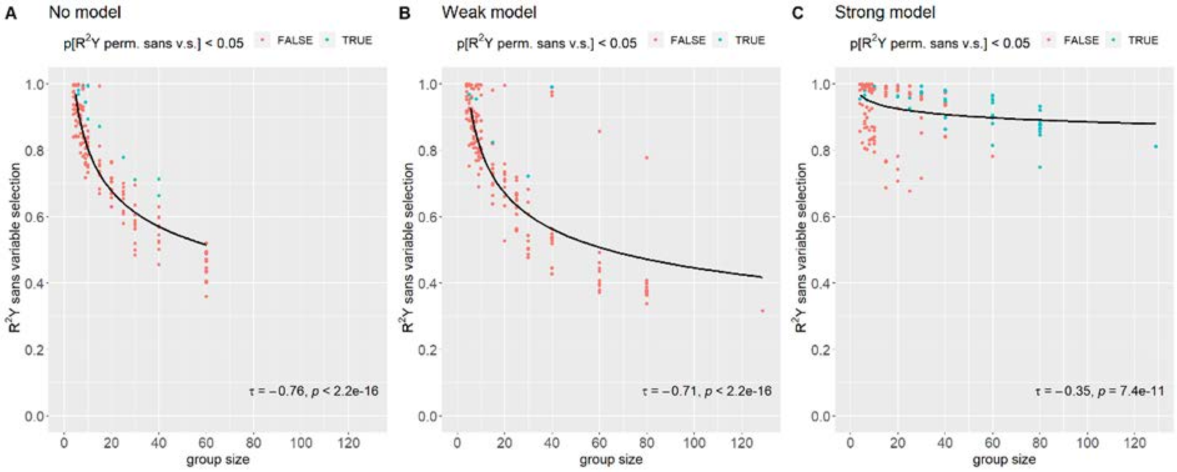


Figure 13. *R2 of OPLS models pre variable selection of 1012 metabolites approached 1 with decreasing sample sizes.*

5.1.1.4 Establishing a threshold for Q^2

Q^2 of models post variable selection compared to Q^2 of models pre variable selection are compared in figure 15. The difference between Q^2 post and pre variable selection in green

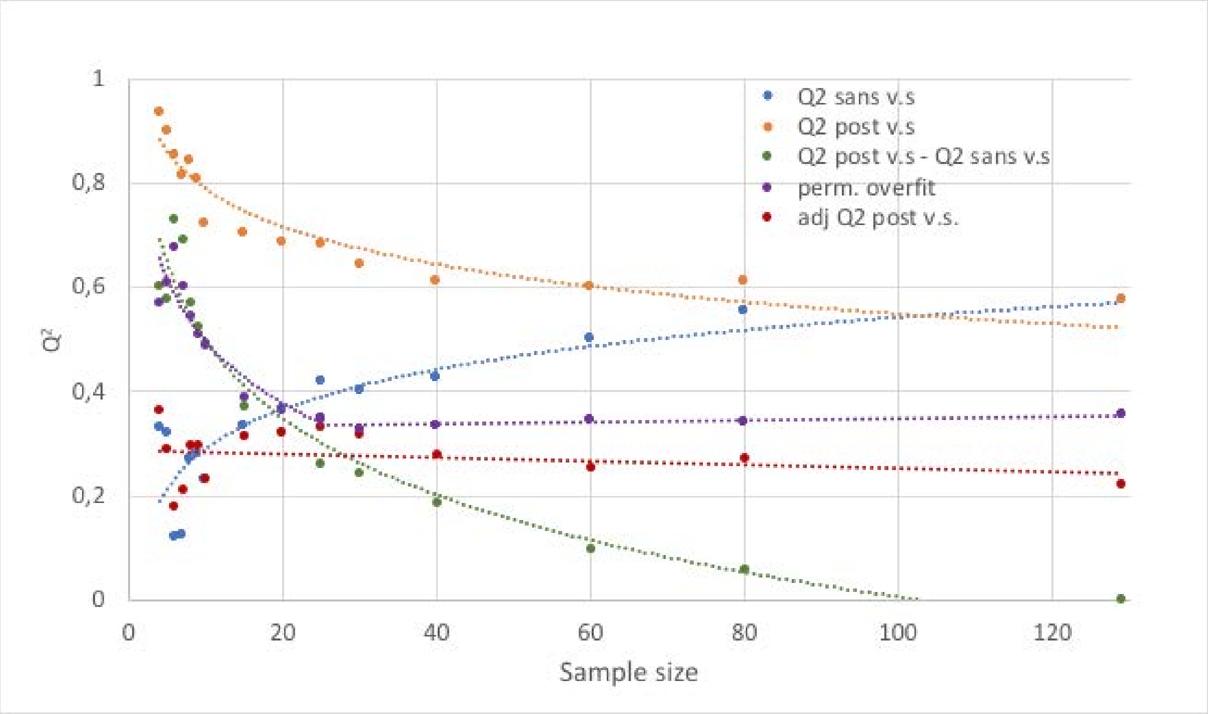


Figure 15. *Q^2 versus group size of models comparing the groups of E+P versus Nonuser, age group 61-65, showing Q^2 of models sans variable selection in blue, Q^2 of models post variable selection in yellow and the difference between Q^2 sans and post variable selection in green. Permutated overfit in purple is calculated subtracting the average of the permutated models sans variable selection from the Q^2 of the permutated models over variable selection. Adjusted Q^2 post variable selection in red is obtained by subtracting the overfit from Q^2 post variable selection.*

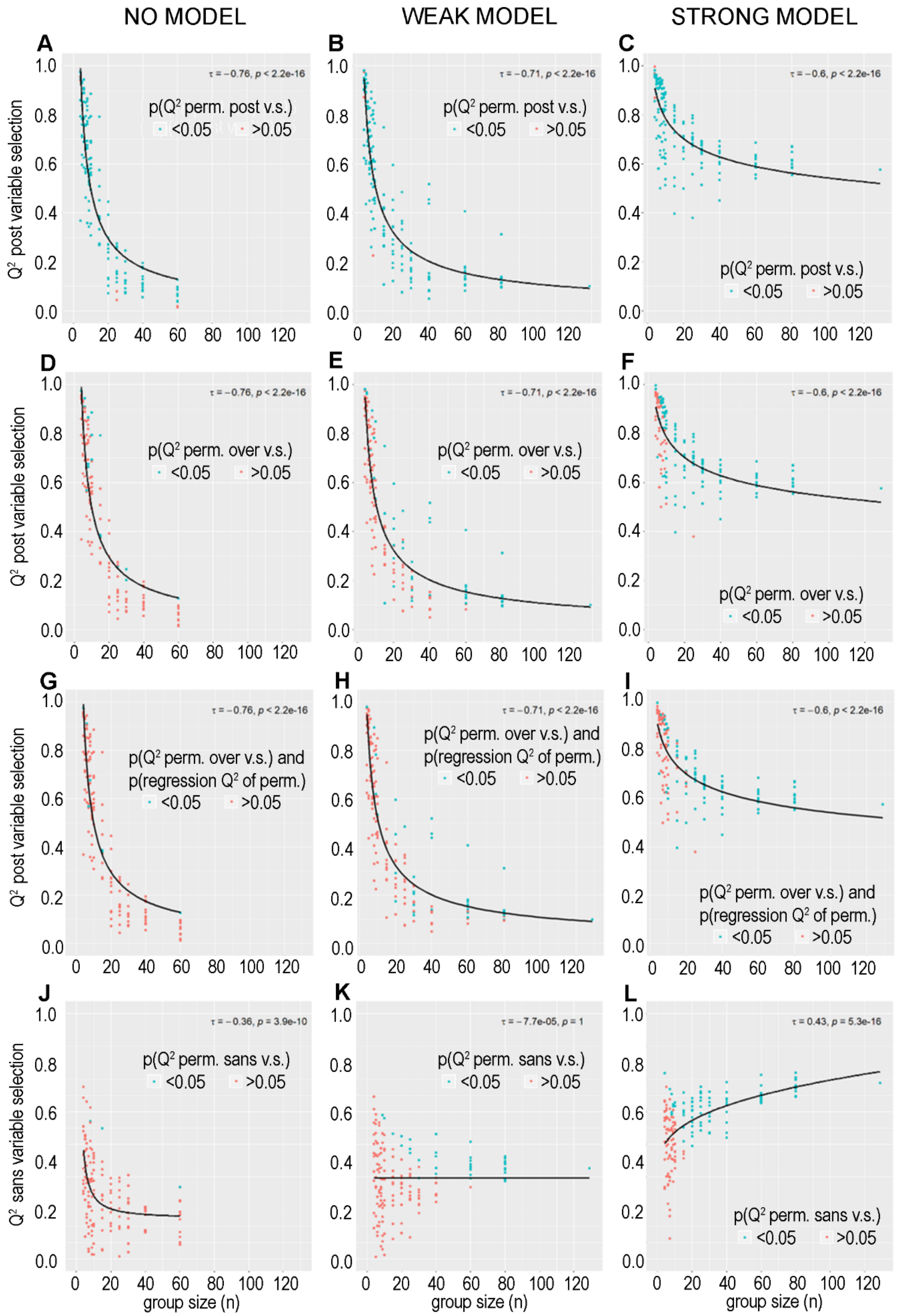


Figure 14. *Q^2 of models using Modelling strategy 3 over subsets of groups with altering size from a strong, a weak and a model on randomly selected subjects. The power to discriminate*

between strong, weak model and model on random data failed using permutations post variable selection demonstrated by insignificant models across all samples sizes (A-C), colored by p[Q^2 permutated post variable selection]<0.05). Permutations over variable selection discriminated well between resulted in the majority of the models in the weak- and non-models were insignificant (D-E), whereas the majority of the strong models were significant as long as groups were larger than 8 (F colored by p(Q^2 permutated over variable selection). Including the significance of the correlation of the permutation correlation resulted in the fraction of random models being significant corresponding to significance level and decreasing the number of significant strong models slightly (G-I). Permutation sans variable selection displayed decreasing Q^2 with decreasing group size (L) and resulted in the weak and non models being insignificant while strong models were mainly significant at group size larger than 8 (J-K, colored by p(Q^2 permutated sans variable selection).

show that the Q^2 is increased more using small sample sizes. This represents the overestimate of the model but may also contain a true increase in predictability thanks to variable selection.

That Q^2 is not enhanced in the largest sample sizes indicate that this increase is only an overestimate. One way to estimate the overestimate is to use permuted data, subtracting Q^2 of permutated model pre variable selection from Q^2 of permutated model post variable selection as proposed by Lindgren et al. This overestimation in permutated models shown in purple in figure 15 and represent an overfit in the actual dataset under hand. To establish a Q^2 that was not overestimated we subtracted the permutated overfit from Q^2 post variable selection yielding a Q^2 adjusted for overestimation shown in red in figure 15.

Investigating that the overestimate of R^2 and Q^2 was not due to larger number of orthogonal components the number of orthogonals were plotted over the sample sizes (Figure 10 Paper 1). This show that the number of orthogonals were not decreased with decreasing sample sizes.

5.1.2 Discussion

Orthogonal projections to latent structures discriminant analysis (OPLS-DA) is a supervised multivariate method adapted from partial least square (PLS). The workflow and package roplspvs were developed in the R programming platform to create an automatic and user-friendly tool for performing the analysis and selecting variables optimized for pathway analysis and biomarker discovery workflows.

The choice of using R to produce the package gives access to a large number of packages covering statistical and bioinformatics tools. It is highly flexible but may involve an initial obstacle for the user who has no previous programming experience. The workflow is particularly useful when the analysis is planned to be repeated on new data, but is more time-consuming if a single analysis is to be performed

Due to the risk of overfitting OPLS models when using small sample sizes, roplspvs includes thorough significance testing using permutations pre, post, and over variable selection, ensuring that the models are performing significantly better than models created at random. Finally, the advantage of performing permutations over variable selection was demonstrated using the same dataset.

Reducing sample sizes resulted in increasing R^2 and Q^2, which should not be the case, as smaller sample sizes statistically provide a lower statistical power.

It was also shown that creating random groups from a homogenous group of subjects consisting of women aged 61 to 65 and modeling these groups using variable selection resulted in significant models using permutation post variable selection. We also showed that the models were deemed significant using a sample size ranging from four to 80 subjects in each group. This means that the procedure of variable selection has identified the biological variation that exists in this group. The permutations post variable selection identify these as significant without taking into account that these alterations were selected from a large set of variables.

Using permutations over variable selection takes into account that the variables in the model were selected from a large number of variables. It tests whether the models are performing better post variable selection than models produced from random data. Here, 13% of all models comparing nonusers with nonusers were significant (Figures 14D) over all the different sample sizes used. Using a significance level of 5%, we should find 5% significant models. That 13% is found to be significant indicates that the model tested has been optimized more than the models created using random data. This may be because p(corr) has been optimized for the tested model, while the permutated models are not optimized and instead use the same p(corr) as the unpermutated model. To circumvent this effect, it is possible to study how the model statistics are affected by how similar the permutated group labels are to the correct unpermutated group labels. If the permutated group labels happens to be assigned with the "correct" unpermutated group labels the model statistics should be the same, and the more similar the permutated and unpermutated labels are, the more the model statistics should approach the statistics of the original model. Therefore, applying a regression over the permutations should result in a possitve regression coefficient for a significant model. When comparing nonusers versus nonusers, the regression coefficient (>0.1) and the p-value for the regression coefficient over the permutations (<0.05) were also considered, with 6% of the models being significant (Figures 14G).

Studying the strong model comparing E+P users versus nonusers of aged 61 to 65, the majority of the models using small sample sizes were significant. However, reducing the sample sizes also resulted in many insignificant models. The reason for this may be that the subjects sampled in the small sample size did not represent the alteration present in the whole group, as the smaller the sample size is, the greater the risk for biased sampling within the group. Small sample sizes also increase the risk of the permutated group labels by accident happening to be the same as the unpermutated group labels. If more that 5% of the permutated group labels are accidently assigned a correct label, the model would be deemed insignificant, resulting in a false negative. An improvement to the script may therefore be to remove the correctly labeled permutations. For now, the permutation plot or table in the html report or Rdata files can be investigated manually to check for correct labels.

In conclusion roplspvs is a user friendly and automated workflow and package to run OPLS models including permutation over the variable selection and provides an estimate of the overfit due to many variables and few samples as well as an adjusted Q^2 providing a threshold for the predictability.

5.2 PAPER II

5.2.1 Results

The goal of this project was to study the miRNA cargo of extracellular vesicles (EVs) enriched for exosomes, isolated from BAL fluid from COPD patients versus control groups using the Karolinska COSMIC cohort. To avoid the strong confounding effect of current smoking, the aim was to compare the COPD patients with a group of smokers who had not yet developed COPD. The smoking effect was also studied by comparing the smokers to healthy never-smokers.

5.2.1.1 Univariate and multivariate analysis

A two-way ANOVA linear model was applied using the Limma Bioconductor package in R. The p-values were adjusted using the Benjamini–Hochberg method of false discovery rate (FDR)[181]. Given that the COSMIC cohort is designed to investigate gender differences in smoking-induced COPD, analyses were performed both on joint gender groups, and stratified by gender. The joint gender p-values were adjusted over the variables and over the joint gender comparisons, and the gender-separated p-values were adjusted over the gender-separated comparisons. This resulted in large alterations when comparing COPD patients with healthy individuals, with 44 significantly altered miRNAs (FDR < 0.05). Comparing smokers with healthy individuals resulted in 40 miRNAs being significantly altered (FDR < 0.05). Finally, comparing COPD patients with smokers resulted in no significant alterations.

Stratification by gender in the comparisons of COPD patients versus healthy individuals resulted in 14 and 21 significantly altered miRNAs (FDR < 0.05) in females and males, respectively. The same gender stratification when comparing smokers with healthy individuals resulted in 26 and 14 significantly altered miRNAs (FDR < 0.05), respectively.

Using the multivariate OPLS method via the SIMCA tool resulted in highly significant models comparing joint gender COPD versus healthy individuals including 41 miRNAs in the model ($R^2 = 0.75$, $Q^2 = 0.71$, p[CV-ANOVA] = $2*10^{-9}$). The significance of the model was also confirmed using roplspvs, which resulted in p(R^2 and Q^2 perm. over v.s.) ≤ 0.001. The gender-stratified comparisons did not generate any improvement in significance of the model, and the miRNAs that were altered in the joint model, to a large extent, also drove the gender-separated models. It was therefore concluded that gender was not a major driver in these comparisons.

Likewise, when comparing smokers with healthy individuals, the models were not improved following stratification by gender. The majority of the miRNAs driving both joint and stratified models were the same, with no strong gender dependence. The joint gender model was highly significant, with 36 miRNAs contributing to the model ($R^2 = 0.69$, $Q^2 = 0.65$, p[CV-ANOVA] = $1*10^{-8}$, p[Q^2 perm. over v.s.] ≤ 0.001).

In contrast, multivariate models to compare the COPD and smoking groups did not generate any significant model at the joint gender level. There was no significant model for females, but for males, a significant group separation was observed ($R^2 = 0.41$, $Q^2 = 0.31$, p[CV-ANOVA] = $4.9*10^{-2}$) based on 25 miRNAs.

The model sans v.s. had a Q^2 of −0.16 and a p(Q^2 sans v.s.) of 0.579, which means that the model before variable selection had no predictivity and was insignificant. One reason for developing the roplspvs script described in paper I was to evaluate whether OPLS on the verge of significance, such as the OPLS models comparing males with COPD and male smokers here, can be considered true positives based on permutating analysis. The permutations resulted in p(Q^2 over v.s.)=0.11, whereas the model sans variable selection had a Q^2= −0.16 with p(Q^2 sans v.s.)=0.58. Calculating the overfit of the model as proposed in Paper I, by the difference between average permutated Q^2 post v.s. and average permutated Q^2 sans v.s., resulting in an overfit due to small sample sizes of 0.26. Subtracting the overfit from the Q^2 post v.s. of 0.31 resulted in the model just reaching over the threshold (adjusted Q^2=0.05).

5.2.1.2 Validation using qRT-PCR

The miRNA were validated using qRT-PCR detecting 11 miRNA above LLOQ.

5.2.1.3 Pathway analysis

Pathway analysis by overrepresentation analysis (ORA) using the miRNA driving the OPLS models revealed significantly enriched pathways (p<0.001) (Table E5 paper II). Comparing COPD versus Healthy revealed that 18 pathways were enriched, 17 were enriched when comparing smokers versus healthy, and comparing male COPD patients versus smokers revealed that 21 were enriched. The p53 signaling pathway, as well as other cell growth and death pathways, were more related to smoking, and pathways related to autophagy, mitophagy, and tight junctions possibly being related to COPD pathology (Figure 16)

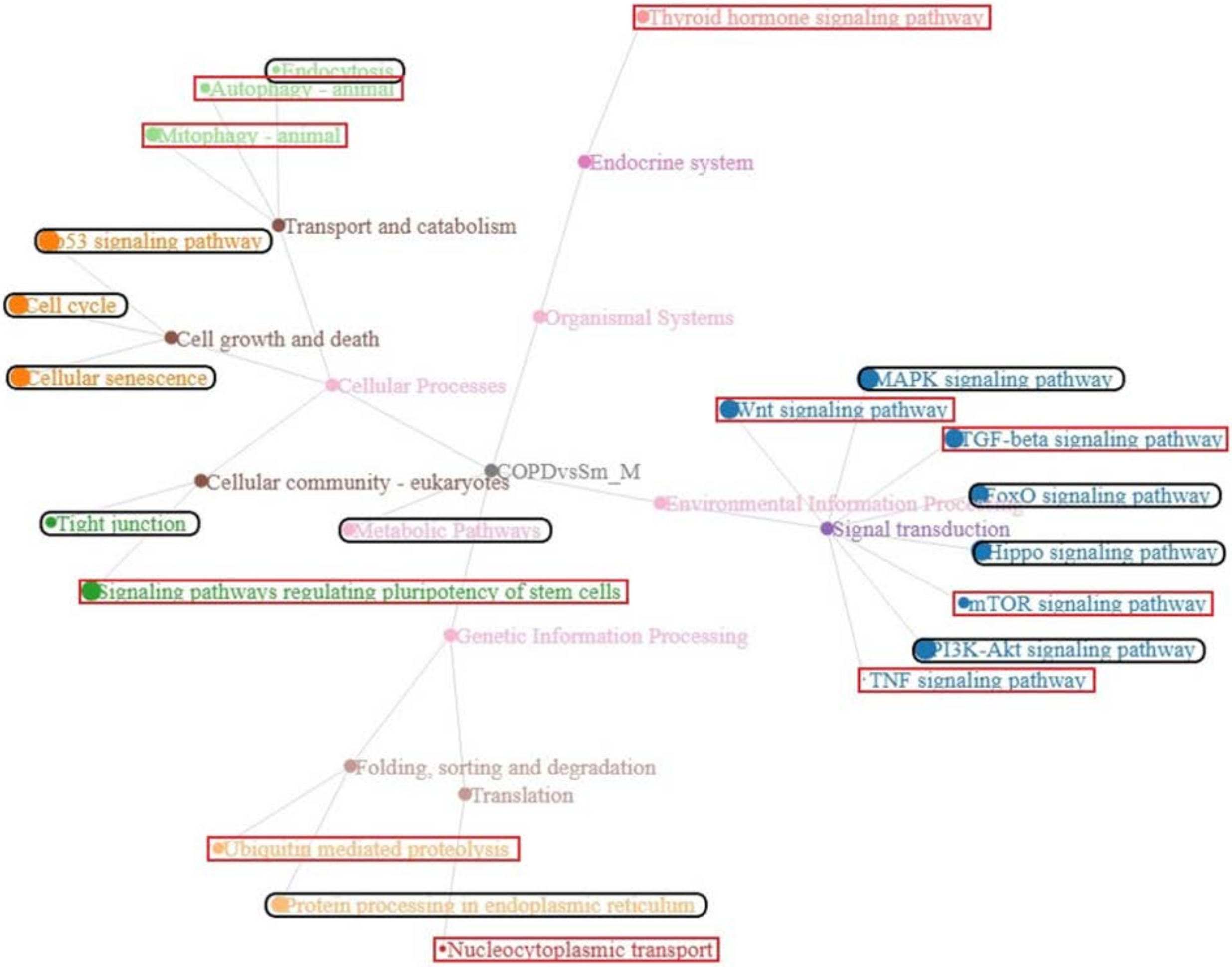


Figure 16. *KEGG pathways targeted by the miRNAs driving the OPLS models comparing male COPD patients versus male healthy controls. Related pathways are shown in the same color. Pathway uniquely altered in COPD patients compared to Smokers without COPD are circled by red rectangles. Pathways altered in both COPD vs Healthy, Smokers vs Healthy and COPD vs Smokers male are circled by black ovals.*

5.2.1.4 Flow cytometry of BALF-EVs

Flow cytometry showed that surface markers CD9, CD86 and HLA-DR were more frequent in EVs from smokers compared to both COPD patients and to healthy controls.

5.2.1.5 Levels of miRNA in EVs compared to epithelial cells and BAL cells

All 134 miRNA detected in the BALF-EVs were also detected in BAL cells and all but one of them were also detected in BEC cells. Enrichment of miRNA in EVs compared to BAL cells and BEC were investigated to decipher which cells the EVs were derived from. This showed that twelve miRNAs were enriched in EVs in BAL compared to BAL cells and 6 miRNAs were enriched compared to BEC (Figure E8A and B in paper II) without being related to abundance. The enrichment factor between BALF-EVs and BAL cells were significantly lower in COPD compared to Smokers for 7 miRNAs.

5.2.1.6 YBXI protein

Packaging of miRNA into sEVs is an active process and the RNA-binding protein Y-box protein 1 (YBX1) is required for the packaging of specific miRNAs [148, 149, 182]. It was shown that YBX1 levels were decreased in the smokers compared to healthy individuals and further decreased in COPD patients (Figure E9A paper II) and that this alteration was mainly driven by males (Figure E9B paper II)

5.2.1.7 RNase treated samples

With the aim to investigate whether the detected miRNAs were located inside the EVs or on the surface, two pools of samples were prepared: one from healthy individuals and one from smokers. Half of the pooled samples were treated with RNase to degrade surface miRNA. After enrichment of sEVs using differential centrifugation and saccharose gradient, the miRNAs were sequenced using next-generation sequencing.

The sequenced miRNAs were aligned to hairpin and mature miRNAs, resulting in a coverage of 45-514 thousand counts and an alignment rate of 0.4-3.6%. Of the remaining reads, 5-20% aligned to human ribosomal RNA, 15% aligned to genomes, and 20-65% stayed unaligned.

Investigating the differences between the treated and untreated samples revealed that the miRNA contents in both samples from healthy individuals and samples from smokers were mostly unaffected by the RNase treatment (Figure 9A and C in paper II) and that the miRNA that remained detected both before and after RNase treatment correlated well (Figure 9B and D in paper II). Four miRNAs were degraded during RNase treatment in each pool, but there were also two and three, respectively, detected in the RNase-treated pools that were not detected before RNase treatment.

MiRNA contents in samples from both healthy individuals and smokers correlated well between the RNase-treated and untreated samples.

5.2.2 Discussion

Several studies have shown that EVs are important factors in respiratory lung diseases. Alterations have been shown in cystic cystic fibrosis [130], sarcoidosis [127] and asthma [113, 183]. It has been shown that smoke induces EV release from lung epithelial cells which in its turn secreted full-length CCN1 and facilitated interleukin (IL)-8 and vascular endothelial growth factor (VEGF) release, as well as increased protease and matrix metalloproteinase (MMP)-1 production [184].

A study has earlier shown small effects of COPD on EVs in BAL compared to healthy subjects and no effect of smoking [185]. By comparison large effects of COPD and smoking on the miRNA cargo of sEVs as well as on surface markers are shown in this study. 44 miRNAs significantly altered (fdr<0.05) when comparing COPD patients with healthy individuals' huge alteration which involves 39% of the 105 miRNA detected above the LLQ. Most of these alterations were shown to be due to smoking, with 40 miRNAs (35% of detected miRNAs) were altered due to smoking alone (FDR<0.05). Smoking has also earlier been found to have strong effect on the proteome in the same cohort, with 50% of the proteins being altered in BAL cells [38].

Using the overrepresentation analysis the p53 pathway was found to be the most significantly affected pathway in COPD patients, both compared to smokers with normal lung function, and to healthy never-smokers. Also, other pathways related to p53 pathways were indicated to be affected such as cell cycle-, cellular senescence and FoxO signaling pathways. P53 mRNA has earlier been found to be increased primary human bronchial epithelial cells (HBECs) from COPD patients compared to cells from smokers [186]. Also, cells that lack p53 has been shown to not secrete sEVs and resume secretion if the p53 target gene transcript designated tumor suppressor activated pathway-6 (TSAP6) is upregulated [122].

TNF signaling pathway, which was also targeted by miRNA from male COPD versus smokers, regulates NF kappa B which is a known player in COPD [187]. NF kappa B has also been suggested to be involved in both regulating and being regulated by p53 [188].

The differential comparison of male current-smoker COPD patients with smokers with normal lung function represent the effect of COPD pathology, without confounders from smoking. In summary the model of COPD versus non-COPD smokers had an adjusted Q^2 that showed a predictive power just above the threshold, a significance using CV-ANOVA just below the 5% significance level, and a significance at a 11th percentile when using the relatively stringent permutation over variable selection. It is challenging to study COPD because of the large confounding effect of smoking that is obvious from the large alteration between smokers and healthy individuals. Also, considering the importance of EVs has on the immune system. A study indicates that EVs are involved in the damaging effect of cigarette smoke on epithelial cells [184]. We therefore considered a 11% significance level taken together with model falling above the threshold for random model to be sufficient evidence for a true positive, and thus performed downstream pathway analysis on this relatively weak male model of COPD versus smokers. However, the results should be considered with caution.

The 25 miRNA driving the model correlated significantly with FEV_1% predicted (r=0.65, p=0.04) (paper II Figure 4A).

The male alteration of autophagy and mitophagy is in line with gender specific alterations that were earlier observed in the same cohort with proteome alterations in the phagocytotic pathways in BAL cells from females [189] and lysosomal pathway in the same females using gel electrophoresis instead of mass spectrometry [89]. Both could be connected to defect mitochondria with dysregulation of macroautophagy. Macroautophagy is one of 3 autophagic pathways and has been connected to COPD [190]. The alterations in mTOR signaling pathway which regulates autophagy further support this connection. Both sEVs and lysosomal degradation originates from the late endosomes further strengthens the link.

MicroRNA-423-5p, which was increased in COPD patients compared to healthy in our study, has been identified to promote Autophagy in Cancer Cells [191]. This microRNA was also one out of the three identified as altered, however in opposite direction, in earlier study on miRNA in Evs in BAL from COPD patients [185]

YBX1 protein has been shown to be required for packaging of sEVs [148, 149]. Here stepwise decreasing levels of YBX1 protein in smokers and further decreased in COPD both in joint gender and in males indicating a dysfunction of packaging of sEVs in COPD.

The presence of Alix and the tetraspanins CD9, CD63 and CD81, and absence of Calnexin, indicates that sEVs were enriched. CD9 was increased in smokers both compared to COPD patients and compared to healthy. This is interesting as mice deficient in both CD9 and CD81 have been shown to develop emphysema-like condition [192]. It has also been suggested that CD9 has a protective role against inflammation induced by smoking [193]. The decreased amount of CD9 in COPD patients compared to smokers may thus indicate loss of protection against the effect of smoke in COPD. Deficiency of tetraspanins CD9 and CD81 has been associated with increased MMP-2 and MMP-9 production [194], both associated with alveolar destruction in emphysema [195].

In conclusion, large alterations were observed in miRNA content in sEVs due to COPD and smoking compared with healthy never-smokers that were related to p53 pathways. Also, possible minor alterations in COPD compared to smokers in males which was linked to decrease in mitophagy and autophagy pathways. It was also proposed that the protective role against smoke, that CD 9 has been suggested to have, may be EV-mediated. Altogether this indicates that sEVs may have a protective role against smoking and that this is impaired in COPD effecting packaging of miRNA and secretion of sEVs.

5.3 PAPER III

5.3.1 Results

In paper III, the LUNAPRE cohort was used to study the effect of bronchopulmonary dysplasia (BPD) in preterm born young adults. The LUNAPRE cohort is described in detail in table E1 in paper III and consists of a group of preterm born with BPD (gestational week 24-31) and one without BPD as well as healthy controls and asthmatics. Previous study on this cohort has shown that the BPD group had lower FEV1 than healthy controls, asthmatics and preterm born [162].

The composition of surface markers of immune cells collected by bronchoalveolar lavage was studied in young prematurely born adults with a history of BPD, compared to control groups of prematurely born without BPD, as well as healthy and asthmatics born at term. Increased numbers of CD8+ T-cells were identified together with decreased numbers of CD4+ T-cells, indicating immunological alterations that correspond to the alterations found in previous studies investigating COPD caused by smoking [196].

The alterations were studied using the nonparametric Wilcoxon test and FDR analysis, adjusting for multiple comparisons of groups and many variables. Effect sizes were also calculated (Figure E5 in paper III).

Comparing the BPD group with healthy controls, the CD8$^+$ T-cells were higher (p = 0.005; P<0.03; Figure 17B). and the CD4$^+$ T-cells was lower (p=0.014; Figure 17A) then in healthy controls. This result in a CD4/CD8 ratio in the BPD group being lower compared to healthy controls (p = 0.007; Figure 17C). The monoclonal antibody CD69 is a marker for T-cell activation. Comparing the preterm group with healthy controls, the proportion of CD69$^+$CD4$^+$ T-cells was lower (p = 0.01; Figure E2A paper III). When comparing the BPD group with the preterm group, the proportion of CD69$^+$CD8$^+$ T-cells increased (p = 0.01; Figure 17D).

Regarding FoxP3+ T-cells, the BPD group as well as the preterm group had lower percentage of FoxP3+CD4+ T-cells than the asthma group and the preterm group had lower percentage FoxP3+CD8+ T-cells than the asthma and healthy groups.

ELISA was used to analyze the CD8 specific markers of activity granzyme B and perforin was analyzed. The BPD group and the preterm group had lower concentration of granzyme B than healthy group (p=0.4, Figure 5A paper III) and asthma group (p=0.02 for both Figure 5B paper III). No significant alterations were observed for perforin over the groups.

Significant correlations to FEV1/FVC z-scores were observed for T-CD8+ cells (r=-0.41, P=0.01; Figure E3B paper III) for the pooled preterm groups. There were also significant correlations for the pooled preterm group between FEV1 z-scores and T-CD4+ cells (*r*=0.38, *P*=0.026, Figure 6A paper III) as well as T-CD8+ cells (r=-0.44, P=0.009; Figure 6B paper III). Further, significant correlations to reversibility FEV1 z-scores were observed for T-CD4+ cells (*r*=0.49, *P*=0.005) as well as T-CD8+ cells (r=-0.47, P=0.007); (Figure E4A and B paper III). Finally, PD 20 also correlated significantly with for CD4+ T-cells (r=0.47, P=0.02) as well as CD8+ T-cells (r=-0.60, P=0.002); (Figure E4 D and E paper III) with the pooled preterm groups.

Decreased levels were observed in naïve T-cells and central memory cells in the preterm group (Figure 3A and B in paper II)

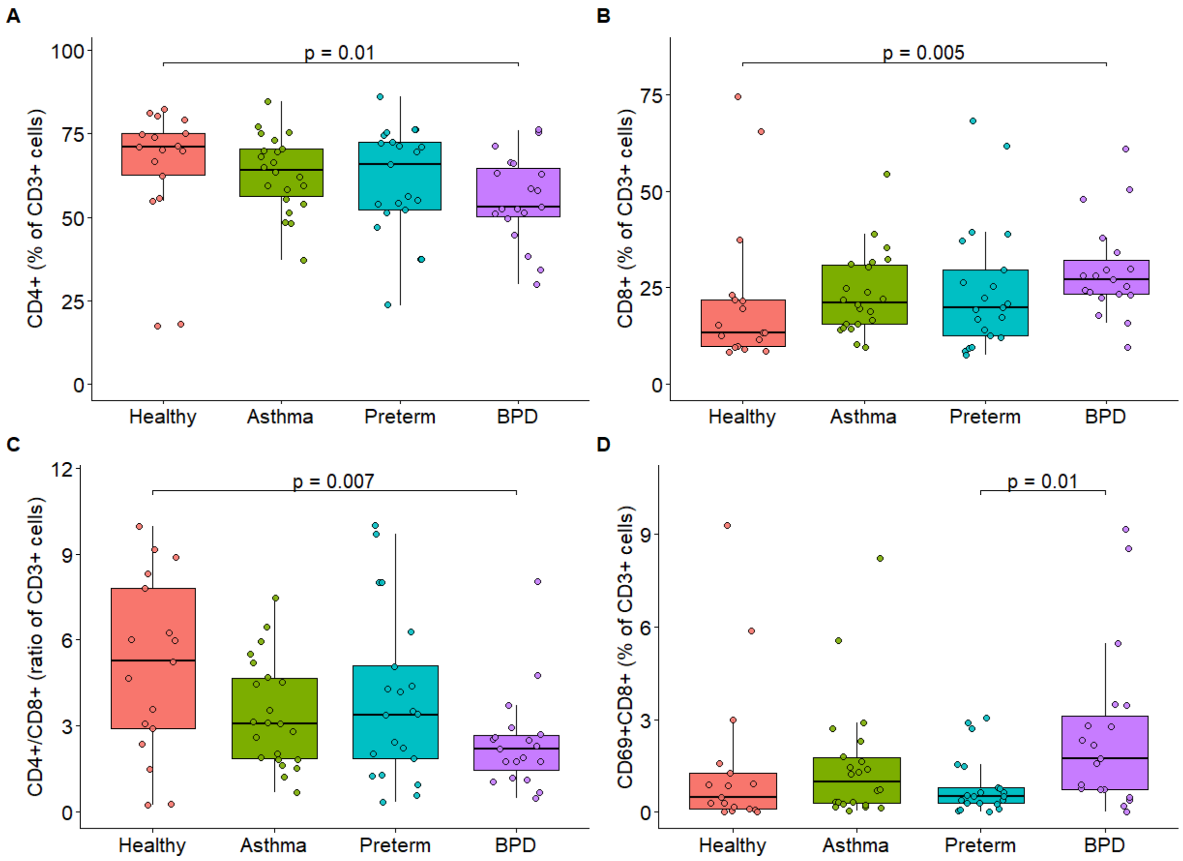


Figure 17. *Comparing percentage CD4$^+$, CD8$^+$, ratio CD4$^{+/}$ CD8$^+$ and CD69$^+$CD8$^+$ of CD3$^+$ cells in BPD, Preterm, Asthma and healthy groups analysed using FACS. Significance using wilcoxon rank sum test is indicated.*

5.3.2 Discussion

In this paper the earlier known decrease in FEV$_1$ in young adults born preterm with BPD were linked to alterations in T-cell subsets. BPD was linked to elevated levels of CD8+ T-cells and decreased levels of CD4+ T-cells. This also correlated with poorer lung function with elevated percentage of CD8+ T-cells, and better lung function with more CD4+ T-cells. Higher levels observed for activated T-cells in the BPD group were also in line with the mentioned effects.

Altogether this points in the direction that preterm born subjects with BPD have more CD8+ T-cells and lower CD4+ T-cells in their lungs. This indicate a disturbed balance between CD4+ T-cells and CD8+ T-cells.

There is also indication that preterm birth in itself without BPD have altered T-cell content in their lungs, as shown by alteration in the preterm group alone.

A number of studies have seen increased levels of FoxP3+CD4+ T-cells in BAL from smokers mainly with bronchitis [197] [198]

These results correlate with findings in patients with COPD related to smoking. The elevated CD8+ T-cells in BAL in young adults born preterm who developed BPD is in line with previous studies, indicating higher levels of CD8+CD103+ T-cells in BAL cells of smokers with and without COPD [197]. This strengthens the view that CD8+ T-cells have a tissue-damaging effect, with a negative correlation to FEV_1 [196].

There were also contradictory results, with decreased FoxP3+CD4+ cells in BPD and preterm compared to asthma, while the levels of FoxP3+CD4+ were increased in smokers compared with never-smokers [197]. It has been proposed that FoxP3+CD4+ has a protective role. It is plausible that smokers with COPD have a protective effect of FoxP3+CD4+ that is lacking in individuals born very-to-extremely preterm, both with and without BPD. This would make this group particularly susceptible to smoking-induced COPD, and life style choices may be of great imporantance to these individuals.

5.4 PAPER IV

5.4.1 Results

Multiomics analysis was used to subphenotype asthma patients from the U-BIOPRED (Unbiased BIOmarkers in PREDiction of respiratory disease outcomes) cohort using omics data obtained from blood and urine. Transcriptomics using microarrays from Affymetrix in whole blood, plasma lipidomics using mass spectrometry (MS), urine metabolomics using MS, and serum somalogics using SOMAmers from 223 severe asthmatics who do not smoke, 77 severe asthmatics who smoke, and 72 mild to moderate asthmatics were subjected to similarity network fusion (SNF) followed by consensus clustering. This resulted in 16, eight and four stable clusters.

To try to extract the variables that drove the separation of the clusters more easily and also to extract the clinical differences between the clusters, more homogenous clusters were created. Subjects were identified that grouped together independently on the number of clusters in an attempt to weed out the core subjects that are driving the clusters. First, subjects that cluster similarly between the eight and 16 clusters were selected. If less than 50% of the subjects in a group in cluster 16 also clustered together in cluster eight, the whole cluster 16 group was removed. Next, subjects that clustered similarly between clusters one and eight were selected. The remaining subjects were called focused subjects, and created so-called focused clusters. The procedure is shown in Figure 2 in paper IV.

The clinical variables most related to the clinical picture of asthma were selected by clinicians in the U-BIOPRED consortium. This preselection increases the possibility of finding statistically significant alterations in the clinical picture. The univariate analysis using a statistical method suitable for the selected clinical features was performed, finding a large number of clinical alterations between the clusters (Table 1 in paper IV).

To further investigate the clinical differences between the groups, OPLS modeling using roplspvs in paper I was performed. Pairwise comparisons were performed between all groups for the three different levels of clustering granularities, with four, eight and 16 clusters respectively, using both the selected clinical variable data set, as well as using the full clinical variable data set. Creating models for all these comparisons resulted in a vast number of models. After attempts to perform this task using the SIMCA software, it became evident that this would be too time-consuming. It was also noticed that there was a need for performing permutation over variable selection, to control for the large differences in group sizes between the different cluster granularities, as well as between the full- and focused cohorts. This was one of the reasons for developing roplspvs workflow in paper I. The application on clinical data also required making it compatible with both quantitative and qualitative data in the independent variables. Creating models including all clinical variables was very time-consuming, and with 1000 permutations the UPPMAX server was needed to perform parallel computing.

The predictivity post variable selection for models of stable clusters on all clinical data over four, eight and 16 clusters are shown in Table 3 paper IV. Using only the focused subjects modeling the 16 cluster groups, 69% of the models were significant. For the eight clusters 76% of focused subjects were significant, and for the four clusters 67% of models were significant. This means that dividing the asthmatics into 16 clusters still gave approximately as many significant models as when using four clusters.

Initially, the models were created comparing groups in clusters four, eight and 16 using the procedure of selecting the number of orthogonals by optimizing Q^2 post variable selection. After editing the script to use the number of orthogonals set by ropls, the 16 cluster models were also created using this method performing the analysis including all subjects and not only focused subjects. These 16 cluster models including all subjects are compared to the 16 clusters including only stable subjects and also showing models pre variable selection is shown in Table 1.

Using all subjects to create models of 16 clusters, 57% of the models pre variable selection were significant (p[Q^2 permutated sans v.s.]<0.05), which resulted in 60% of the models post variable selection being significant (p[Q^2 permutated over v.s.]<0.05). This can be compared to using only focused clusters with 51% of the models pre variable selection being significant and as mentioned 69% post variable selection.

This indicates that only using the focused subjects did not enhance the models to any large extent. Among the models including all subjects, the number of significant models varied greatly between the cluster groups.

Table 1

OPLS model sans variable selection using strategy 3 for focused subjects and all subjects comparing 16-cluster groups showing Q2 of significant models

		cluster	16:1		16:2		16:3		16:4		16:5		16:6	16:7		16:8		16:9		16:10	16:11		16:12		16:13		16:14	16:15		16:16	
		subjects	all	foc.	all	foc.	all	foc.	all	foc.	all	foc.	all	all	foc.	all	foc.	all	foc.	all	all	foc.	all	foc.	all	foc.	all	all	foc.	all	foc.
cluster	n all	n foc.	25	17	26	17	16	10	21	17	29	20	17	15	7	19	11	24	18	18	20	13	19	15	36	27	26	27	8	34	24
16:1	25	17	▨	▨			0.33	0.39	0.36	0.41	0.27	0.41	0.24	0.35		0.29	0.20	0.31	0.37	0.35			0.29	0.29	0.12	0.32	0.28	0.18		0.40	0.59
16:2	26	17			▨	▨	0.18	0.35	0.30	0.27		0.31		0.22		0.15	0.38		0.36	0.33					0.16	0.42	0.10	0.29	0.40	0.44	0.50
16:3	16	10	0.33	0.39	0.18	0.35	▨	▨	0.21							0.35	0.34			0.28	0.25				0.22	0.32		0.50	0.42		
16:4	21	17	0.36	0.41	0.30	0.27	0.21		▨	▨	0.35	0.29		0.31		0.21			0.16	0.24	0.20		0.31	0.29	0.15	0.42	0.37	0.22		0.22	0.23
16:5	29	20	0.27	0.41		0.31			0.35	0.29	▨	▨				0.17	0.28			0.14	0.20				0.35	0.30					
16:6	17	▨	0.24	▨		▨		▨					▨		▨		▨		▨			▨		▨	0.17	▨		0.14	▨		▨
16:7	15	7	0.35		0.22				0.31					▨	▨						0.28				0.19			0.19			
16:8	19	11	0.29	0.20	0.15	0.38	0.35	0.34	0.21		0.17	0.28				▨	▨								0.24	0.24				0.19	0.26
16:9	24	18	0.31	0.37		0.36				0.16								▨	▨		0.16	0.18			0.29	0.38		0.14			0.22
16:10	18	▨	0.35	▨	0.33	▨	0.28	▨	0.24	▨	0.14	▨			▨		▨		▨	▨	0.24	▨		▨	0.29	▨		0.27	▨	0.17	▨
16:11	20	13					0.25		0.20		0.20			0.28				0.16	0.18	0.24	▨	▨			0.17		0.12	0.32	0.36	0.28	0.39
16:12	19	15	0.29	0.29					0.31	0.29													▨	▨	0.08			0.20			0.29
16:13	36	27	0.12	0.32	0.16	0.42	0.22	0.32	0.15	0.42	0.35	0.30	0.17	0.19		0.24	0.24	0.29	0.38	0.29	0.17		0.08		▨	▨	0.24	0.30	0.31	0.32	0.44
16:14	26	▨	0.28	▨	0.10	▨		▨	0.37	▨		▨			▨		▨		▨		0.12	▨		▨	0.24	▨	▨	0.31	▨		▨
16:15	27	8	0.18		0.29	0.40	0.50	0.42	0.22				0.14	0.19				0.14		0.27	0.32	0.36	0.20		0.30	0.31	0.31	▨	▨	0.26	0.33
16:16	34	24	0.40	0.59	0.44	0.50			0.22	0.23						0.19	0.26		0.22	0.17	0.28	0.39		0.29	0.32	0.44		0.26	0.33	▨	▨

OPLS model post variable selection using strategy 3 for focused subjects and all subjects comparing 16-cluster groups showing Q2 of significant models

		cluster	16:1		16:2		16:3		16:4		16:5		16:6	16:7		16:8		16:9		16:10	16:11		16:12		16:13		16:14	16:15		16:16	
		subjects	all	foc.	all	foc.	all	foc.	all	foc.	all	foc.	all	all	foc.	all	foc.	all	foc.	all	all	foc.	all	foc.	all	foc.	all	all	foc.	all	foc.
cluster	n all	n foc.	25	17	26	17	16	10	21	17	29	20	17	15	7	19	11	24	18	18	20	13	19	15	36	27	26	27	8	34	24
16:1	25	17	▨	▨	0.62		0.66	0.74	0.77	0.77	0.61	0.74	0.54	0.62	0.75	0.58	0.72	0.63	0.72	0.53					0.58	0.77	0.54	0.62		0.65	0.72
16:2	26	17	0.62		▨	▨	0.55		0.62	0.71		0.71		0.62	0.75	0.71	0.85	0.72		0.58				0.66	0.63	0.72	0.59	0.63	0.88	0.64	0.74
16:3	16	10	0.66	0.74	0.55		▨	▨	0.52		0.60			0.73	0.84	0.60	0.77	0.72		0.77	0.65			0.74	0.59	0.68	0.66	0.65	0.91	0.68	0.68
16:4	21	17	0.77	0.77	0.62	0.71	0.52		▨	▨	0.55	0.62		0.71			0.74	0.54	0.54	0.51			0.64		0.51	0.63		0.68	0.78		0.58
16:5	29	20	0.61	0.74		0.71	0.60		0.55	0.62	▨	▨		0.61		0.64	0.74				0.61	0.68	0.60		0.52	0.67				0.51	0.60
16:6	17	▨		▨		▨		▨		▨		▨	▨	0.76	▨		▨		▨			▨		▨		▨			▨	0.75	▨
16:7	15	7	0.62	0.75	0.62	0.75	0.73	0.84	0.71		0.61		0.76	▨	▨	0.72						0.84	0.67		0.55	0.69				0.66	
16:8	19	11	0.58	0.72	0.71	0.85	0.60	0.77		0.74	0.64	0.74		0.72		▨	▨		0.78						0.51	0.78	0.69		0.83	0.64	0.75
16:9	24	18	0.63	0.72		0.72	0.72		0.54	0.54							0.78	▨	▨			0.65				0.68			0.75	0.62	0.65
16:10	18	▨	0.53	▨	0.58	▨	0.77	▨	0.51	▨		▨			▨		▨		▨	▨	0.53	▨		▨	0.60	▨		0.51	▨		▨
16:11	20	13					0.65				0.61	0.68			0.84				0.65	0.53	▨	▨			0.66		0.50	0.62	0.82	0.56	0.67
16:12	19	15				0.66		0.74	0.64		0.60			0.67									▨	▨	0.54	0.71			0.81	0.62	0.72
16:13	36	27	0.58	0.77	0.63	0.72	0.59	0.68	0.51	0.63	0.52	0.67		0.55	0.69	0.51	0.78		0.68	0.60	0.66		0.54	0.71	▨	▨	0.56	0.72	0.78	0.54	0.66
16:14	26	▨	0.54	▨	0.59	▨	0.66	▨		▨		▨			▨	0.69	▨		▨		0.50	▨		▨	0.56	▨	▨	0.57	▨	0.59	▨
16:15	27	8	0.62		0.63	0.88	0.65	0.91	0.68	0.78							0.83		0.75	0.51	0.62	0.82		0.81	0.72	0.78	0.57	▨	▨	0.56	0.77
16:16	34	24	0.65	0.72	0.64	0.74	0.68	0.68		0.58	0.51	0.60	0.75	0.66		0.64	0.75	0.62	0.65		0.56	0.67	0.62	0.72	0.54	0.66	0.59	0.56	0.77	▨	▨

foc.: focused clusters　Q²: 0.4-0.6　Q²: 0.6-0.7　Q²: 0.7-0.8　Q²: 0.8-0.9　Q²: 0.9-1　▨ Comparison does not exist

The number of significant models differed between the clusters indicating that some clusters were more related to specific clinical features. Cluster group 13 had 9 models post variable selection that were significant, and also had significant models pre variable selection for focused subjects. Including all subjects, the number of significant models in both pre and post variable selection increased to 13 out of 15 models. Group 16 also had 13 models that were significant post variable selection, but only six of these were also significant pre variable selection. This can be compared to cluster 12, which only had one model that was significant in both pre and post variable selection.

SUS plots were created after developing a script to create SUS plots in R, including variables from both models to be compared.

In addition to the OPLS models, decision trees were used for predictive modeling of the focused clusters to compare the clinical data. Using the selected set of variables resulted in 75% accuracy when including all groups in the eight-cluster granularity, and 89% for the four-cluster granularity. This showed that the number of neutrophils was important at the eight-cluster level, while the number of eosinophils were important at the four-cluster level.

5.4.2 Discussion

Despite many treatment alternatives for asthma, there are still a sizeable group of asthma patients who do not get relief from the treatments currently available. Many of these patients are classified as severe asthmatics, and experience debilitating symptoms and frequent exacerbations. The overall aim of the U-BIOPRED consortium was to sub-group asthmatics in a data-driven manner based on molecular data, with the aim of identifying the mechanism that differs between the newly identified subgroups of asthmatics. This would allow possible new treatments to be identified. However, it is also of great interest to identify any clinical characteristics associated with these newly identified sub-groups of patients, in order to create synergy between novel molecular findings and established clinical features of asthma. As in any type of clustering, the result will be a core of individuals that define the core clinical features, and a number of less well characterized subjects that may lay in between two or several clusters. Here we applied a method to choose patients who are similar to each other, as determined by them clustering in a similar pattern across the three levels of clustering granularities created, referred to here as the focused clusters.

The clinical differences were investigated using OPLS modeling resulting in significant differences between cluster groups in 4 group, as well as in 8 group and 16 group clusters using OPLS modeling. Similar proportion of models were found significant using the different number of groups in clusters. This indicate that there were clinical differences in between all cluster groups. The significance was to a large portion confirmed using permutation pre variable selection. Comparing the models of 16 clusters using all subjects resulted in only slightly larger number of significant models post variable selection but slightly smaller number pre variable selection. This could be an effect of a slight overestimate of models post variable selection using fewer subjects in the focused clusters. It could also be due to a real advantage of selecting variables using the more homogenous subjects in the focused cluster producing more significant models post variable selection. The clinical differences were also investigated using random forest for predictive modelling. This also confirmed that there were clinical differences.

This study is important as it identifies clusters of asthmatics with diverse clinical characteristics. This has previously been performed for COPD [160] in our group, which has shown that integration of multiple omics platforms from same subject may improve the statistical power significantly, and accordingly the sample sizes may be reduced without a penalty in the accuracy of group classification [160]. In that study, the number of subjects in the groups could be decreased to n=6 when integrating data from seven omics platforms, while still maintaining 95% accuracy. Here, four platforms were used, obtaining 89% accuracy when clustering the subjects into eight cluster groups, with the resulting group sizes ranging from n=10 to n=37. The slightly inferior performance may be due to the larger variability allowed in the inclusion criteria in the U-BIOPRED cohort, where e.g. oral corticosteroids were allowed. Also, the multicenter study design which involves samples being collected by different people using different equipment is bound to add to the overall variance of the study. In comparison, the Karolinska COSMIC cohort that the multi-omics power calculation study was performed on is a single center study with inclusion and exclusion criteria designed to create a homogenous study group.

The weakness of this study is that no validation cohort is available to test the identified clustering models. This is a common dilemma with extensive multi-omics cohorts such as U-BIOPRED and the Karolinska COSMIC cohort.

However, the fact that the clinical differences observed between the groups identified by the multi-molecular clustering are highly significant provides a means of validation. Using univariate analysis, the most significant differences were the number of neutrophils, the use of daily oral corticosteroids, and alterations in alpha 1 microglobulin levels. While these are known variable of importance in asthma phenotyping, the more complex clinical picture provided by the significant OPLS models may be more informative.

The clinical characteristics of several of the newly identified sub-groups, particularly in the 16-granularity clusters, are likely to describe actual asthmatic endotypes that should be treated differently. Thus, this represents a small first step towards being able apply a more personalized treatment for patients with asthma.

This paper show that there are eight or even 16 different sub-clusters of asthmatics with distinct clinical properties. This is an important step towards personalized treatment of asthma, which consists of many disease endotypes.

5.5 PAPER V

5.5.1 Results

Metabolomics analysis was performed on urine for 310 severe asthmatics who do not smoke (SANS), 108 severe asthmatics who smoke or have smoked, 87 mild to moderate asthmatics (MMAs), and 100 healthy controls from the U-BIOPRED adult cohort.

To investigate which types of molecules were connected to alterations due to asthma, hierarchical cluster analysis was performed. This clustering of metabolite abundance identified seven groups of metabolites with alterations between the asthmatic groups.

Using univariate methods, significant alterations ($FDR < 0.05$) in 40 metabolites from urine were found to be due to asthma. Severe asthmatics showed lower abundance of metabolites in clusters A (amino acids) and B (amino acids) and higher abundance in clusters C (carnitines), D (mixture of metabolites), and F (dietary and drug metabolites). To investigate how the metabolites were affected by asthma treatment, the analysis was stratified by treatment. Alterations in 23 metabolites due to oral corticosteroid (OCS) use were identified. Most of the metabolites were affected by OCS treatment, but clusters C and F were identified to be unaffected.

Additionally, the multivariate method principal components–canonical variate analysis (PC-CVA) showed that the strongest alterations that were not affected by OCS use were found in carnitine levels.

Gene set variation analysis (GSVA) enrichment score (ES) was used to connect the genes to enrichment in pathways. The results showed that β-oxidation decreased with asthma severity in sputum and that fatty acid metabolism decreased with severity in sputum and in bronchial brushings. It was also shown that the carnitine transporter SLC22A5 was lower in severe asthmatics.

To investigate what drove the alterations in carnitine levels, the entire collection of clinical data from the U-BIOPRED cohort was used. The carnitines were converted into a carnitine component, and the subjects with the highest versus the lowest quartiles were compared. Using roplspvs, this resulted in a significant ($p[Q^2$ perm. over v.s.$] < 0.05$) model, with $R^2 = 0.30$ and $Q^2 = 0.21$. The loading plot (Figure 18) shows that gender was the largest driver for separating subjects with high and low carnitine levels. In the MMA, sans, and healthy groups, the carnitine levels were higher in males (Figure E5A in paper V), and the alterations in carnitine levels between asthma groups were driven mainly by males (Figure E5B in paper V). The carnitine transporter SLC22A5 decreased with asthma severity in both sputum and bronchial brushings, and these alterations correlated with the FEV1% of predicted value. The pattern was also followed for TAC classifications, where TAC1 and 2 were lower compared with TAC3 in sputum fatty acid metabolism, and the sputum SLC22A5 expression was also lower. When comparing groups of different T-helper cells class 2 (TH2) classifications, the connection was not as clear. The low TH2 group had significantly lower carnitine scores than the healthy controls did, but the high TH2 group did not show an altered carnitine score. A clear decrease was also observed for SLC22A5.

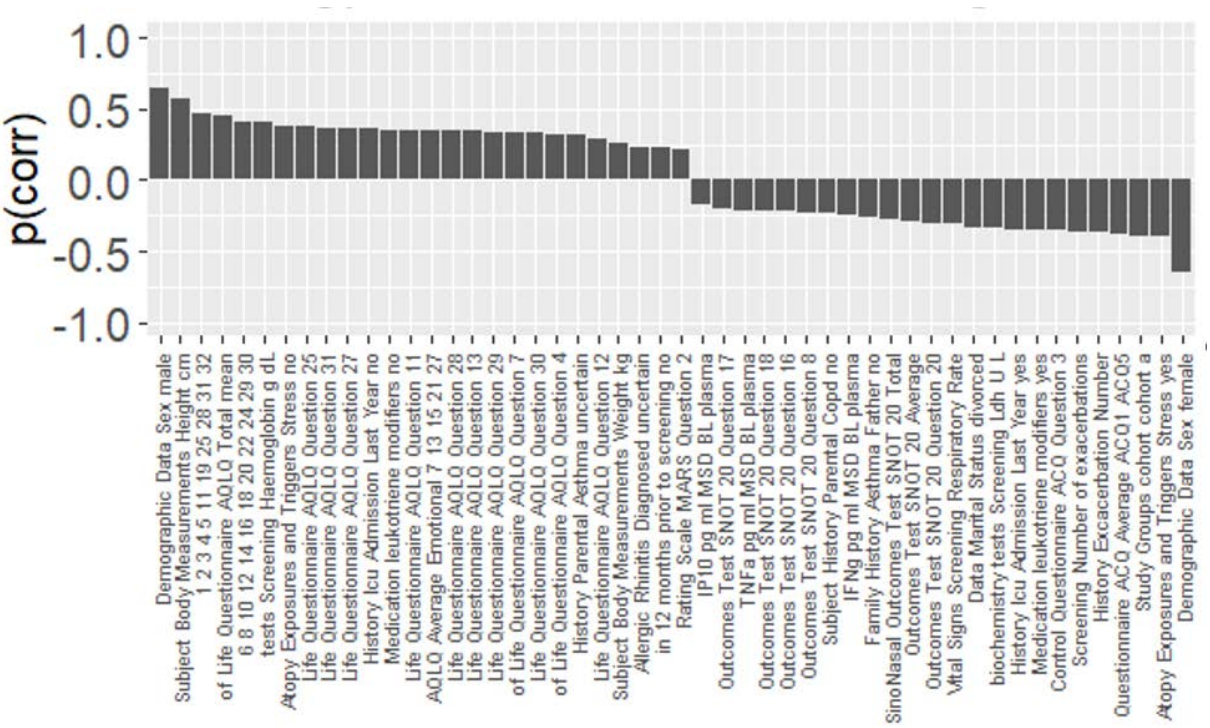


Figure 18. *Loading plot of clinical variables driving the separation in OPLS model comparing subjects with highest quartile to lowest quartile of carnitine levels. P(corr): Scaled loading of clinical variables*

5.5.2 Discussion

The decreased levels of carnitines in severe asthmatics were identified that was independent of OCS treatment.

The reduced carnitine levels, reduced β-oxidation and fatty acid metabolism, and lower levels of carnitine transporter SLC22A5 together with the lower carnitine levels in severe asthmatics indicated mitochondrial dysfunction.

Both free carnitines and acetylated carnitines have previously been shown to be decreased in females [199]. Oral contraceptive use further reduces levels of free carnitines and acetyl carnitines compared with nonuse [200]. As mentioned in paper I, the carnitines as a group, specifically the acyl carnitines, were also decreased in by hormone replacement therapy in postmenopausal women [201]. Acyl carnitines were also, together with 12-hydroxyeicosatetraenoic acid, the most driving metabolites separating patients with COPD from smokers [90]. In the same cohort, the ratio between medium and long-chain carnitines was lower ($p < 0.05$) in females with COPD versus female smokers, indicating fatty acid β-oxidation. Women are also more affected by COPD, and women display a faster decline in lung function after menopause [202]. This suggests that carnitines might have a role in the known connection between lung disease and sex hormones [203]. One study showed that reduced L-carnitines levels were associated with emphysema progression in a mouse model, and that supplementation improved their lung function [204].

In conclusion, this represents the first large-scale study identifying urine metabolites associated with OCS- and asthma disease, with carnitines as the largest OCS-independent drivers. This suggests that mitochondrial functions might be disturbed in asthma and that carnitine supplementation could be a potential therapy.

5.6 PAPER VI

5.6.1 Results

The proteome stability in EDTA-plasma was investigated with multiplexed MS-based proteomics using tandem mass tags (TMT). Blood samples from healthy donors, which had been stored for one, three, eight, 24, and 36 hours at 24°C before centrifugation, were analyzed using both univariate and multivariate statistical methods. Multivariate analysis using PCA and OPLS modelling in SIMCA software was performed to compare samples from each time point with the one-hour sample (baseline) as well as with each adjacent time point. The analysis was performed both on joint gender groups and stratified by gender. The analyses were performed both at the protein level, at the tryptic peptide level, and using also semi tryptic peptides.

PCA analysis showed that the sets had a tendency to cluster, indicating batch effects (Figure 1 in paper VI). We therefore attempted to correct for theses batch effects using different methods for normalization, including the selection of stable peptides as normalization factors, using quantile normalization as well as normalizing across batches and subjects. No improvement in model performances was observed after normalizations, and analyses were performed without further normalization. Also, PCA showed that all subjects were within Hotelling's T^2 95% confidence interval and was not altered by normalization indicating that no normalization was need.

Analysis using protein level information revealed no significant models comparing the time points to each other. To force a model a method of overfitting the models before variable selection was applied. This removes the advantage of using multivariate methods and approaches a method using univariate methods to filter variables.

Comparing models in terms of protein level, resulted in significant models (CV-ANOVA<0.05) for all timepoints comparing to baseline (Table 1 paper V1). Comparing adjacent timepoints resulted in significant models (CV-ANOVA<0.05) only for the 24-hour sample compared to the 36-hour sample (Table 2 in paper VI). The models comparing time points stratified by gender did not result in better performing models.

When comparing timepoints using peptide level information, better performing models were obtained when including only tryptic peptides, with all models comparing against baseline as well as adjacent time points being significant. After stratifying by gender, models comparing against baseline were also largely significant, but as no major improvement was achieved by gender stratification, analysis of the joint gender groups was prioritized.

To also include degradation-products of proteins degrading during the delay of processing, semi-tryptic peptide levels were also investigated. Semi-tryptic peptide models were created, but no enhancement of models compared to tryptic peptide models was observed.

As the OPLS models of the joint gender comparing tryptic peptide levels were the most significant, we compared the peptides that were driving these models. Comparing peptide level models comparing 3, 8, 24 and 36 hours to one hour samples resulted in 11, 7, 18 and 5 peptides respectively driving the OPLS models (Table 3 in paper VI). The same comparisons resulted in 3, 1, 3, and 34 peptides respectively being altered ($p<0.05$) using univariate analysis (data not shown). Comparing adjacent time points i. e. 3 versus 1 hour, 8 versus 3 hours, 24 versus 8 hours and 36 versus 24 hours resulted in 11, 7, 14, and 6 peptides respectively driving the OPLS models (Table 4 in paper VI). Using univariate analysis comparing the same adjacent time points resulted in 3, 1, 1 and 20 peptides respectively being significantly ($p<0.05$) altered using univariate analysis (data not shown).

The peptide with sequence NIQSLEVIGK was shared by all models (Figure 19) and was also significant ($p<0.05$) comparing peptide levels at all time points to one hour sample using univariate analysis. The stability profile showed increasing levels over time (Figure 20A).

MFLSFPTTK was shared driving models between 3, 8 and 24 hours to 1 hour sample (Figure 19).

The peptide with sequence IDSLLENDR was the peptide most driving the model between 3-hour and 1 hour sample (Figure 19) and also driving the gender separated models. The degradation profile had a minimum of peptide levels at 3 hours (not shown).

The primary driver of the separation between 8- versus 1-hour time points was AQGYSGLSVK, and it was also one of the most driving peptides of the separation between the 24- versus 1 hour samples. The degradation profile showed increasing levels of peptide during the 36-hour period (Figure 20B).

DYIEFNK was the most driving out of the 5 peptides driving the separation between 1 and 36 hours (Figure 19). LLGEVDHYQLALGK was share between the peptides modeling the separation of 1 versus 24-hour and 1 versus 36-hour models (Figure 19).

Taken together that no significant models were obtained without using an overfitted model, and that the number of peptides significantly altered ($p<0.05$) was very low with NIQSLEVIGK being the only significantly altered after adjusting the p-value, contribute to the conclusion that the proteins were stable over the time period studied of 36 hours. Also, calculating relative standard deviation (RSD) over all subjects and all time points at 22°C

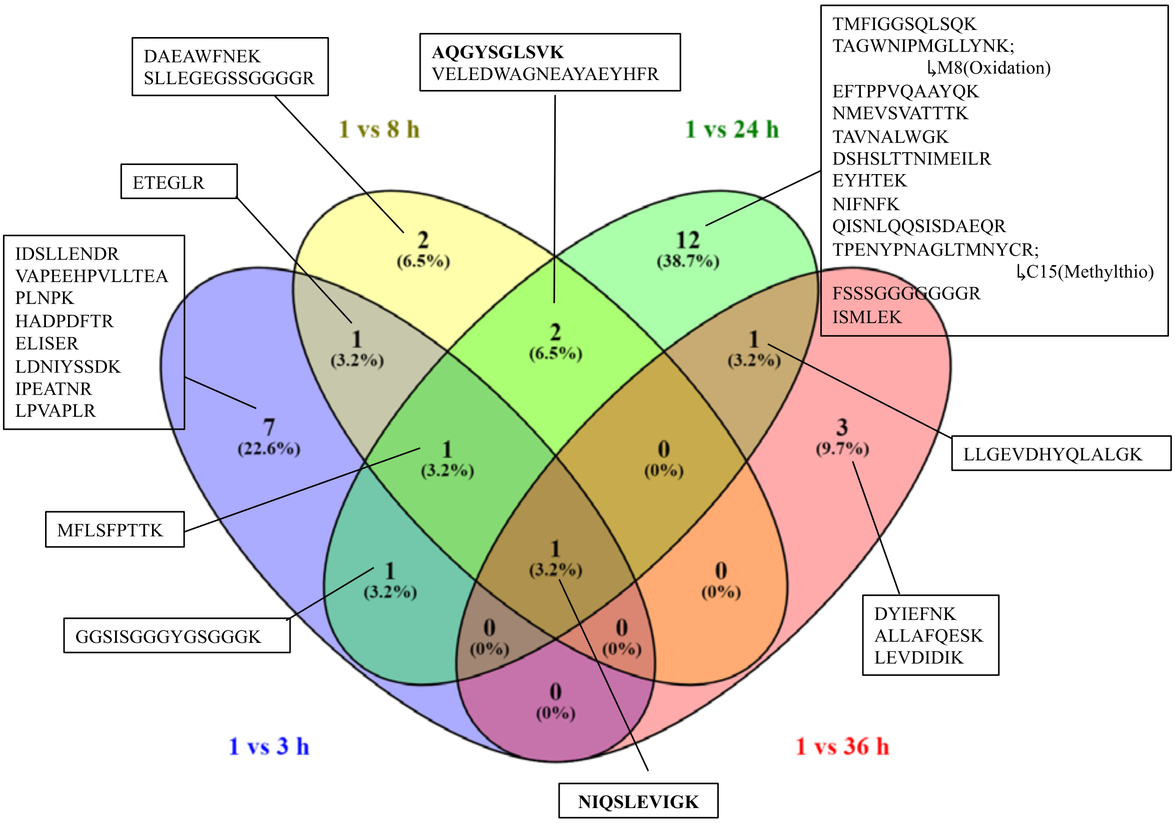


Figure 19. *Venn diagram showing OPLS models comparing ratio of peptide levels at 3, 8, 24 and 36 hours to 1 hour storage at 22°C. The sequences and numbers of shared and unique tryptic peptides driving the models are displayed.*

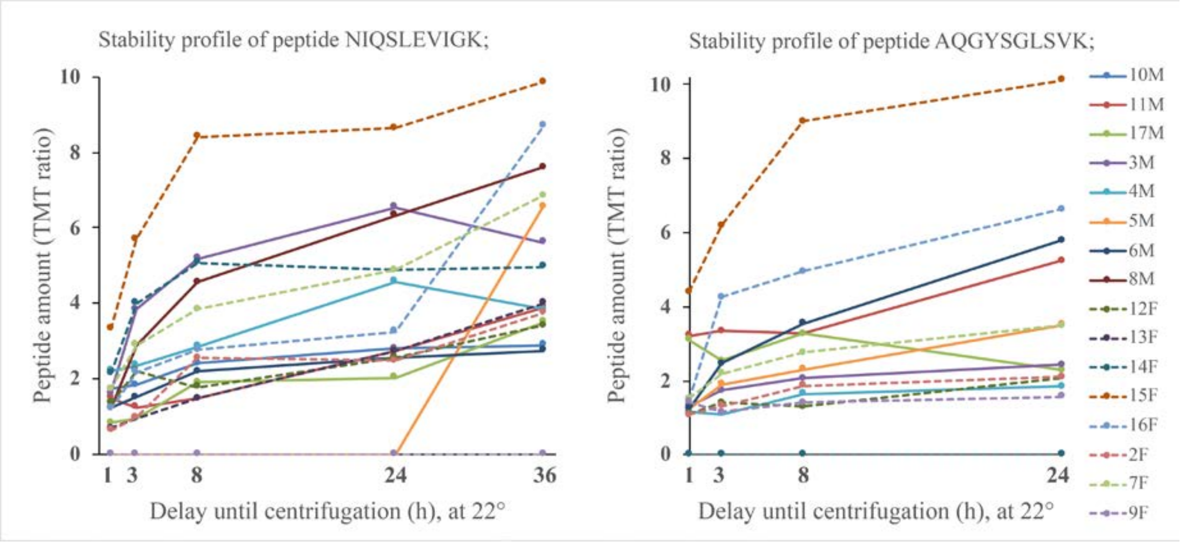


Figure 20. *Stability profiles for peptide with sequence NIQSLEVIGK which was significantly altered at all time points (p<0.05) and included in all OPLS models comparing to levels at 1 hour and AQGYSGLSVIK which was altered at 24 hours (p<0.05) compared to levels at 1 hour and included in OPLS model comparing 8 and 24 hours to 1 hour.*

resulted in 95.5% of the peptides having RSD lower than 0.3 (Figure E3 in paper VI) also confirmed that the proteins were stable.

5.6.2 Discussion

Plasma samples have sometimes been processed with delays before being stored in biorepositories. As such, the aim of this study aimed at finding biomarkers for degradation to be able to determine if samples have been exposed to processing delays. The stability of proteins in EDTA-plasma was assessed to investigate effect on the proteome due to delays in processing before freezing. Mass spectrometry with isobaric tandem mass tags (TMT) was used to quantify protein degradation after incubating samples at 1, 3, 8, 24 and 36 hours in room temperature. Analysis was performed both at the protein and peptide level, as well as by including semi tryptic peptides in the evaluation.

OPLS modeling and univariate methods were used to compare the levels of peptides at all time points to the baseline levels at 1 hour, and also comparing all adjacent time points to each other. No significant models were obtained using the regular workflow, showing that the peptides and proteins were stable in plasma over the time period studied of 36 hours at 22°C. Taken together, the number of peptides altered between the time points were very limited when assessed with univariate statistics, and only one peptide was altered after adjusting the p-value (FDR< 0.05). The fact that more than 95% of the peptides had RSD below 0.3 further confirmed that most peptides were stable.

Despite that most peptides were stable, further analysis using multivariate and univariate methods was performed to try to identify also minor alterations, and to find biomarkers of such processes. Models were created using an overfitting procedure in the models pre variable selection. NIQSLEVIGK was the only peptide driving the separation between all time points compared to the one hour baseline sample (Figure 18). The peptide is derived from the pro-platelet basic protein (PPBP, a.k.a. CXCL7, Uniprot accession no. P02775). PPBP is released from platelets when they are activated, and can be cleaved into 10 polypeptide chains with different biological functions, all containing the NIQSLEVIGK sequence. PPBP is precursor of the 2 platelet alpha-granule proteins, the platelet basic protein (PBP) as well as the connective tissue-activating peptide III (CTAP3) which are further processed into beta-thromboglobulin (TGB) and neutrophil-activating peptide-2 (NAP2). NIQSLEVIGK showed increasing levels over the time period studied. NIQSLEVIGK is also a biomarker for preeclampsia, a complication in pregnant women, which has been validated [205].

IDSLLENDR showed quickly decreasing amounts and was the most significantly altered tryptic peptide after 3 hours. It is derived from Clusterin (P10909) which is a protein involved in apoptosis [206].

AQGYSGLSVK is a degradation product from the adhesive glycoprotein thrombospondin-1 (P07996) which is involved in cell-to-cell interactions and platelet aggregation [207]. This peptide was altered both at the 8h and 24 h time points as compared to the baseline.

DYIEFNK was the peptide most driving the model comparing 36-hour samples to 1 hour sample and is derived from Zinc-alpha-2-glycoprotein (P25311) which is a protein stimulating lipolysis.

Shen et al [167] has analyzed the same samples using multiplex proximity extension assay which uses antibodies. They found that 40 proteins out of 139 analyzed were altered using Bonferroni corrected p-values ($p<0.05$). They suggest that the altered peptides were a result of proteins leaking out of cells into plasma during delay to processing, which explain the increasing levels we observed in the AQGYSGLSVK and NIQSLEVIGK peptides. They did not identify alterations in PPBP, as it was not included in their panel. This suggests that there is a need to run both targeted and non-targeted proteomics methods when searching for biomarkers.

This study showed that proteins in EDTA-plasma were stable over the 36-hour time period studied at 22°C and also samples at 4°C were stable as analyzed using TMT-MS. This means that the effect of processing delays up to 36 hours before processing and storing samples in freezers probably has minor effect on the result if protein levels are to be studied. Still, a few altered proteins were identified using the multivariate method OPLS. If these specific protein classes or platelet-related processes are to be studied, it is important to add freezer as soon as possible since release of these proteins occur within the 3-hour time frame. Given the robustness of the ROC curves generated with the identified peptides (AUC= 0.89-0.98, Figure 2 paper VI), the findings from this project have the potential to be formatted into biomarkers of stability to create a model that can predict how long the samples have been delayed before freezing. This would have potential to contribute to better utilization of biobank repositories, which are currently underused.[208-210]

6 CONCLUSIONS

The three diseases studied in this thesis – asthma, COPD, and BPD – are all obstructive lung diseases. They are mainly caused by environmental factors, but are also influenced by genetic factors. Both COPD and asthma are so-called umbrella diseases, which means that they show a whole range of endotypes with some being treatable while others show symptoms despite treatment with currently used therapeutics.

Alterations between subphenotypes of asthma and COPD were identified using several platforms, including miRNA in EVs, metabolome, and lymphocyte composition, as well as clinical data. The alterations in miRNA cargo of sEVs in smokers with COPD compared with healthy never smokers – as well as alterations in metabolome when comparing severe asthmatics with mild asthmatics – were connected to alterations in pathways, indicating the mechanism for the endotypes.

The workflow and package in paper I created to perform OPLS-DA analysis in R show the need for better approaches to significance testing than the commonly used permutation post variable selection, since this approach was shown to produce significant models using random data. Performing permutations sans variable selection was shown to better separate significant models from insignificant models. Finally, permutation over variable selection ensures that the variable selected can still be used to separate the groups significantly. It was also shown how decreasing sample sizes resulted in increasing R^2 and Q^2 of the models, and that permutations over variable selection are a means of testing whether the models are still significant when the sample sizes are small. The tool is user friendly also for people who are not familiar with R, and provides a tool to efficiently compare many study groups; these stratifications can be used on all kinds of data in a reproducible workflow, offering the possibility to provide a useful and necessary tool for extracting biomarkers that can be used not only for airway diseases, but also for other diseases.

The workflow was applied to groups of subjects affected by COPD and BPD, as well as asthmatics, to identify alterations in a range of compartments.

In paper II the alterations in miRNA cargo in EVs due to COPD and smoking linked to the p53 signaling pathway as well as other cell growth and cell death pathways seemed mostly related to smoking, while pathways related to autophagy, mitophagy, and tight junctions were more often linked to COPD alone. These pathways have protective effects in other diseases, and it is conceivable that the effects of smoking observed in the sEVs could contribute to protecting against the effect of smoking. This could open up a new area of treatment for COPD, and would be enormously welcome as there is currently no cure for this deadly disease.

In paper III the elevated CD8+ T-cells and reduced levels of CD4+ T-cells in BAL among young adults born preterm who developed BPD identified in paper III indicate the same damaging effect of CD8+ T-cells in both COPD caused by smoking and young adults born preterm who developed BPD. Multivariate analysis using OPLS with stratification by gender indicated that the alterations in CD4+ and CD8+ were mainly driven by females. This is in line with female infants with BPD needing oxygen for a longer time than male infants [211]. The understanding of the BPD mechanism is a step towards managing the malfunction of the lungs in youths born with BPD.

In paper IV, the roplspvs method developed in paper I was used to identify clinical differences in clusters of asthmatics that had been clustered by means of integrated multi-omics data into four, eight and 16 clusters. The clinical differences that were identified by pairwise OPLS comparisons were confirmed using permutation pre, post, and over variable selection in the roplspvs workflow. This indicate that asthmatics may be divided into separate endotypes and is a step towards more individualized treatment of asthma.

Our finding in paper V, that carnitines were the most altered metabolites not related to OCS demonstrates that severe asthma may be linked to mitochondrial dysfunction. The gender differences identified indicate that carnitine levels may connect sex hormones and lung disease. The potential therapeutic target provided by carnitines is an important contribution to the field, as severe asthmatics lack efficient treatment for their symptoms on top of the incurable disease that asthma represents.

In paper VI, the findings that most proteins in plasma were stable at 22°C for 36 hours may contribute to increased use of biobank repositories in a range of biological areas. However, given that proteins related to platelet activation were identified as altered as quickly as withing 3 hours raises some concerns. In studies associated with these processes or proteins, it may be important to assess the quality of biobank samples used, or to place samples in freezers as soon as possible.

7 POINTS OF PERSPECTIVE

The studies in this thesis are a small part of a huge work to find the mechanisms of all the endotypes of asthma, COPD and BPD that affect people. Analyzing all different kinds of omics data and decipher the mechanisms of these diseases is an enormous challenge. The technology improvements have taken huge steps in recent years enabling analyses that were not possible a few years ago. Next generation sequencing together with development of computer technology has allowed to routinely perform whole genomes sequencing. Sensitivity of analysis has developed to the degree that it is now possible to analyze contents of single cells.

This revolution in omics data analysis including more platforms result in a bottleneck in analyzing all data. The challenge of putting data together from many platforms to understand the big picture of systems biology has just begun. New tools are being developed enabling easier and more robust analysis.

One challenge is to know that the findings are truly significant. This is challenging when the number of variables studied are in another dimension than when the statistical tools were developed making the risk of false findings increasing largely.

This has brought a growing need to validate findings in other cohorts and this also increases the need for collaborations. For this reason, among others, has led to collecting samples to be studied in biobanks which is accessible for all researchers. A challenge is the need for each platform to have their own quality controls.

Both collecting data and analyzing data use a lot of computer power this has brought computer clusters into frequent use. This brings another need for large collaborations.

Last but not least has the knowledge bank grown enormously bringing another need for collaborations as one person cannot know everything. With collaborations there is a great chance of solving many puzzles.

The long-term goal is to identify endotypes enabling individualized medicine to find diagnostics and therapeutics for everyone with asthma or COPD. As these diseases affect around 300 million people each and COPD is the third or fourth leading cause of death, this would substantially enhance the quality of life for a large number of people all over the globe.

8 ACKNOWLEDGEMENTS

Thank you to all participants of the three cohorts that I have studied. Thank you all who have contributed with money to research. Without your contributions this research could not have been performed. In addition, I would like to thank all my friends and family for the support that I have received performing this study and I would especially like to thank:

My main supervisor, Åsa Wheelock, for giving me the opportunity to do and for being my great support throughout my PhD studies. Thank you for giving me energy when I needed and encouragement when I lost fate in myself. Thank you for understanding my needs when I did not myself. Thank you for giving me freedom to investigate the models and spend longer time than I ever thought was needed into my scripting. Thank you for encouraging me to celebrate successes and share my troubles. Thank you also for nice small talks in between work.

My co-supervisor Susanne Gabrielsson, for introducing me into the world of exosomes and having me at your lab meetings. Thank you also for your patients with the manuscript that I never finished.

My co-supervisor Magnus Sköld for showing me the bronchoscopy procedure. Thank you for your interest in my progress and continuously asking how things were going.

My co-supervisor Eva Berggren-Broström for your warm engagement and patience. Thank you for showing me the neonatal ward. Thank you also for the company and encouraging words on our trip to Gothenburg.

Johan Grunewald head of Respiratory Medicine Unit for setting a nice, warm and open-minded atmosphere at the unit.

Peter Savolainen my mentor for pushing me to make timelines. Thanks also for all the very concrete advice finishing the thesis up.

Bengt Sennblad my mentor at NBIS for your advice on bioinformatics and R programming. Thank you also for nice peptalk.

My colleagues at the Respiratory Medicine Unit for making my years at the Unit memorable. Benita D for making me feel welcome at the department and helping with all practicalities and introducing med to the Öron näs o hals unit with whom we have shared lunches with lots of laughter. Benita E for guidance in the laboratory performing endless amounts of ficol separations and for your always very frank and good advice on how to solve problems of all diverse natures. Maria for your great cheering while writing my thesis. It meant a lot to me and made me believe that I would succeed when I doubted it. Thank you also for proof reading. Benedikt Z for our nice PhD fikas sharing the difficulties and pleasures of being a PhD student. Thank you also for nice feedback on my presentations. Tracy for sharing my experience of mindfullness practice and sharing wise thoughts and advice about data handling and life in general. Tina Heyder for paving the way as a PhD student and introducing me to the labwork of isolating HLA peptides. Emil for sharing your thoughts of the science of medicine. Emma Ringquist for introducing me to FACS analysis and showing me your lab at Huddinge. Anders Lindén for your interest and questions. Jing for showing your great commitment to science. MingXing for introducing me to the OPLS analysis workflow and giving me advice also after moving from our group. Jan Wahlström for always being there for a chat and for giving nice feedback on halftime review presentation. Ylva for showing me that it is possible to enjoy presenting a thesis. Avinash for being a great desk neighbor. Karlhans for sharing your perspectives.

Emma Karlsson for taking care of all practicalities of administration leaving me to spend more time on research. Thank you also for taking your time for small talks.

Research nurses Helen Blomqvist, Margitha Dahl and Gunnel de Forest for collecting samples and always meeting up with smiles and being nice company at the clinical seminar series.

LUNAPRE team Petra for showing the way to a PhD with encouragement.

Biobank team Gunnel and proteomics Carina and Britt-Marie for having me visit the Karolinska biobank and the proteomics facility in Gothenburg.

My friends Elisabeth, Lars, Sara for always being there. And for all our fun new year's parties!

Annika, Jonas, Linn and André Solehav for all midsummers, dinners and fun together. Annette for being my private mentor telling me how to think of my PhD studies and what is important. For all our nice lunches. And for being my great support and friend.

Family Fredrik, Erika, Rikard, Christina, Martina and Alexander Lindvall Kiss for all the fun parties on the bridge!

Annika and Tomas Liljenberg for the greatest party. I had so much fun!

Cia for all windsurfing we did together.

Sonja, Laura, Vincent and Michelle von Zeipel, what a skitrip we had together!

Lena W and Lena H for keeping my head cool in Mälaren and for being my extra fantastic cheering mentors. Hope for many more swims this winter!

Monica, Tina, Fredrik, Helena and Johan. Thank you for being the best stepfamily I could have! Thanks for all fun together!

My aunt Ulla for all the laughs, thoughtfulness and love! Jill and Jens with families for all good times at Lindskär over the years. Hope many more will come.

My extra parents Marylee and Dick Vanderschuere, thank you for welcoming me and my whole family to your home. We had the most wonderful time and the greatest memories forever.

My most precious brother Peter and sister Helena for always being there no matter what. For making me take care of myself, helping with taking care of our summer house and helping our parents so I could focus on my thesis. Peter for giving me frank and to the ground observations and telling me how things are and always with a smile. Helena for always having new ideas and encouraging me more than anyone else. Thank you both for all skiing, walks, hikes and skating. I could go on forever. Kennet for being the greatest brother in law always being there with a smile and a hand. Emil for being my wonderful sisterson and godson! Hugo for being my lovely sisterson! I am so happy that it is you all we share summer place with!

My dearest father Richard Stenbeck for making me always feel loved, always being there for me and supportive of everything I do. Thanks for always being energetic taking the whole family to sailing every summer, to skiing and to skating. You are also the best advisor when it comes to economics, history, politics mm mm

My outstanding mother Britt-Marie Stenbeck for always being the most curious about my private life, always being the most caring and thoughtful mother. I am utterly impressed with how you have handled your difficulties during the recent years, with thankfulness in the most difficult times.

Johanna for being the loveliest daughter. For taking care of all humans, animals and plants. For always having new projects! I love you so much!

My lovely son David for being so caring and thoughtful. Thank you for being down to earth and sharing your bright views. I love you so much!

My lovely son Martin for taking care of all spiritual matters! Always giving new perspectives to things. I love you so much!

Last and most my beloved husband Joakim without whom this would not have been completed. You are my support in everything from practicalities to take care of me my mind, body and soul. I love you forever!

9 REFERENCES

1. Adeloye, D., et al., *Global and regional estimates of COPD prevalence: Systematic review and meta-analysis.* J Glob Health, 2015. **5**(2): p. 020415.

2. Mortality, G.B.D. and C. Causes of Death, *Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013.* Lancet, 2015. **385**(9963): p. 117-71.

3. Adeloye, D., et al., *Global, regional, and national prevalence of, and risk factors for, chronic obstructive pulmonary disease (COPD) in 2019: a systematic review and modelling analysis.* Lancet Respir Med, 2022. **10**(5): p. 447-458.

4. Reitsma, M.B., et al., *Spatial, temporal, and demographic patterns in prevalence of smoking tobacco use and attributable disease burden in 204 countries and territories, 1990–2019: a systematic analysis from the Global Burden of Disease Study 2019.* The Lancet, 2021. **397**(10292): p. 2337-2360.

5. Asher, M.I., et al., *Trends in worldwide asthma prevalence.* Eur Respir J, 2020. **56**(6).

6. Wenzel, S.E., *Asthma phenotypes: the evolution from clinical to molecular approaches.* Nat Med, 2012. **18**(5): p. 716-25.

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

8. Macleod, M.R., et al., *Biomedical research: increasing value, reducing waste.* Lancet, 2014. **383**(9912): p. 101-4.

9. Fanelli, D., *"Positive" results increase down the Hierarchy of the Sciences.* PLoS One, 2010. **5**(4): p. e10068.

10. Prinz, F., T. Schlange, and K. Asadullah, *Believe it or not: how much can we rely on published data on potential drug targets?* Nat Rev Drug Discov, 2011. **10**(9): p. 712.

11. Ioannidis, J.P., *Microarrays and molecular research: noise discovery?* Lancet, 2005. **365**(9458): p. 454-5.

12. Chalmers, I. and P. Glasziou, *Avoidable waste in the production and reporting of research evidence.* Lancet, 2009. **374**(9683): p. 86-9.

13. Ioannidis, J.P., *Why most published research findings are false.* PLoS Med, 2005. **2**(8): p. e124.

14. Bakker, M., et al., *Recommendations in pre-registrations and internal review board proposals promote formal power analyses but do not increase sample size.* PLoS One, 2020. **15**(7): p. e0236079.

15. Wheelock, A.M. and C.E. Wheelock, *Trials and tribulations of 'omics data analysis: assessing quality of SIMCA-based multivariate models using examples from pulmonary medicine.* Mol Biosyst, 2013. **9**(11): p. 2589-96.

16. Ståhle, L. and S. Wold, *Partial least squares analysis with cross validation for the two class problem: A Monte Carlo study.* Journal of Chemometrics, 1987. **1**.

17. Pinto, R.C., J. Trygg, and J. Gottfries, *Advantages of orthogonal inspection in chemometrics.* Journal of Chemometrics, 2012. **26**(6): p. 231-235.

18. Bylesjo, M., et al., *OPLS discriminant analysis: Combining the strengths of PLSDA and SIMCA classification.* J Chemometr, 2006. **20(8-10)**.

19. Kirwan, G.M., et al., *Building multivariate systems biology models.* Anal Chem, 2012. **84**(16): p. 7064-71.

20. Reinke, S.N., et al., *OnPLS-Based Multi-Block Data Integration: A Multivariate Approach to Interrogating Biological Interactions in Asthma.* Anal Chem, 2018. **90**(22): p. 13400-13408.

21. Cloarec, O., et al., *Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic 1H NMR data sets.* Anal Chem, 2005. **77**(5): p. 1282-9.

22. Brereton, R.G. and G.R. Lloyd, *Partial least squares discriminant analysis: taking the magic away.* Journal of Chemometrics, 2014. **28**(4): p. 213-225.

23. Eriksson, L., et al., *Orthogonal PLS (OPLS) Modeling for Improved Analysis and Interpretation in Drug Design.* Mol Inform, 2012. **31**(6-7): p. 414-9.

24. Eriksson, L., et al., *Multi- and Megavariate Data Analysis : Part I: Basic Principles and Applications*. 2006: Umetrics Inc. 425.

25. Bylesjö, M., et al., *K-OPLS package: Kernel-based orthogonal projections to latent structures for prediction and interpretation in feature space.* BMC Bioinformatics, 2008. **9**(1): p. 106.

26. Considine, E.C., et al., *Critical review of reporting of the data analysis step in metabolomics.* Metabolomics, 2017. **14**(1): p. 7.

27. Triba, M.N., et al., *PLS/OPLS models in metabolomics: the impact of permutation of dataset rows on the K-fold cross-validation quality parameters.* Mol Biosyst, 2015. **11**(1): p. 13-9.

28. Ruiz-Perez, D., et al., *So you think you can PLS-DA?* BMC Bioinformatics, 2020. **21**(Suppl 1): p. 2.

29. Bevilacqua, M. and R. Bro, *Can We Trust Score Plots?* Metabolites, 2020. **10**(7).

30. D'Ascenzo, N., et al., *Metabolomics of blood reveals age-dependent pathways in Parkinson's Disease.* Cell Biosci, 2022. **12**(1): p. 102.

31. Delgado-Dolset, M.I., et al., *Understanding uncontrolled severe allergic asthma by integration of omic and clinical data.* Allergy, 2022. **77**(6): p. 1772-1785.

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

33. Indahl, U.G. and T. Naes, *Evaluation of alternative spectral feature extraction methods of textural images for multivariate modeling.* Journal of Chemometrics, 1998. **12**(4): p. 261-278.

34. Pitman, E.J.G., *Significance Tests Which May be Applied to Samples From any Populations.* Supplement to the Journal of the Royal Statistical Society, 1937. **4**(1): p. 119-130.

35. Szymanska, E., et al., *Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies.* Metabolomics, 2012. **8**(Suppl 1): p. 3-16.

36. *WHO 2012, accessed 28 May 2018.* http://www.who.int/pmnch/media/news/2012/201204_borntoosoon-report.pdf.

37. Galindo-Prieto, B., L. Eriksson, and J. Trygg, *Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS).* Journal of Chemometrics, 2014. **28**.

38. Yang, M., et al., *Long-term smoking alters abundance of over half of the proteome in bronchoalveolar lavage cell in smokers with normal spirometry, with effects on molecular pathways associated with COPD.* Respir Res, 2018. **19**(1): p. 40.

39. Wheelock, C.E., et al., *Systems biology approaches and pathway tools for investigating cardiovascular disease.* Mol Biosyst, 2009. **5**(6): p. 588-602.

40. Worley, B. and R. Powers, *PCA as a practical indicator of OPLS-DA model reliability.* Curr Metabolomics, 2016. **4**(2): p. 97-103.

41. Heinze, G., C. Wallisch, and D. Dunkler, *Variable selection – A review and recommendations for the practicing statistician.* Biometrical Journal, 2018. **60**(3): p. 431-449.

42. Hastie, T., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2009.

43. Ambroise, C. and G.J. McLachlan, *Selection bias in gene extraction on the basis of microarray gene-expression data.* Proc Natl Acad Sci U S A, 2002. **99**(10): p. 6562-6.

44. Christin, C., et al., *A critical assessment of feature selection methods for biomarker discovery in clinical proteomics.* Mol Cell Proteomics, 2013. **12**(1): p. 263-76.

45. Lindgren, F., et al., *Model validation by permutation tests: Applications to variable selection.* Journal of Chemometrics, 1996. **10**(5-6): p. 521-532.

46. Shi, L., et al., *Variable selection and validation in multivariate modelling.* Bioinformatics, 2019. **35**(6): p. 972-980.

47. Gibson, G.J., *Respiratory Medicine*. 2003: Saunders.

48. Tam, A., et al., *The airway epithelium: more than just a structural barrier.* Ther Adv Respir Dis, 2011. **5**(4): p. 255-73.

49. Yang, J., et al., *The development and plasticity of alveolar type 1 cells.* Development, 2016. **143**(1): p. 54-65.

50. Mason, R.J., *Biology of alveolar type II cells.* Respirology, 2006. **11 Suppl**: p. S12-5.

51. Mills, C.D., *M1 and M2 Macrophages: Oracles of Health and Disease.* Crit Rev Immunol, 2012. **32**(6): p. 463-88.

52. Evren, E., E. Ringqvist, and T. Willinger, *Origin and ontogeny of lung macrophages: from mice to humans.* Immunology, 2020. **160**(2): p. 126-138.

53. Azizi, M.H., T. Nayernouri, and F. Azizi, *A brief history of the discovery of the circulation of blood in the human body.* Arch Iran Med, 2008. **11**(3): p. 345-50.

54. Joshi, S. and S. Kotecha, *Lung growth and development.* Early Hum Dev, 2007. **83**(12): p. 789-94.

55. Fujiwara, T., et al., *Artificial surfactant therapy in hyaline-membrane disease.* Lancet, 1980. **1**(8159): p. 55-9.

56. Blencowe, H., et al., *National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications.* Lancet, 2012. **379**(9832): p. 2162-72.

57. Jackson, R.A., et al., *Perinatal outcomes in singletons following in vitro fertilization: a meta-analysis.* Obstet Gynecol, 2004. **103**(3): p. 551-63.

58. Goldenberg, R.L., et al., *Epidemiology and causes of preterm birth.* Lancet, 2008. **371**(9606): p. 75-84.

59. Kamath, B.D., et al., *Neonatal mortality from respiratory distress syndrome: lessons for low-resource countries.* Pediatrics, 2011. **127**(6): p. 1139-46.

60. Hjalmarson, O., *Epidemiology and classification of acute, neonatal respiratory disorders. A prospective study.* Acta Paediatr Scand, 1981. **70**(6): p. 773-83.

61. Laughon, M., et al., *Patterns of respiratory disease during the first 2 postnatal weeks in extremely premature infants.* Pediatrics, 2009. **123**(4): p. 1124-31.

62. Lange, P., et al., *Lung-Function Trajectories Leading to Chronic Obstructive Pulmonary Disease.* N Engl J Med, 2015. **373**(2): p. 111-22.

63. Avery, M.E., et al., *Is chronic lung disease in low birth weight infants preventable? A survey of eight centers.* Pediatrics, 1987. **79**(1): p. 26-30.

64. Roberts, D. and S. Dalziel, *Antenatal corticosteroids for accelerating fetal lung maturation for women at risk of preterm birth.* Cochrane Database Syst Rev, 2006(3): p. CD004454.

65. Neu, J., *Gastrointestinal development and meeting the nutritional needs of premature infants.* Am J Clin Nutr, 2007. **85**(2): p. 629S-634S.

66. Been, J.V., et al., *Chorioamnionitis alters the response to surfactant in preterm infants.* J Pediatr, 2010. **156**(1): p. 10-15 e1.

67. Shiou, S.R., et al., *Synergistic protection of combined probiotic conditioned media against neonatal necrotizing enterocolitis-like intestinal injury.* PLoS One, 2013. **8**(5): p. e65108.

68. Lohmann, P., et al., *The airway microbiome of intubated premature infants: characteristics and changes that predict the development of bronchopulmonary dysplasia.* Pediatr Res, 2014. **76**(3): p. 294-301.

69. Brostrom, E.B., et al., *Obstructive pulmonary disease in old age among individuals born preterm.* Eur J Epidemiol, 2013. **28**(1): p. 79-85.

70. Grydeland, T.B., et al., *Quantitative computed tomography: emphysema and airway wall thickness by sex, age and smoking.* Eur Respir J, 2009. **34**(4): p. 858-65.

71. Gan, W.Q., et al., *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**: p. 52.

72. Lisspers, K., et al., *Gender differences among Swedish COPD patients: results from the ARCTIC, a real-world retrospective cohort study.* NPJ Prim Care Respir Med, 2019. **29**(1): p. 45.

73. Centers for Disease, C. and Prevention, *Annual smoking-attributable mortality, years of potential life lost, and productivity losses--United States, 1997-2001.* MMWR Morb Mortal Wkly Rep, 2005. **54**(25): p. 625-8.

74. Lozano, R., et al., *Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010.* Lancet, 2012. **380**(9859): p. 2095-128.

75. Quanjer, P.H., et al., *Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations.* Eur Respir J, 2012. **40**(6): p. 1324-43.

76. Lamprecht, B., et al., *COPD in never smokers: results from the population-based burden of obstructive lung disease study.* Chest, 2011. **139**(4): p. 752-763.

77. Hagstad, S., et al., *COPD among non-smokers - report from the obstructive lung disease in Northern Sweden (OLIN) studies.* Respir Med, 2012. **106**(7): p. 980-8.

78. Vogelmeier, C.F., et al., *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): p. 557-582.

79. Galban, C.J., et al., *Computed tomography-based biomarker provides unique signature for diagnosis of COPD phenotypes and disease progression.* Nat Med, 2012. **18**(11): p. 1711-5.

80. Spiro, S.G., G.A. Silvestri, and A. Agustí, *Clinical Respiratory Medicine E-Book: Expert Consult - Online and Print*. 2012: Elsevier Health Sciences.

81. Kew, K.M., S. Dias, and C.J. Cates, *Long-acting inhaled therapy (beta-agonists, anticholinergics and steroids) for COPD: a network meta-analysis.* Cochrane Database Syst Rev, 2014(3): p. CD010844.

82. Barnes, P.J., *How corticosteroids control inflammation: Quintiles prize lecture 2005.* British Journal of Pharmacology, 2006. **148**(3): p. 245-254.

83. Barnes, P.J., *Cellular and molecular mechanisms of chronic obstructive pulmonary disease.* Clin Chest Med, 2014. **35**(1): p. 71-86.

84. Barnes, P.J., *Immunology of asthma and chronic obstructive pulmonary disease.* Nat Rev Immunol, 2008. **8**(3): p. 183-92.

85. Baraldi, E. and M. Filippone, *Chronic lung disease after premature birth.* N Engl J Med, 2007. **357**(19): p. 1946-55.

86. Brusselle, G.G., G.F. Joos, and K.R. Bracke, *Chronic Obstructive Pulmonary Disease 1 New insights into the immunology of chronic obstructive pulmonary disease.* Lancet, 2011. **378**(9795): p. 1015-1026.

87. Caramori, G., et al., *Nuclear localisation of p65 in sputum macrophages but not in sputum neutrophils during COPD exacerbations.* Thorax, 2003. **58**(4): p. 348-51.

88. Valko, M., et al., *Free radicals and antioxidants in normal physiological functions and human disease.* Int J Biochem Cell Biol, 2007. **39**(1): p. 44-84.

89. Kohler, M., et al., *Gender differences in the bronchoalveolar lavage cell proteome of patients with chronic obstructive pulmonary disease.* J Allergy Clin Immunol, 2013. **131**(3): p. 743-51.

90. Naz, S., et al., *Metabolomics analysis identifies sex-associated metabotypes of oxidative stress and the autotaxin-lysoPA axis in COPD.* Eur Respir J, 2017. **49**(6).

91. Balgoma, D., et al., *Linoleic acid-derived lipid mediators increase in a female-dominated subphenotype of COPD.* Eur Respir J, 2016. **47**(6): p. 1645-56.

92. Schatz, M. and C.A. Camargo, Jr., *The relationship of sex to asthma prevalence, health care utilization, and medications in a large managed care organization.* Ann Allergy Asthma Immunol, 2003. **91**(6): p. 553-8.

93. Yung, J.A., H. Fuseini, and D.C. Newcomb, *Hormones, sex, and asthma.* Ann Allergy Asthma Immunol, 2018. **120**(5): p. 488-494.

94. Reddel, H.K., et al., *Global Initiative for Asthma Strategy 2021: executive summary and rationale for key changes.* Eur Respir J, 2022. **59**(1).

95. Rackemann, F.M., *A working classification of asthma.* Am J Med, 1947. **3**(5): p. 601-6.

96. Pavlidis, S., et al., *"T2-high" in severe asthma related to blood eosinophil, exhaled nitric oxide and serum periostin.* Eur Respir J, 2019. **53**(1).

97. Peters, M.C., et al., *Measures of gene expression in sputum cells can identify TH2-high and TH2-low subtypes of asthma.* J Allergy Clin Immunol, 2014. **133**(2): p. 388-94.

98. Peters, M.C., et al., *Refractory airway type 2 inflammation in a large subgroup of asthmatic patients treated with inhaled corticosteroids.* J Allergy Clin Immunol, 2019. **143**(1): p. 104-113 e14.

99. Woodruff, P.G., et al., *T-helper type 2-driven inflammation defines major subphenotypes of asthma.* Am J Respir Crit Care Med, 2009. **180**(5): p. 388-95.

100. Shi, H., et al., *Infiltration of eosinophils into the asthmatic airways caused by interleukin 5.* Am J Respir Cell Mol Biol, 1997. **16**(3): p. 220-4.

101. Boonpiyathad, T., et al., *Immunologic mechanisms in asthma.* Semin Immunol, 2019. **46**: p. 101333.

102. Price, D., M. Fletcher, and T. van der Molen, *Asthma control and management in 8,000 European patients: the REcognise Asthma and LInk to Symptoms and Experience (REALISE) survey.* NPJ Prim Care Respir Med, 2014. **24**: p. 14009.

103. Dagher, R., et al., *Novel mechanisms of action contributing to benralizumab's potent anti-eosinophilic activity.* Eur Respir J, 2022. **59**(3).

104. Kuo, C.S., et al., *T-helper cell type 2 (Th2) and non-Th2 molecular phenotypes of asthma using sputum transcriptomics in U-BIOPRED.* Eur Respir J, 2017. **49**(2).

105. Heaney, L.G., et al., *Research in progress: Medical Research Council United Kingdom Refractory Asthma Stratification Programme (RASP-UK).* Thorax, 2016. **71**(2): p. 187-9.

106. Kowal, J., M. Tkach, and C. Thery, *Biogenesis and secretion of exosomes.* Curr Opin Cell Biol, 2014. **29**: p. 116-25.

107. Caby, M.P., et al., *Exosomal-like vesicles are present in human blood plasma.* Int Immunol, 2005. **17**(7): p. 879-87.

108. Lasser, C., et al., *Human saliva, plasma and breast milk exosomes contain RNA: uptake by macrophages.* J Transl Med, 2011. **9**: p. 9.

109. Michael, A., et al., *Exosomes from human saliva as a source of microRNA biomarkers.* Oral Dis, 2010. **16**(1): p. 34-8.

110. Pisitkun, T., R.F. Shen, and M.A. Knepper, *Identification and proteomic profiling of exosomes in human urine.* Proc Natl Acad Sci U S A, 2004. **101**(36): p. 13368-73.

111. Admyre, C., et al., *Exosomes with immune modulatory features are present in human breast milk.* J Immunol, 2007. **179**(3): p. 1969-78.

112. Admyre, C., et al., *B cell-derived exosomes can present allergen peptides and activate allergen-specific T cells to proliferate and produce T(H)2-like cytokines.* Journal of Allergy and Clinical Immunology, 2007. **120**(6): p. 1418-1424.

113. Levanen, B., et al., *Altered microRNA profiles in bronchoalveolar lavage fluid exosomes in asthmatic patients.* J Allergy Clin Immunol, 2013. **131**(3): p. 894-903.

114. Admyre, C., et al., *Exosomes with major histocompatibility complex class II and co-stimulatory molecules are present in human BAL fluid.* European Respiratory Journal, 2003. **22**(4): p. 578-583.

115. Wolfers, J., et al., *Tumor-derived exosomes are a source of shared tumor rejection antigens for CTL cross-priming.* Nat Med, 2001. **7**(3): p. 297-303.

116. van Niel, G., et al., *Intestinal epithelial cells secrete exosome-like vesicles.* Gastroenterology, 2001. **121**(2): p. 337-49.

117. Raposo, G., et al., *B lymphocytes secrete antigen-presenting vesicles.* J Exp Med, 1996. **183**(3): p. 1161-72.

118. Blanchard, N., et al., *TCR activation of human T cells induces the production of exosomes bearing the TCR/CD3/zeta complex.* J Immunol, 2002. **168**(7): p. 3235-41.

119. Admyre, C., et al., *Direct exosome stimulation of peripheral human T cells detected by ELISPOT.* Eur J Immunol, 2006. **36**(7): p. 1772-81.

120. Thery, C., et al., *Molecular characterization of dendritic cell-derived exosomes. Selective accumulation of the heat shock protein hsc73.* J Cell Biol, 1999. **147**(3): p. 599-610.

121. Valadi, H., et al., *Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells.* Nat Cell Biol, 2007. **9**(6): p. 654-9.

122. Yu, X., S.L. Harris, and A.J. Levine, *The regulation of exosome secretion: a novel function of the p53 protein.* Cancer Res, 2006. **66**(9): p. 4795-801.

123. Hurley, J.H., *ESCRT complexes and the biogenesis of multivesicular bodies.* Curr Opin Cell Biol, 2008. **20**(1): p. 4-11.

124. Lopez-Verrilli, M.A. and F.A. Court, *Exosomes: mediators of communication in eukaryotes.* Biol Res, 2013. **46**(1): p. 5-11.

125. Hutagalung, A.H. and P.J. Novick, *Role of Rab GTPases in membrane traffic and cell physiology.* Physiol Rev, 2011. **91**(1): p. 119-49.

126. Weber, T., et al., *SNAREpins: minimal machinery for membrane fusion.* Cell, 1998. **92**(6): p. 759-72.

127. Qazi, K.R., et al., *Proinflammatory exosomes in bronchoalveolar lavage fluid of patients with sarcoidosis.* Thorax, 2010. **65**(11): p. 1016-24.

128. Zhang, R., et al., *Serum long non coding RNA MALAT-1 protected by exosomes is up-regulated and promotes cell proliferation and migration in non-small cell lung cancer.* Biochem Biophys Res Commun, 2017. **490**(2): p. 406-414.

129. Garcia-Contreras, M., et al., *Plasma-derived exosome characterization reveals a distinct microRNA signature in long duration Type 1 diabetes.* Sci Rep, 2017. **7**(1): p. 5998.

130. Szul, T., et al., *Toll-Like Receptor 4 Engagement Mediates Prolyl Endopeptidase Release from Airway Epithelia via Exosomes.* Am J Respir Cell Mol Biol, 2016. **54**(3): p. 359-69.

131. Karlsson, M., et al., *"Tolerosomes" are produced by intestinal epithelial cells.* Eur J Immunol, 2001. **31**(10): p. 2892-900.

132. Escudier, B., et al., *Vaccination of metastatic melanoma patients with autologous dendritic cell (DC) derived-exosomes: results of thefirst phase I clinical trial.* J Transl Med, 2005. **3**(1): p. 10.

133. Morse, M.A., et al., *A phase I study of dexosome immunotherapy in patients with advanced non-small cell lung cancer.* J Transl Med, 2005. **3**(1): p. 9.

134. Besse, B., et al., *Dendritic cell-derived exosomes as maintenance immunotherapy after first line chemotherapy in NSCLC.* Oncoimmunology, 2016. **5**(4): p. e1071008.

135. Esser, J., et al., *Exosomes from human macrophages and dendritic cells contain enzymes for leukotriene biosynthesis and promote granulocyte migration.* J Allergy Clin Immunol, 2010. **126**(5): p. 1032-40, 1040 e1-4.

136. Martinez-Bravo, M.-J., et al., *Pulmonary sarcoidosis is associated with exosomal vitamin D–binding protein and inflammatory molecules.* Journal of Allergy and Clinical Immunology, 2017. **139**(4): p. 1186-1194.

137. Kulshreshtha, A., et al., *Proinflammatory role of epithelial cell-derived exosomes in allergic airway inflammation.* J Allergy Clin Immunol, 2013. **131**(4): p. 1194-203, 1203 e1-14.

138. Valadi, H., et al., *Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells.* Nature Cell Biology, 2007. **9**(6): p. 654-U72.

139. Aliotta, J.M., et al., *Exosomes induce and reverse monocrotaline-induced pulmonary hypertension in mice*, in *Cardiovasc Res*. 2016. p. 319-30.

140. Aliotta, J.M., et al., *Induction of pulmonary hypertensive changes by extracellular vesicles from monocrotaline-treated mice.* Cardiovasc Res, 2013. **100**(3): p. 354-62.

141. Genschmer, K.R., et al., *Activated PMN Exosomes: Pathogenic Entities Causing Matrix Destruction and Disease in the Lung.* Cell, 2019. **176**(1-2): p. 113-126 e15.

142. Cech, T.R. and J.A. Steitz, *The noncoding RNA revolution-trashing old rules to forge new ones.* Cell, 2014. **157**(1): p. 77-94.

143. Winter, J., et al., *Many roads to maturity: microRNA biogenesis pathways and their regulation.* Nat Cell Biol, 2009. **11**(3): p. 228-34.

144. Goodwin, A.J., *MicroRNA Analysis in Acute Lung Injury*, in *Acute Lung Injury and Repair: Scientific Fundamentals and Methods*, L.M. Schnapp and C. Feghali-Bostwick, Editors. 2017, Springer International Publishing: Cham. p. 161-177.

145. Friedman, R.C., et al., *Most mammalian mRNAs are conserved targets of microRNAs.* Genome Res, 2009. **19**(1): p. 92-105.

146. Arroyo, J.D., et al., *Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma.* Proc Natl Acad Sci U S A, 2011. **108**(12): p. 5003-8.

147. Mitchell, P.S., et al., *Circulating microRNAs as stable blood-based markers for cancer detection.* Proc Natl Acad Sci U S A, 2008. **105**(30): p. 10513-8.

148. Shurtleff, M.J., et al., *Broad role for YBX1 in defining the small noncoding RNA composition of exosomes.* Proc Natl Acad Sci U S A, 2017. **114**(43): p. E8987-E8995.

149. Shurtleff, M.J., et al., *Y-box protein 1 is required to sort microRNAs into exosomes in cells and in a cell-free reaction.* Elife, 2016. **5**.

150. Montecalvo, A., et al., *Mechanism of transfer of functional microRNAs between mouse dendritic cells via exosomes.* Blood, 2012. **119**(3): p. 756-66.

151. Wang, K., et al., *Export of microRNAs and microRNA-protective protein by mammalian cells.* Nucleic Acids Res, 2010. **38**(20): p. 7248-59.

152. Bard, M.P., et al., *Proteomic analysis of exosomes isolated from human malignant pleural effusions.* Am J Respir Cell Mol Biol, 2004. **31**(1): p. 114-21.

153. Fujita, Y., et al., *Suppression of autophagy by extracellular vesicles promotes myofibroblast differentiation in COPD pathogenesis.* J Extracell Vesicles, 2015. **4**: p. 28388.

154. Olave, N., et al., *Regulation of alveolar septation by microRNA-489.* Am J Physiol Lung Cell Mol Physiol, 2016. **310**(5): p. L476-87.

155. Ogata, H., et al., *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Res, 1999. **27**(1): p. 29-34.

156. Vastrik, I., et al., *Reactome: a knowledge base of biologic pathways and processes.* Genome Biol, 2007. **8**(3): p. R39.

157. Ni, Y., et al., *Gene set enrichment analysis: A genome-wide expression profile-based strategy for discovering functional microRNA-disease relationships.* J Int Med Res, 2018. **46**(2): p. 596-611.

158. Backes, C., et al., *miEAA: microRNA enrichment analysis and annotation.* Nucleic Acids Res, 2016. **44**(W1): p. W110-6.

159. Wang, B., et al., *Similarity network fusion for aggregating data types on a genomic scale.* Nat Methods, 2014. **11**(3): p. 333-7.

160. Li, C.X., et al., *Integration of multi-omics datasets enables molecular classification of COPD.* Eur Respir J, 2018. **51**(5).

161. Zizzo, A.N., et al., *Similarity Network Fusion: A Novel Application to Making Clinical Diagnoses.* Rheum Dis Clin North Am, 2018. **44**(2): p. 285-293.

162. Um-Bergstrom, P., et al., *Pulmonary outcomes in adults with a history of Bronchopulmonary Dysplasia differ from patients with asthma.* Respir Res, 2019. **20**(1): p. 102.

163. Sandberg, A., et al., *Assessing recent smoking status by measuring exhaled carbon monoxide levels.* PLoS One, 2011. **6**(12): p. e28864.

164. Karimi, R., et al., *Lung density on high resolution computer tomography (HRCT) reflects degree of inflammation in smokers.* Respir Res, 2014. **15**: p. 23.

165. Schildge, J., C. Nagel, and C. Grun, *Bronchoalveolar lavage in interstitial lung diseases: does the recovery rate affect the results?* Respiration, 2007. **74**(5): p. 553-7.

166. Shaw, D.E., et al., *Clinical and inflammatory characteristics of the European U-BIOPRED adult severe asthma cohort.* Eur Respir J, 2015. **46**(5): p. 1308-21.

167. Shen, Q., et al., *Strong impact on plasma protein profiles by precentrifugation delay but not by repeated freeze-thaw cycles, as analyzed using multiplex proximity extension assays.* Clin Chem Lab Med, 2018. **56**(4): p. 582-594.

168. Merikallio, H., et al., *Smoking-associated increase in mucins 1 and 4 in human airways.* Respir Res, 2020. **21**(1): p. 239.

169. Thevenot, E.A., et al., *Analysis of the Human Adult Urinary Metabolome Variations with Age, Body Mass Index, and Gender by Implementing a Comprehensive Workflow for Univariate and OPLS Statistical Analyses.* J Proteome Res, 2015. **14**(8): p. 3322-35.

170. Wold, H., *Nonlinear Estimation by Iterative Least Squares Procedures in: David, FN (Hrsg.), Festschrift for J.* Neyman: Research Papers in Statistics, London, 1966.

171. Wiklund, S., et al., *Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models.* Anal Chem, 2008. **80**(1): p. 115-22.

172. Dweep, H. and N. Gretz, *miRWalk2.0: a comprehensive atlas of microRNA-target interactions.* Nat Methods, 2015. **12**(8): p. 697.

173. Xie, C., et al., *KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W316-22.

174. Chevillet, J.R., et al., *Quantitative and stoichiometric analysis of the microRNA content of exosomes.* Proc Natl Acad Sci U S A, 2014. **111**(41): p. 14888-93.

175. Hill, A.F., et al., *ISEV position paper: extracellular vesicle RNA analysis and bioinformatics.* J Extracell Vesicles, 2013. **2**.

176. Blank, A. and C.A. Dekker, *Ribonucleases of human serum, urine, cerebrospinal fluid, and leukocytes. Activity staining following electrophoresis in sodium dodecyl sulfate-polyacrylamide gels.* Biochemistry, 1981. **20**(8): p. 2261-7.

177. Kottel, R.H., et al., *Serum ribonuclease activity in cancer patients.* Br J Cancer, 1978. **38**(2): p. 280-6.

178. Robinson, M.D. and A. Oshlack, *A scaling normalization method for differential expression analysis of RNA-seq data.* Genome Biol, 2010. **11**(3): p. R25.

179. Dillies, M.A., et al., *A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis.* Brief Bioinform, 2013. **14**(6): p. 671-83.

180. Naz, S., et al., *Development of a Liquid Chromatography-High Resolution Mass Spectrometry Metabolomics Method with High Specificity for Metabolite Identification Using All Ion Fragmentation Acquisition.* Anal Chem, 2017. **89**(15): p. 7933-7942.

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

182. Liu, X.M., L. Ma, and R. Schekman, *Selective sorting of microRNAs into exosomes by phase-separated YBX1 condensates.* Elife, 2021. **10**.

183. Torregrosa Paredes, P., et al., *Bronchoalveolar lavage fluid exosomes contribute to cytokine and leukotriene production in allergic asthma.* Allergy, 2012. **67**(7): p. 911-9.

184. Moon, H.G., et al., *CCN1 secretion and cleavage regulate the lung epithelial cell functions after cigarette smoke.* Am J Physiol Lung Cell Mol Physiol, 2014. **307**(4): p. L326-37.

185. Kaur, G., et al., *Distinct Exosomal miRNA Profiles from BALF and Lung Tissue of COPD and IPF Patients.* Int J Mol Sci, 2021. **22**(21).

186. Gogebakan, B., et al., *The role of bronchial epithelial cell apoptosis in the pathogenesis of COPD.* Mol Biol Rep, 2014. **41**(8): p. 5321-7.

187. Di Stefano, A., et al., *Increased expression of nuclear factor-kappaB in bronchial biopsies from smokers and patients with COPD.* Eur Respir J, 2002. **20**(3): p. 556-63.

188. Webster, G.A. and N.D. Perkins, *Transcriptional cross talk between NF-kappa B and p53.* Molecular and Cellular Biology, 1999. **19**(5): p. 3485-3495.

189. Yang, M., et al., *Proteomic profiling of lung immune cells reveals dysregulation of phagocytotic pathways in female-dominated molecular COPD phenotype.* Respir Res, 2018. **19**(1): p. 39.

190. Haspel, J.A. and A.M. Choi, *Autophagy: a core cellular process with emerging links to pulmonary disease.* Am J Respir Crit Care Med, 2011. **184**(11): p. 1237-46.

191. Stiuso, P., et al., *MicroRNA-423-5p Promotes Autophagy in Cancer Cells and Is Increased in Serum From Hepatocarcinoma Patients Treated With Sorafenib.* Mol Ther Nucleic Acids, 2015. **4**: p. e233.

192. Takeda, Y., et al., *Double deficiency of tetraspanins CD9 and CD81 alters cell motility and protease production of macrophages and causes chronic obstructive pulmonary disease-like phenotype in mice.* J Biol Chem, 2008. **283**(38): p. 26089-97.

193. Takeda, Y., et al., *Preventive Role of Tetraspanin CD9 in Systemic Inflammation of Chronic Obstructive Pulmonary Disease.* Am J Respir Cell Mol Biol, 2015. **53**(6): p. 751-60.

194. Sugiura, T. and F. Berditchevski, *Function of alpha3beta1-tetraspanin protein complexes in tumor cell invasion. Evidence for the role of the complexes in production of matrix metalloproteinase 2 (MMP-2).* J Cell Biol, 1999. **146**(6): p. 1375-89.

195. Atkinson, J.J., et al., *The role of matrix metalloproteinase-9 in cigarette smoke-induced emphysema.* Am J Respir Crit Care Med, 2011. **183**(7): p. 876-84.

196. O'Shaughnessy, T.C., et al., *Inflammation in bronchial biopsies of subjects with chronic bronchitis: inverse relationship of CD8+ T lymphocytes with FEV1.* Am J Respir Crit Care Med, 1997. **155**(3): p. 852-7.

197. Mikko, M., et al., *Increased intraepithelial (CD103+) CD8+ T cells in the airways of smokers with and without chronic obstructive pulmonary disease.* Immunobiology, 2013. **218**(2): p. 225-31.

198. Roos-Engstrand, E., et al., *Influence of smoking cessation on airway T lymphocyte subsets in COPD.* COPD, 2009. **6**(2): p. 112-20.

199. Lambert, M.E., et al., *Serum carnitine levels in normal individuals.* JPEN J Parenter Enteral Nutr, 1988. **12**(2): p. 143-6.

200. Ruoppolo, M., et al., *Serum metabolomic profiles suggest influence of sex and oral contraceptive use.* Am J Transl Res, 2014. **6**(5): p. 614-24.

201. Stevens, V.L., et al., *Serum metabolomic profiles associated with postmenopausal hormone use.* Metabolomics, 2018. **14**(7): p. 97.

202. Triebner, K., et al., *Menopause Is Associated with Accelerated Lung Function Decline.* Am J Respir Crit Care Med, 2017. **195**(8): p. 1058-1065.

203. Tam, A., et al., *Sex Differences in Airway Remodeling in a Mouse Model of Chronic Obstructive Pulmonary Disease.* Am J Respir Crit Care Med, 2016. **193**(8): p. 825-34.

204. Conlon, T.M., et al., *Metabolomics screening identifies reduced L-carnitine to be associated with progressive emphysema.* Clin Sci (Lond), 2016. **130**(4): p. 273-87.

205. Zhang, Q., et al., *Serum proteomics reveals systemic dysregulation of innate immunity in type 1 diabetes.* J Exp Med, 2013. **210**(1): p. 191-203.

206. Jones, S.E. and C. Jomary, *Clusterin.* Int J Biochem Cell Biol, 2002. **34**(5): p. 427-31.

207. Isenberg, J.S., et al., *Thrombospondin-1 stimulates platelet aggregation by blocking the antithrombotic activity of nitric oxide/cGMP signaling.* Blood, 2008. **111**(2): p. 613-23.

208. Cadigan, R.J., et al., *Neglected ethical issues in biobank management: Results from a U.S. study.* Life Sci Soc Policy, 2013. **9**(1): p. 1.

209. Cadigan, R.J., et al., *Underutilization of specimens in biobanks: an ethical as well as a practical concern?* Genet Med, 2014. **16**(10): p. 738-40.

210. Wheelock, A.M., et al., *The EuPA Biobank Initiative: Meeting the future challenges of biobanking in proteomics & systems medicine.* J Proteomics, 2015. **127**(Pt B): p. 414-6.

211. Collaco, J.M., A.D. Aherrera, and S.A. McGrath-Morrow, *The influence of gender on respiratory outcomes in children with bronchopulmonary dysplasia during the first 3 years of life.* Pediatr Pulmonol, 2017. **52**(2): p. 217-224.