The Bronchoalveolar Lavage Proteome: Phenotypic associations to smoking and divergence towards development of COPD

Plymoth, Amelie

2006

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain.
• You may freely distribute the URL identifying the publication in the public portal.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
The Bronchoalveolar Lavage Proteome

Phenotypic associations to smoking and divergence towards development of COPD

Av

Amelie Plymoth

Fakultetsopponent:
Doc. Carol L. Nilsson
National High Magnetic Field Laboratory
Tallahassee, Florida, USA
Title and subtitle

The Bronchoalveolar Lavage Proteome- Phenotypic associations to smoking and divergence towards development of COPD

Abstract

Chronic obstructive pulmonary disease (COPD) is currently the world’s fourth leading cause of death and its prevalence is increasing. The leading cause of COPD is smoking and an estimated 600 million people in the world suffer from COPD which makes it the world’s most common chronic disease.

The overall aim of this thesis was to explore and characterize the bronchoalveolar lavage (BAL) proteome of never-smokers and smokers. The basic hypotheses were that the BAL proteome reflect a subjects smoking habits and that the proteome of smokers susceptible to COPD development is specific. We have chosen to analyze BAL because of its ability to address secreted and extracellular proteins present within the central and descending airways of the studied subjects. In order to relate the measurement of sets of proteins with phenotypes of clinical presentation we have developed and utilized an interdisciplinary toolbox that includes protein separation (two-dimensional gel electrophoresis and liquid chromatography), mass spectrometry identification platforms and statistical methods for multivariate analysis.

The study material used in this thesis consisted of age matched men all born in 1933, living in one city differing by lifelong smoking history, that were compared by clinical function measurements and histological assessment at the same relative time points. The follow up study after 6-7 years identified a group of subjects who had progressed to chronic obstructive pulmonary disease (COPD) GOLD stage 2. These eventual COPD patients commonly shared a distinct protein expression profile at the baseline BAL sample that could be identified using multivariate analysis. This pattern was not observed in BAL samples of asymptomatic smokers free of COPD at the 6-7 year follow-up.

In summary, the results suggested that certain patterns of protein expression occurring in the airways of long term smokers may be detected in smokers susceptible to a progression of COPD disease, and before disease is evidenced clinically.

Key words

COPD, proteomics, mass spectrometry, MALDI, ESI,

Supplementary bibliographical information

Language English

ISSN and key title 1652-8220

ISBN 91-85559-70-9

Number of pages 186

Recipient´s notes

Security classification

Distribution by (name and address)

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

Signature Amelie Plymoth Date 2006-11-01
Amelie Plymoth

The Bronchoalveolar Lavage Proteome

*Phenotypic associations to smoking and divergence towards development of COPD*

Respiratory Medicine and Allergology
Department of Clinical Sciences, Lund
Lund University, Sweden, 2006
To Anders, Margareta, Björn, Ellica and Louis
Abstract

Proteomic analysis of bronchoalveolar lavage (BAL) fluid from smokers at risk of developing chronic obstructive pulmonary disease (COPD) and never smokers is described. COPD is currently the world's fourth leading cause of death and its prevalence is increasing. The leading cause of COPD is smoking and an estimated 600 million people in the world suffer from COPD which makes it the world's most common chronic disease. The aim of this thesis was to explore and characterize the BAL proteome of never smokers and smokers. The hypotheses were that the BAL proteome reflect smoking habits in subjects, and that smokers susceptible to COPD development have a specific proteome. In order to relate the measurement of protein expression with clinical phenotypes we have developed and utilized an interdisciplinary toolbox that includes protein separation (two-dimensional gel electrophoresis and liquid chromatography), mass spectrometry identification and statistical methods for multivariate analysis. The study material used in this thesis consisted of age matched men all born in 1933, living in one city differing by lifelong smoking history. These were compared by clinical function measurements and histological assessment at the same relative time points. A follow up study after 6-7 years identified a group of subjects who had progressed to COPD GOLD stage 2. Those with COPD shared a distinct protein expression profile in the baseline BAL sample which could be identified using multivariate analysis. This pattern was not observed in BAL samples of asymptomatic smokers free of COPD at the 6-7 year follow-up. The results suggest that specific patterns of protein expression occur in the airways of smokers susceptible to COPD disease progression, before the disease is clinically measurable.
List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


Some data not included in the papers are added in this thesis.

---

1 Reprints were made with permission from the publishers
I: © 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
II: © 2006 American Association for Clinical Chemistry
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>v</td>
</tr>
<tr>
<td>List of Papers</td>
<td>vii</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>1</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>5</td>
</tr>
<tr>
<td>2 Background - Airway Diseases</td>
<td>7</td>
</tr>
<tr>
<td>2.1 Smoking induced disease of the lung</td>
<td>7</td>
</tr>
<tr>
<td>2.1.1 Chronic obstructive pulmonary disease</td>
<td>7</td>
</tr>
<tr>
<td>2.1.1.1 Prevalence</td>
<td>9</td>
</tr>
<tr>
<td>2.1.1.2 Costs to society</td>
<td>10</td>
</tr>
<tr>
<td>2.1.1.3 Treatment</td>
<td>11</td>
</tr>
<tr>
<td>3 Proteomics Background</td>
<td>13</td>
</tr>
<tr>
<td>3.1 Proteomic methodologies for bronchoalveolar lavage</td>
<td>14</td>
</tr>
<tr>
<td>3.1.1 Clinical sample acquisition and storage</td>
<td>14</td>
</tr>
<tr>
<td>3.1.1.1 Epithelial lining fluid and bronchoalveolar lavage</td>
<td>14</td>
</tr>
<tr>
<td>3.1.2 Sample preparation and fractionation</td>
<td>16</td>
</tr>
<tr>
<td>3.1.2.1 Separation techniques in BAL Proteomics</td>
<td>16</td>
</tr>
<tr>
<td>3.1.3 Protein identification and quantification</td>
<td>17</td>
</tr>
<tr>
<td>3.1.3.1 Mass spectrometry</td>
<td>19</td>
</tr>
<tr>
<td>3.1.4 Bioinformatics</td>
<td>19</td>
</tr>
<tr>
<td>3.2 Proteomics as a tool to study lung diseases</td>
<td>20</td>
</tr>
<tr>
<td>3.2.1 History of bronchoalveolar lavage proteomics</td>
<td>21</td>
</tr>
<tr>
<td>3.2.2 The bronchoalveolar lavage proteome</td>
<td>21</td>
</tr>
<tr>
<td>3.2.3 BAL protein alterations in lung disease</td>
<td>22</td>
</tr>
<tr>
<td>4 Aims of this thesis</td>
<td>25</td>
</tr>
<tr>
<td>5 Study population</td>
<td>27</td>
</tr>
<tr>
<td>5.1 Clinical outcome at 6-7 years follow up</td>
<td>28</td>
</tr>
<tr>
<td>5.1.1 Prevalence of COPD</td>
<td>30</td>
</tr>
<tr>
<td>5.1.1.1 Prevalence of COPD among the subjects who underwent fiberoptic bronchoscopy</td>
<td>30</td>
</tr>
<tr>
<td>5.1.1.2 Prevalence of COPD in the whole study population</td>
<td>31</td>
</tr>
<tr>
<td>6 Clinical Methods</td>
<td>33</td>
</tr>
<tr>
<td>6.1 Fibre-optic bronchoscopy</td>
<td>33</td>
</tr>
<tr>
<td>6.2 Bronchoalveolar lavage sampling</td>
<td>33</td>
</tr>
<tr>
<td>6.3 Tissue sampling and histology</td>
<td>34</td>
</tr>
</tbody>
</table>
# Proteomic Methods

## 7.1 Sample preparation of bronchoalveolar lavage

7.1.1 Albumin depletion using an anti-HSA column

7.1.2 Desalting and protein concentration techniques

7.1.2.1 Size exclusion

7.1.2.2 TCA precipitation

7.1.2.3 Precipitation with PlusOne 2-D Clean-Up kit

7.1.2.4 Ulramembrane filtration

7.1.2.5 Dialysis

7.2 2-D PAGE of the BAL proteome

7.2.1 Sample loading

7.2.2 First dimension: isoelectric focusing (IEF)

7.2.2.1 IEF with IPGphor

7.2.2.2 IEF on the Multiphor

7.2.3 Second dimension: SDS-PAGE

7.2.4 2-DE gel image analysis

7.3 MALDI-TOF and MALDI-TOF/TOF Protein Analysis

7.4 LC-MS/MS

7.4.1 Sample preparation

7.5 LC-LTQ

7.6 LTQ-FT

7.7 Quantitative Proteomics

7.8 Data Analysis and Interpretation

7.8.1 2-D gel generated data

7.8.2 LC-MS/MS generated data

## Results

8.1 Methodology development for analysis of the BAL proteome

8.1.1 Albumin depletion

8.1.2 Desalting and protein concentration techniques

8.1.3 First dimension: isoelectric focusing (IEF)

8.1.3.1 Sample loading for optimal resolution

8.2 BAL proteome profiling in never smokers and smokers

8.2.1 2-D gels

8.2.1.1 A model for BAL protein differential expression comparison

8.2.1.2 Statistical distributions and separations of SSP profiles of subjects

8.2.1.3 Clinical outcome at 6-7 years follow up

8.3 HPLC-MS/MS BAL proteome analysis

8.3.0.4 BAL protein differential expression comparison by mass spectrometry
8.4 Validation of LC-MS/MS with individual samples and 2-D gels 63
9 Discussion .......................................................... 67
10 Future perspectives ............................................. 73
Summary In Swedish ............................................... 77
Acknowledgements ............................................... 79
References ......................................................... 81
Abbreviations

**2-D PAGE**  Two-Dimensional Polyacrylamide Gel Electrophoresis

**2-DE**  Two-dimensional gel electrophoresis

**ACN**  Acetonitrile

**BAL**  Bronchoalveolar Lavage

**BL**  Bronchial Lavage

**BTS**  British Thoracic Society

**COPD**  Chronic Obstructive Pulmonary Disease

**CTG**  Connective Tissue Growth Factor

**DIGE**  Two Dimensional Difference Gel Electrophoresis

**DL_{CO}**  Diffusion capacity of the lung for carbon monoxide transfer

**DTT**  DiThioThreitol

**ELF**  Epithelial Lining Fluid

**EOF**  Electroendosmotic Flow

**ESI**  Electrospray Ionization

**EST**  Expressed Sequence Tag

**FEV_{1}**  Forced Expiratory Volume in one second

**FT-ICR**  Fourier Transform Ion Cyclotron Resonance

**FVC**  Forced Vital Capacity

**GOLD**  Global initiative for chronic Obstructive pulmonary Lung Disease

**HPLC**  High Performance Liquid Chromatography

**HRCT**  High Resolution Computed Tomography
Abbreviations

HSA  Human Serum Albumin
IEF  Isoelectric Focusing
IPG  Immobilized pH Gradient
LC   Liquid Chromatography
LC-MS Liquid Chromatography Mass Spectrometry
LTQ  Linear ion Trap Quadrupole
MALDI Matrix Assisted Laser Desorption Ionization
Mr   Molecular weight
MS   Mass Spectrometry
m/z  mass-to-charge ratio
PBS  Phosphate Buffered Saline
PCA  Principal Components Analysis
pI   isoelectric point
PLS-DA Partial Least Squares Discriminant Analysis
RSD  Relative Standard Deviation
SDS-PAGE Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis
SEK  SwEdish Krona (Currency)
SLPI  Secretory Leukocyte Protease Inhibitor
SP-A  Surfactant Protein A
SSP  Standard Spot Number
TCA  Trichlorocetic Acid
TIMP Tissue Inhibitor of Metalloproteinase
TLC  Total Lung Capacity
TOF  Time of flight
UMP-CMP Uridine MonoPhoshate-Cytidine MonoPhosphate kinase
VC   Vital Capacity
VEGF  Vascular Endothelial cell Growth Factor

VEGFR-2  Vascular Endothelial Growth Factor Receptor 2 precursor
1. Introduction

Chronic obstructive pulmonary disease (COPD) is currently the world’s fourth leading cause of death and its prevalence is increasing. The leading cause of COPD is smoking and an estimated 600 million people in the world suffer from COPD which makes it the world’s most common chronic disease. Very little information is available today regarding the effects of smoking on the proteome of the respiratory tract and only a few studies have directly compared global patterns of protein expression in the respiratory tract of smokers and non-smokers. The overall aim of this thesis is to describe and discuss the exploration and characterization of the bronchoalveolar lavage (BAL) proteome of never smokers and smokers. The hypotheses were that the BAL proteome reflect smoking habits in subjects, and that smokers susceptible to COPD development have a specific proteome. We have chosen to analyze BAL because of its ability to address secreted and extracellular proteins present within the central and descending airways. In order to relate the measurement of protein expression with clinical phenotypes we have developed and utilized an interdisciplinary toolbox that includes protein separation (two-dimensional gel electrophoresis and liquid chromatography), mass spectrometry identification and statistical methods for multivariate analysis. The study material used in this thesis consisted of age matched men all born in 1933, living in one city but differing by lifelong smoking history. These were compared by clinical function measurements and histological assessment at the same relative time points. A follow up study after 6-7 years identified a group of subjects who had developed COPD GOLD stage 2. The background chapters provide an overview of the covered fields of airway disease and proteomic research. The subsequent sections describe the principles behind the instruments and techniques used for the generation of all experimental data, discussion of the results obtained in the papers and their place in a global context.
2. Background - Airway Diseases

2.1 Smoking induced disease of the lung

Smoking is a world wide problem; currently there are an estimated 1.3 billion smokers in the world. According to the world health organization, five million deaths per year are induced by tobacco consumption, and tobacco-induced diseases are one of four prioritized non-infectious illnesses in the world [1].

Chronic cigarette smoking is a major cause of lethal diseases such as cancers, cardiovascular diseases, pulmonary hypertension, chronic obstructive pulmonary disease (COPD) and pneumonia, and harms nearly every organ of the body. Cigarette smoke contains more than 4700 substances, many of which have a toxic effect upon cells [2]. Within the respiratory tract, in addition to the xenobiotic effects associated with oncogene expression and malignant transformation, smoking drives a chronic inflammatory state that results in the release of mediators such as proteases that cause the destruction of ground matrix and tissue components. The adverse effects of smoking on the lung function in women may be even greater than in men. One possible explanation is that women have smaller lungs and hence the exposure to harmful agents is proportionally larger [3, 4].

Recent studies on human airway epithelial cells obtained by bronchoscopy from smokers and control subjects, have shown that the pattern of gene expression is permanently altered in smokers [5]. These alterations occur in both the qualitative and quantitative levels of important pathways that mediate electron transport and the response to oxidative stress. An additional finding of this study was that a certain subset of the smokers showed a differential response in the gene expression to smoke, possibly indicating a higher susceptibility to disease from smoking.

2.1.1 Chronic obstructive pulmonary disease

Chronic obstructive pulmonary disease (COPD) is a major worldwide health problem [6]. It is a slowly progressive disease characterized by a non fully reversible limitation in the lung airflow [7]. Although COPD affects the lungs, it also has significant extra-pulmonary effects. In the last two years there have been an increasing number of publications emphasizing the systemic nature of COPD [8, 9, 10, 11, 12, 13, 14, 15, 16] and the frequent and important
chronic comorbidities [8, 17, 18, 19, 20] that may contribute significantly to its severity and mortality [8, 21, 22, 23].

Acknowledged systemic effects of COPD are weight loss, nutritional abnormalities, and skeletal muscle dysfunction. Other less well known but potentially important systemic effects includes an increased risk of cardiovascular disease and several neurological and skeletal defects. The mechanisms underlying these systemic effects are unclear, but they are probably interrelated and multi factorial, including systemic inflammation, tissue hypoxia and oxidative stress among others. These systemic effects add to the respiratory morbidity produced by the underlying pulmonary disease and should be considered in the clinical assessment as well as the treatment of affected patients [9]. These different aspects contribute to a poor quality of life, and to added morbidity and mortality [24].

Figure 2.1: Biopsies of peripheral lung tissue from patients with moderate COPD. Bronchioles in the peripheral airways lack cartilage (A) and are in COPD patients often inflamed (bronchiolitis) and sometimes also plugged by mucus and cell debris. The lung parenchyma is responsible for the respiratory gas exchange, which benefits from the large area of the alveoli. In emphysema the alveolar structure is destroyed due to tissue degradation leaving large sacs with a higher volume/area ratio (B). (Figures kindly provided by C-G Löfdahl, lung drawing is a free picture from Oxford illustrated science encyclopedia)

Patients are often not aware of COPD symptoms in spite of a deteriorated lung function as measured e.g. by forced expiratory volume in one second ($FEV_1$). After a period free of symptoms, often lasting many years, symptoms start emerging; chronic cough, mucus hypersecretion and breathlessness associated with emerging airflow obstruction. These symptoms are the results of pathological changes in various structural and functional domains in the lungs (Figure 2.1). Within the lungs there are three main manifestations of the disease; bronchitis with increased production of phlegm in the central airways, severe bronchiolitis in airways with a diameter < 2 mm, and progressive development of emphysema [25, 26].

Another important feature of the disease is the occurrence of exacerbations that are often due to infections in the lower airways [27, 28]. These exacer-
COPD is a complex inflammatory disease that involves many different types of inflammatory and structural cells, all of which have the capacity to release multiple inflammatory mediators [30, 31]. The study of pathogenic mechanisms requires a thorough understanding of functional roles of different cells and proteins involved in normal pulmonary homeostasis as well as specific disease processes. Figure 2.2, describe some possible inflammatory mechanisms involved in the development of COPD. The cigarette smoke causes a chronic inflammation. Furthermore the inflammation is caused by bacteria and virus present within the respiratory mucosa. The toxic cigarette smoke triggers an inflammatory reaction, recruiting preferably macrophages to the airways. The macrophages release a number of neutrophil chemotactic factors, including IL-8, LTB4 and IL-1β. This way, neutrophil granulocytes are recruited to the lungs. The neutrophil granulocytes are also capable of releasing IL-8, which further increases the neutrophil adhesion effect. Activated neutrophil granulocytes release proteases, including neutrophil elastase that break down connective tissue in the lung parenchyma, resulting in emphysema, and also stimulate mucus hypersecretion. These enzymes are normally counteracted by protease inhibitors, including α1-antitrypsin, SLPI, and TIMP. Oxidative stress may also impair the function of antiproteases such as α1-antitrypsin and SLPI and thereby accelerates the breakdown of elastin in lung parenchyma. The epithelial cells are probably affected early on in the inflammatory process in parallel with the macrophages. Besides epithelial cell release of neutrophil chemotactic factors, including IL-8, they also stimulate mucus hypersecretion. Cytotoxic T cells (CD8+) may be recruited and involved in alveolar wall destruction. Fibroblasts may be activated by growth factors releases from macrophages and epithelial cells (CTG, connective tissue growth factor) [32].

COPD is usually a result of cumulative exposure to tobacco smoke, occupational dusts or vapors, as well as indoor or outdoor air pollution. The disease is most common among middle-aged to elderly people. The main risk factors of COPD are increasing age in combination with smoking [33]. The diagnosis is confirmed by spirometry, which measures ventilatory capacity in the lung. The severity grading of COPD in guidelines from example the Global Initiative for Chronic Obstructive Lung Disease (GOLD) and the British Thoracic Society (BTS), is based primarily on the degree of lung function impairment as measured by spirometry [6, 7, 34].

2.1.1.1 Prevalence

Chronic obstructive pulmonary disease is currently the fourth leading cause of death in the world and its prevalence is increasing [35, 36]. An estimated
600 million people in the world suffer from COPD making it the world’s most common chronic disease.

COPD is today the only major public health disease with an increasing mortality. In 2002 COPD was responsible for 2.75 million or 4.8% of all deaths (1.41 and 1.34 million deaths in males and females respectively). By 2020, the mortality is expected to increase to 4-5 million, or 7% of all deaths. COPD is then expected to have risen to third leading cause of death, after only ischemic heart disease and cerebrovascular disease [37]. The number of COPD patients in Sweden today is estimated to be about 500 000. This number is increasing because of a changing smoking behavior 30-50 years ago, combined with the current shape of the age distribution [38].

A number of population studies performed in northern Europe have shown that the prevalence of COPD is low in persons below the age of 50 [39, 40, 41, 42]. From around the age of 50, the prevalence increases considerably, particularly among smokers. Sooner or later 50% or more of smokers develop COPD [41].

### 2.1.1.2 Costs to society

Few studies have used health economic data from representative general population samples to estimate the burden of COPD. In 1991, the total cost to the Swedish society for asthma and COPD was estimated to be SEK 6 billion
A study run between 1994-2002, from the Karolinska Institute presents calculations of the costs to society of COPD based on a representative selection of the general population in Sweden [29, 44]. They present the annual costs for COPD in Sweden to be approximately SEK 9.1 billion, which corresponds to SEK 13 418 per subject with COPD (2002). The total cost was split 42:58 between direct and indirect costs. Direct costs refer to medical care in the form of prevention, diagnostics, treatment, and rehabilitation, etc. Indirect costs consist of the loss of productivity, which has an effect on society due to days off from work, early retirement, and death caused by the disease.

COPD is a largely under diagnosed disease [45, 46]. The large proportion of unidentified cases complicates the estimation of the burden of the disease. Approximately 40% of the total cost is estimated to be accounted for by undiagnosed subjects [44].

2.1.1.3 Treatment

No medications can cure COPD or reverse the loss of lung function caused by smoking. However the symptoms can be improved and the damage to the lungs delayed. The objective of treatments for COPD include slowing the accelerated decline in lung function; relieving symptoms, such as shortness of breath and cough; improving daily lung function; reducing the risk of exacerbations; all which can improve quality of life.

The most important step in treating COPD is early detection and active intervention to stop smoking. This is very important as the damage caused to the lungs by smoking is non-reversible and progressively worsens if smoking continues. Current available medications that are helpful in treating COPD include bronchodilators such as beta-2-agonists and anti-cholinergics, anti-inflammatory agents like corticosteroids and a combination of these. Antibiotics are also used in treating severe exacerbations, most commonly caused by bacterial and viral infections [47, 48]. Other procedures to help COPD patients with severe emphysema are lung volume reduction surgery, which increases the lung elastic recoil, and oxygen therapy. Surgery is not a cure for emphysema but can improve the quality of life especially as it decreases dyspnea. This procedure can be an alternative to or postpone lung transplantation [9, 49, 50, 51].
3. Proteomics Background

Today, Chronic obstructive pulmonary disease (COPD) is diagnosed through spirometry. However, it has become increasingly clear that FEV$_1$ and its change does not fully represent the complex clinical manifestations of COPD [52, 53]. Now the time has come to move from a disease expressed by the degree of airflow limitation, to a broader and more useful characterization of COPD [52]. The presence of inflammation and systemic consequences are important components of the disease that ought to be measured and characterized when describing patients with COPD.

We are witnessing a scientific period where important changes are being made to the way we study diseases. We are beginning to understand different patterns of disease through genomics, proteomics, and metabolomics [54, 55, 56]. The availability of novel and powerful technologies derived from proteomics and functional genomics are adding new dimensions to the analysis of clinically relevant samples [57, 58]. This promises to revolutionize the way diseases will be treated and managed in the future.

Proteomics, i.e., the study of the entire protein complement of the genome in a biological system at a given time point [58, 59], is a relatively new post-genomic science with high potential. In contrast to gene expression studies, proteomics directly addresses the level of gene products in a given cell state and can further characterize protein activities, interactions and subcellular distributions. Proteins and peptides present within clinical samples represent a wealth of information regarding the ongoing processes within cells and tissues in healthy and diseased state. Expression proteomics studies are usually employed to identify proteins that are up- or downregulated in a disease-specific manner for use as diagnostic markers or therapeutic targets [60, 61]. In other words, by studying differences between proteomes in healthy and diseased individuals, we gain information about the state of the disease which can help us develop new preventive measures.

Proteomics is a multidisciplinary approach, which has become practical through improved techniques and methods, including:

(a) advances in genomic sequencing and the growth of gene expressed sequence tags (EST) and protein sequence databases during the 1990s;

(b) introduction of user-friendly, browser-based bioinformatics and
computational methods that facilitate the analysis and interpretation of the abundant data generated by the proteomics experiments;

(c) new methods in protein separation that allow the detection of subtle changes in protein expression, including post translational modifications;

(d) development of oligonucleotide micro array;

(e) advances in mass spectrometry that now permit the identification and relative quantitation of small amounts of nearly any single protein;

(f) and automation and miniaturization that permit high-throughput analysis of clinical samples.

This multidisciplinary approach makes proteomics studies practical for pulmonary researchers from many sources including bronchoalveolar lavage (BAL) fluid, lung tissue, blood, and exhaled breath condensates.

3.1 Proteomic methodologies for bronchoalveolar lavage

A typical work-flow for a proteomics experiment includes the following: (1) sample acquisition and storage, (2) sample preparation and fractionation, (3) protein quantification and identification, and (4) bioinformatics (Figure 3.1).

3.1.1 Clinical sample acquisition and storage

Well-characterized samples from well-designed clinical studies are needed to support clinical proteomics activities [62]. Although the most commonly collected sample is blood, other body fluids may provide additional important opportunities. Bronchoalveolar lavage fluid can provide important information about lung function when studying pulmonary diseases.

For all types of samples, the investigator should be cautious when it comes to collection and the method used for long-term storage in order to prevent protein degradation and modification (e.g. proteolysis).

3.1.1.1 Epithelial lining fluid and bronchoalveolar lavage

The airways and, particularly, the alveoli are covered with a thin layer of epithelial lining fluid (ELF), which is a rich source of many different cells and soluble components of the lung that play important functions by protecting the lung from undue aggressions and preserving its gas-exchange capacity.
Figure 3.1: Typical BAL proteomic work flow, where the intact proteins are separated using 2-D gels and peptides of digested protein mixtures using LC.

[63]. Bronchoalveolar lavage performed during fiber-optic bronchoscopy is the most common way to get samples of ELF [64].

Bronchoalveolar lavage fluid thus collected from healthy individuals contains different compounds: cells, lipids, nucleic acids and proteins/peptides. The cellular content of BAL mainly consists of alveolar macrophages (80-95% of the cells), lymphocytes (<10%), neutrophils (<5%), eosinophils (<5%) and sometimes plasma cells. Squamous epithelial cells, bronchial epithelial cells, type II pneumocytes, basophils and mast cells are also found in BAL [65, 66]. Phospholipids, responsible for the decrease of alveolar surface tension, are synthesized by the lung epithelial cells and represent the main components of the surfactant (90%) [67]. Nucleic acids are found occasionally in BAL during infections by pathogens [68].

The protein composition of the ELF reflects the effects of the external factors that contact the lung and is of primary importance in the early diagnosis, assessment and characterization of lung disorders as well as in the search for disease markers [67]. Numerous chemical, physical and biological exposures and, finally, lung diseases induce biochemical modifications of this biological fluid [64].
3.1.2 Sample preparation and fractionation

A complex biological sample such as bronchoalveolar lavage has a wide dynamic range of protein concentrations. High abundant proteins in BAL are albumin (50%), immunoglobulins A and G (together about 30%), transferrin (5 to 6%) and α1-antitrypsin (3 to 5%) [69]. Other factors that complicate the analysis of the BAL proteome are the highly variable dilution factor, the low protein content and its high salt concentration that comes from the phosphate-buffered saline used for the procedure.

Since no current technologies are capable of simultaneously resolving the high number of proteins across a high dynamic range present in a complex biological sample such as BAL, additional steps are needed to reduce the complexity of the sample.

3.1.2.1 Separation techniques in BAL Proteomics

The most commonly used analytical tools in classical proteome research are multidimensional protein separation by two-dimensional polyacrylamide gel electrophoresis (2-D PAGE), or by liquid chromatography (LC). In contrast to 2-D gel electrophoresis where intact proteins are separated, liquid chromatography can separate either whole proteins or peptides of digested protein mixtures. Both 2-D PAGE and LC are followed by mass spectrometry (MS) identification.

Two-dimensional gel electrophoresis

The most widely used BAL proteome profiling method has been 2-D gel electrophoresis separation. 2-D PAGE enables the separation of complex mixtures of proteins according to isoelectric point (pI) and molecular weight (Mr). In the first dimension, proteins are separated on the basis of their isoelectric point and in the second dimension; the focused proteins are further separated according to their molecular weight. Once stained, the resulting two-dimensional gels deliver a map of intact proteins, which reflects changes in protein expression level, isoforms or post-translational modifications [70].

A typical BAL 2-D gel map comprises about 1000 protein spots corresponding to about 100 different proteins. Several classes of proteins are systematically underrepresented in 2-D PAGE; this limitation is relevant for many of the potential proteins of interest in pulmonary research [71]. Important reasons for this problem are that 2-D PAGE discriminates against low abundance proteins, very small and very large proteins, alkaline proteins and hydrophobic proteins, membrane proteins and heavily glycosylated proteins. Other reasons for this problem are limitations in the amount of protein that can be loaded on a preparative 2-D gel, the dynamic range of the current staining procedures and poor recovery of the peptides from the gels. Generally, proteins that are visualized in 2-D gels by conventional staining methods are high-abundance
Proteomic methodologies for bronchoalveolar lavage

5-50 ng (corresponding to 100-1000 fmol for a 50 kDa protein, an amount visible by silver staining and fluorescence detection) is considered necessary for successful mass spectrometry identification of gel-embedded proteins [72].

**Liquid chromatography**

Due to the huge dynamic range of concentrations of BAL proteins and the inherent limitation of the 2-D gel electrophoresis-based technique, the 2-D PAGE approach has only been able to identify a limited number of relatively high abundance proteins without prior sample fractionation [73]. Analysis of complex samples requires the introduction of an efficient separation step prior to mass spectrometric detection. One of the most powerful separation methods for peptides is liquid chromatography. To provide a more comprehensive BAL proteome database, studies are emerging that use liquid chromatography coupled with tandem mass spectrometry for protein identification and semi-quantitation [74, 75, 76].

The separation of complex peptide mixtures is usually performed by a combination of several separation techniques. In the first dimension size exclusion, ion exchange, hydrophobic interaction or affinity can be used. This is followed by a last step of reversed liquid chromatography, which can easily be coupled with mass spectrometry.

### 3.1.3 Protein identification and quantification

The rapid evolution of mass spectrometry has made it a key technique for the investigation of the proteome [58, 77]. Today, mass spectrometry is one of the most powerful tools in proteomics.

**Figure 3.2:** Ionization techniques used in this thesis. In MALDI, the peptides are embedded in a crystal matrix and ionized by a laser pulse (A). In ESI, the sample enters through a flow stream (often from the HPLC) and passes through the needle held at a high voltage. The needle tip aligns with a nozzle inlet and a fine jet is formed, which breaks up into small charged droplets. Due to solvent evaporation and subsequent droplet fission caused by charge repulsion, gas-phase ions are formed and transported into the mass analyzer (B). (Adapted from [58])
In a typical protein identification workflow, a protein is first digested using a proteolytic enzyme e.g. trypsin, which cleaves reproducibly at the C-terminal end of arginine and lysine. The resulting peptides are then ionized and detected using mass spectrometry, which resolves ions based on their mass-to-charge ratio (m/z). A detector registers the number of ions at each m/z value. Two of the most commonly used ionization techniques, are matrix-assisted laser-desorption ionization (MALDI) and electrospray ionization (ESI) (Figure 3.2). The mass analyzers used in this thesis are time-of-flight (TOF), ion trap and Fourier transform ion cyclotron resonance (FT-ICR) (Figure 3.3).

Multiple strategies can be used in order to obtain quantitative data on protein expression. Relative quantitation can be obtained from 2-D gels by comparing presence, absence, and staining intensity of the individual stained protein spots. The two Dimensional difference gel electrophoresis (DIGE) technique allows up to three samples, labeled with different fluorescent dyes, to be analyzed at the same time [78].

Several methods allow the relative quantification of proteins to be made through mass spectrometry. Stable isotopes can be incorporated into the sample polypeptides through metabolic labeling [79, 80], enzymatically directed methods [81, 82, 83], and chemical labeling using externally introduced tags [84, 85, 86]. For absolute quantitation, a known amount of heavy standard (identical) peptide for each peptide to be analyzed needs to be included in

---

**Figure 3.3:** Mass spectrometers used in this thesis. (A) Reflector time-of-flight matrix-assisted laser-desorption ionization (MALDI-TOF). (B) TOF-TOF instrument with an incorporated collision cell between the two TOF sections. (C) Linear ion trap with electrospray ionization. (D) Fourier transform ion cyclotron resonance mass spectrometer (FT-ICR MS) with electrospray ionization. (Adapted from [58])
Each experiment [87, 88]

One promising alternative label-free protein quantitation method is mass spectral peak intensities of peptide ions, which correlate well with protein abundances in complex samples [89]. Another label-free method, termed spectral counting, compares the number of MS/MS spectra assigned to each protein. An advantage of spectral counting is that relative abundances of different proteins can in principle be measured [90]. One of the differential analysis softwares available at the market is DeCyderTM MS, which is a novel software program for fully automated differential expression analysis based on LC-MS/MS data.

3.1.3.1 Mass spectrometry

The coupling of liquid chromatography with mass spectrometry had a great impact on the identification and quantitation of proteins and has proven to be an important alternative method to 2-D gels. Liquid chromatography can be coupled to both ESI and MALDI.

To explore the context of global protein expression patterns in complex clinical samples, one frequently used strategy involves an approach known as “shotgun sequencing”, which was pioneered by Yates and Hunt [91, 92, 93]. The shotgun proteomic strategy based on digesting proteins into peptides and sequencing them using tandem mass spectrometry and automated database searching, has become the method of choice for identifying proteins in most large scale studies. Shotgun sequencing allows the rapid identification of hundreds of protein identities present in high to medium abundance [94, 95, 96]. This tandem mass spectrometry based techniques may have advantages over gel-based techniques in speed, sensitivity, reproducibility, and applicability to different samples and concentrations [72, 92, 96, 97, 98, 99].

3.1.4 Bioinformatics

Because of the complexity of the proteome, large amounts of data are generated from one set of proteomic experiments. Therefore, bioinformatics where high-throughput data is stored and analyzed play a key role in proteomic studies and is often the rate-limiting step [100, 101]. With improvements in genomic and protein databases, and with more powerful computational search engines, MS-based approaches are now the method of choice for most protein identification. Protein search programs used in this thesis are SEQUEST [102, 103] and Mascot [104]. These are examples of algorithms/programs that interface experimental MS spectra with the theoretical spectra derived from protein databases. Since clinical proteomics experiments may result in tens of thousands of data points, bioinformatical challenges are the data storage, the analysis and the description of the large amount of information into a compre-
hensive model. This includes the development of methods for data comparison between different research groups, and the integration of gene ontologies [105]. In the functional annotation cataloge, Gene Ontology, biological functions are described by a set of defined terms organized in hierarchical structure. Statistical tools used for data exploration are for example unsupervised and supervised multivariate analysis, hierarchical clustering and the classification tool Random Forest [106, 107, 108].

The goal of most clinical proteomics experiments is to identify biomarkers specific to disease state. However, the broad spectrum of proteomic results of hundreds to thousands of proteins makes the interpretation challenging. Only a few of these proteins are likely to be biomarkers for disease. Therefore, special attention should be paid to well-focused scientific questions in the study design.

3.2 Proteomics as a tool to study lung diseases

The large vascular and air space surface area of the lung is the primary target of ambient air pollutants that can produce various health effects, including decrement of lung function, induce bronchoconstriction as in asthma attacks and increase the risks of acquiring respiratory infections. Whether caused by chemical, physical or biological environmental compounds, damage to the lungs and airways often leads to morbidity or mortality [63, 109]. The complexity of the lung makes it a promising, but also challenging, target for proteomics.

The proteomes of the lung (tissue proteome, cell proteome, BAL proteome) are a dynamic group of specialized proteins probably related to pulmonary, cellular and integrated functions. The complexity of the lung proteomes are essentially due to the presence of multiple different types of cells and the large contact with environmental compounds, such as industrial pollutants or pathogens [109, 110]. Many lung diseases are essentially multi-factorial diseases, and thus require a global analysis of such disorders.

During the development of proteomics over the last two decades, there have been numerous attempts to apply proteomic technologies to pulmonary medicine, with the aim to define groups of proteins significantly associated with disease or clinical outcome. Results from these studies indicate that the clinical applicability will increase in the near future. In combination with genomic and transcriptomic approaches, protein profile analysis of clinical samples from well characterized patients and subjects will lead to a better understanding of the mechanisms of lung diseases, and contribute to the discovery of diagnostic tools and new therapeutics for the treatment of diseases [111].
3.2.1 History of bronchoalveolar lavage proteomics

The first two-dimensional gel database of the bronchoalveolar lavage was published by Bell in 1979 [112]. They investigated alveolar proteinosis in smokers compared to non-smokers. At that time most of the proteins found in BAL were identified as serum proteins. Later, Lenz and others confirmed the research potential of the BAL proteome. In 1990 this group published a method for 2-D PAGE of BAL fluid from dogs and later compared protein patterns in BAL fluid proteins from patients with idiopathic pulmonary fibrosis, sarcoidosis, and asbestosis with normal controls [113, 114]. The results of these early studies provided a basic understanding of the protein composition of BAL fluid, but many of the characterized spots could not be identified, which limited the values of these results for clinical medicine. Since then, gradual progress in staining and imaging techniques, improvements in protein identification by mass spectrometry and standardization have made it possible to identify the most abundant BAL proteins and refine the information on proteomic changes in different disease states.

The most important technical progress made in 2-D PAGE BAL proteomics lies in the method used for isoelectrical focusing. A major contribution to the variability observed formerly in 2-D gel patterns was the carrier ampholyte that determined the first-dimensional separation. Limitations included discontinuities in the gel gradient, drifts in the focusing pattern and difficulty to produce precast gels, thereby lowering intra and inter-laboratory reproducibility. The introduction of immobilized pH gradients (IPG) in 1983 [115] overcame these drawbacks. The pH gradient, created by covalently incorporating acidic and basic buffering groups into the polyacrylamide gel, is an integral part of the gel matrix. Additional improvements of the human 2-D BAL proteome map can be achieved by using narrow-range IPG strips covering such pI interval as 3.5-6.7. Instead of wide range pI 3-10. This enables the detection of lower abundant proteins in BAL. The developments of BAL proteome master gels have increased the clinical relevance of 2-D PAGE BAL studies. The improvement in protein spot detection has been shown to be more significant for the protein spots present exclusively in BAL than for the spots present in both BAL and serum. This finding suggests that many of the BAL fluid specific proteins, which are likely to be of pulmonary origin, are low-abundance proteins.

3.2.2 The bronchoalveolar lavage proteome

Soluble proteins are the most promising elements of epithelial lining fluid leading to accurate diagnosis of lung disorders. This is because BAL fluid samples contain a large number of different proteins, the level of which change with varying physiological conditions. Soluble proteins in BAL may origi-
nate from a broad range of sources, such as diffusion from serum across the air-blood barrier and the production by different cell types present; bronchial epithelial cells, pulmonary T cells, alveolar macrophages, alveolar TI and TII cells and Clara cells [116, 117].

Many proteins identified in BAL are also found in plasma. However, among these proteins, some are synthesized also in the lung. More interestingly, a certain number of proteins are present in a high concentration in BAL and only minute in plasma, detectable by the use of specific antibodies. This suggests that they are specifically produced in the airways. These proteins are therefore good candidates for becoming lung-specific biomarkers.

Several intracellular proteins that are specifically found in BAL and not in plasma, probably originate from cellular lysis consecutive to cellular turnover and death. This may be more pronounced in the BAL compartment then in peripheral blood. Interestingly, some of the lung-specific proteins are also found in serum, as a result of their spontaneous leakage across the air-blood barrier, and thereby provide another mean of assessing the integrity of the barrier, less invasive than by measuring the levels of these proteins in BAL [116, 117].

3.2.3 BAL protein alterations in lung disease

Many studies of BAL differential-display proteome analysis have been dedicated to different lung pathologies. Information is available on change in 2-D PAGE protein patterns of BAL for smoking [113, 118, 119, 120, 121], sarcoidosis [69, 113, 118, 119, 122, 123, 124, 125, 126], idiopathic pulmonary fibrosis [69, 113, 118, 119, 122, 123, 124, 125], lupus erythematosus [69, 123], Wegener’s granulomatosis [69, 123], hypersensitivity pneumonitis [113, 118, 119, 122, 123, 127], lipoid pneumonia [69, 123], chronic eosinophilic pneumonia [69, 123], alveolar proteinosis [112, 128], bacterial pneumonia [69], other infectious, malignancies and immunosuppression [129, 130], cystic fibrosis before and after α1-antiprotease treatment [131], asbestosis [69, 123], pulmonary transplantation [132] and acute lung injury [133]. Recently, information from BAL differential-display proteome analysis that uses liquid chromatography coupled with tandem mass spectrometry are also available for; COPD [134], and in combination with SDS-PAGE fractionation for asthma [76].

In response to various inflammatory stimuli, lung endothelial cells, alveolar and airway epithelial cells, and activated alveolar macrophages produce nitric oxide and superoxide, products that may react to form peroxynitrates. Peroxynitrate can nitrate and oxidize amino acids in various lung proteins, such as surfactant protein A (SP-A), and inhibit their function. It has been shown that the nitration and oxidation of a variety of alveolar proteins is associated with diminished function in vitro; in addition, both modifications have been
identified in proteins sampled from patients with acute lung injury using immunoassays [135, 136]. So far there have not been many studies published on the detection of post-translational modifications of BAL proteins. However, there is growing interest in identifying proteins in BAL which undergo post-translational modifications. It appears obvious that such unidentified proteins, present amongst the wide variety of proteins in BAL, could serve as useful and potentially novel biomarkers for diagnosing lung diseases. Today proteomic approaches provide a promising tool for this.
4. Aims of this thesis

The overall aim of this thesis was to explore and characterize the BAL proteome of never smokers and smokers. The hypotheses were that the BAL proteome reflect smoking habits in subjects, and that smokers susceptible to COPD development have a specific proteome.

To achieve this, the thesis has the following specific aims:

- To develop methods for the identification and quantification of proteins in bronchoalveolar lavage samples to achieve a higher level of annotation of the expression map of this proteome (paper I, II, III).
- To determine whether relative qualitative and quantitative differences in protein expression could be related to smoke exposure (paper II, III).
- To provide a comprehensive qualitative proteomic analysis of BAL fluid protein expression from never smokers and from smokers at risk of developing chronic obstructive pulmonary disease, and relate the proteome findings to development of COPD (paper II, III).
- To link candidate biomarkers to the pathophysiology of COPD (paper II, III).
5. Study population

The male 60 year old subjects studied were recruited from the randomized population study “Men born in Göteborg in 1933” [137, 138]. In 1983 a random sample was drawn of half of all the men born in Göteborg in 1933, who were still residing in Göteborg (n=1016) (Figure 5.1). In 1993 a subset of the study population of 879 subjects, who were still alive and living in Göteborg, were recruited for examination.

![Subject recruitment diagram](image)

*Figure 5.1: Subject recruitment from the WHO study “Men born in Göteborg in 1933”, a population based study of health and disease.*

Of the 602 responding men, 532 were evaluated with spirometry. Of these 112 were smokers and 198 were never smokers. 222 were former smokers, and
not further evaluated. All smokers and a random sample of 60 lifelong never
smokers were invited to an examination of their lungs. Among the subjects
originally included in the study, subjects were excluded if they had
1. Any airway disease for which they had sought medical attention (n=11).
2. Congestive heart failure or unstable angina pectoris (n=4).
3. Any other severe disease (n=5).
4. Scoliosis or other diseases with deformity of the thorax (n=0).
5. Any kind of infection during the four weeks prior to the examination (n=0).
6. Treatment with corticosteroids, N-Acetylcysteine or acetylsalicylic acid
   less than 4 weeks before examination (n=0).

Twenty subjects were excluded for above mentioned medical conditions
and four smokers had quit smoking before they came to the examination and
were therefore excluded. Further 56 subjects were not willing to take part
in the study for other reasons. The drop-out frequency was equally distributed
among smokers and never smokers. The initial study (trial I) included 92 men,
58 smokers and 34 never smokers. The subjects were examined in the period
February 1994 to July 1995: they were thus 61 or 62 years old. A subset of the
subjects participating volunteered to undergo fiberoptic bronchoscopy. This
group included 48 subjects: 30 chronic smokers (15 light and 15 heavy smok-
ers); and 18 never smokers. The definition of light smokers was ≥15 smoked
cigarettes/day with a median number of 22 pack years (range 9-45). Defini-
tion of heavy smokers was ≥15 smoked cigarettes/day with a median number
of 45 pack years (range 31-79). Pack years were calculated by multiplying
the number of packs of cigarettes smoked per day by the number of years the
person had smoked.

The unpublished data presented in this thesis also contains patients taking
part in a clinical study where BAL was taken with a similar method. They
were 5 moderate COPD patients (GOLD stage 2) and 2 mild COPD patients
(GOLD stage 1).

The studies were approved by the ethics committees of the Sahlgrenska
Hospital (Gbg M 117-01) and Lund University Hospital in Sweden (LU 689-
01), and informed consent was obtained.

5.1 Clinical outcome at 6-7 years follow up

All subjects included in trial I were asked to participate in a follow-up study
(trial II), 6-7 years later (Figure 5.1). The follow-up examinations took place
in 2000 and 2001, at which time the subjects were 67 or 68 years old. Median
follow-up duration was 6 years and 3 months (75 months), with a range of 5
years to 6 years and 11 months (60-83 months) [139].

Sixty-eight subjects came to trial II. Twenty-four subjects, 7 never smok-
### Table 5.1: Lung function measurements of the study subjects.

<table>
<thead>
<tr>
<th>COPD patient nr(^3)</th>
<th>Visit 1(^1)</th>
<th>Visit 2(^2)</th>
<th>Visit 1</th>
<th>Visit 2</th>
<th>(\Delta)^3</th>
</tr>
</thead>
<tbody>
<tr>
<td>243</td>
<td>70</td>
<td>61</td>
<td>94</td>
<td>70</td>
<td>-24</td>
</tr>
<tr>
<td>301</td>
<td>61</td>
<td>46</td>
<td>86</td>
<td>64</td>
<td>-22</td>
</tr>
<tr>
<td>326</td>
<td>69</td>
<td>62</td>
<td>84</td>
<td>64</td>
<td>-20</td>
</tr>
<tr>
<td>501</td>
<td>67</td>
<td>65</td>
<td>67</td>
<td>66</td>
<td>-1</td>
</tr>
<tr>
<td>621</td>
<td>64</td>
<td>54</td>
<td>88</td>
<td>64</td>
<td>-24</td>
</tr>
<tr>
<td>817</td>
<td>64</td>
<td>53</td>
<td>83</td>
<td>79</td>
<td>-4</td>
</tr>
<tr>
<td>855</td>
<td>58</td>
<td>54</td>
<td>73</td>
<td>62</td>
<td>-11</td>
</tr>
<tr>
<td>mean</td>
<td>65</td>
<td>56</td>
<td>82</td>
<td>67</td>
<td>-15.1/pt</td>
</tr>
</tbody>
</table>

- Asymptomatic Smokers\(^5\)

| mean                   | 76             | 73             | 98      | 94      | -3.4/pt       |

1. The initial study took place during the period February 1994 to July 1995.
3. Delta is defined as the numerical difference: \(\Delta = \text{Visit2} - \text{Visit1}\).
4. Seven of the smokers developed moderate COPD and 14 of the smokers remained asymptomatic at follow up.
ers and 17 smokers, were lost to follow-up. Reasons in the case of smokers: death (carcinoma of the lung, bladder and stomach, brain tumor, cerebral infarction, myocardial infarction, bronchopneumonia and suicide, n=7), various disorders (psychiatric disorder, lymphoma, sequelae due to coronary bypass operation, leg fracture, n=4), moved from the area (n=3), not willing to participate (n=3). Reasons in the case of never smokers: dementia (n=1), stroke (n=2), moved from the area (n=2), not willing to participate (n=2). The final material comprised 68 subjects, giving a follow-up rate of 74%.

5.1.1 Prevalence of COPD

![Figure 5.2: Prevalence of COPD among the subjects who underwent fiberoptic bronchoscopy.](image)

COPD was classified according to the criteria developed by GOLD [6, 7]. GOLD define COPD as $FEV_1/FVC < 70\%$. The severity of COPD is defined by level of $FEV_1$ in % of predicted. GOLD defines:

1. mild COPD as $FEV_1 \geq 80\%$ predicted,
2. moderate COPD as $50\% \leq FEV_1 < 80\%$ predicted,
3. severe COPD as $30\% \leq FEV_1 < 50\%$ and
4. very severe COPD as $FEV_1 < 30\%$.

5.1.1.1 Prevalence of COPD among the subjects who underwent fiberoptic bronchoscopy

Of the subjects who underwent fiberoptic bronchoscopy, a total of 3 of the 24 light and heavy smokers who participated in trial II, were found to have developed mild COPD (GOLD stage 1) (Figure 5.2). Seven of the light and heavy smokers, were found to have developed moderate COPD (GOLD stage 2). None of the never smoking subjects developed either mild or moderate COPD.
As shown in Table 5.1, the subjects who developed COPD had developed significant airway obstruction as seen in the decline in $FEV_1$ (% predicted) and $FEV_1/FVC$ scores, compared to the other smokers.

### 5.1.1.2 Prevalence of COPD in the whole study population

In the whole study population, a total of 5 of the 40 light and heavy smokers who participated in trial II, were found to have developed mild COPD (GOLD stage 1) (Figure 5.3). Fourteen of the light and heavy smokers, were found to have developed moderate COPD (GOLD stage 2). None of the never smoking subjects developed either mild or moderate COPD.

Our study suggests that almost 50% of smokers may develop COPD, not 15-20% as was proposed in the literature decades ago [140] and recently as well [141]. This higher COPD percentage is in concordance with the findings of Lundbäck et al [41].

The prevalence of COPD, among smokers, according to the guidelines of today seems considerably higher than that generally reported in current literature, and over 50% of the smokers developed COPD. The study illustrates the large dependency of the criteria of disease, the age composition and the smoking habits of the study population when measuring the prevalence of COPD.
6. Clinical Methods

6.1 Fibre-optic bronchoscopy

The bronchoscopies were performed at the Sahlgrenska hospital by physician Ann Ekberg-Jansson as described previously [137]. All bronchoscopies were made between 08.00 and 10.00. Premedication was given with diazepam 5 mg orally followed by 0.5 ml morphine-scopolamin intramuscularly. In some cases, additional diazepam (2.5-5.0 mg) was given intravenously during the bronchoscopic procedure. To avoid unexpected bronchoconstriction during the procedure, terbutalin 0.25 mg/dose 2 × 3 in a nebulizer was given. Local anesthesia was given initially with 1% tetracaine-spray in the mouth and laryngeal tract, followed by additional anesthesia applied through the bronchoscope channel for the lower respiratory tract. The bronchoscopy was performed transorally with an Olympus flexible fiberoptic bronchoscope (Loympus optical Co., Japan). The subjects were examined in a supine position. Oxygen saturation was measured with an Ohmedia Pulse Oximeter during the bronchoscopy, and supplemental oxygen was given at a rate of 2-3 L/minute through a nasal catheter when needed.

6.2 Bronchoalveolar lavage sampling

Bronchoalveolar lavage was made as a fractionated small volume lavage 10 ml +3 × 50 ml of sterile and endotoxin-free phosphate buffered saline (PBS). The first 10 ml were processed separately and were denoted bronchial lavage (BL). The rest of the lavage, denoted bronchoalveolar lavage (BAL), was pooled into a sterile siliconised bottle and transported on ice immediately to the laboratory. At the laboratory, the total volume of the lavage was measured and centrifuged at 250 × g and 4°C for 10 minutes. The supernatant was centrifuged a second time at 10000 × g and frozen at −70°C.

The protein concentrations in the recovered BAL were determined using the Coomassie plus protein assay reagent (Pierce, IL, USA) with BSA as a standard, according to Bradford [142]. The total protein concentration of the recovered BAL samples was not significantly different between subjects but the recovery of BAL was lower in the smoking cohort (Table 6.1).
6.3 Tissue sampling and histology

Peripheral bronchial biopsy specimens were sampled from each of the subject groups with an alligator forceps from the main carina between the right and left lung. The biopsy specimens were gently removed from the forceps and immediately placed in a sterile and moistened chamber and transported to the laboratory where they were frozen immediately in melting propane previously cooled in liquid nitrogen. Frozen biopsy specimens were kept in liquid nitrogen until sectioned. The samples were attached to the specimen holder of a cryostat microtome (Microm, HM 500 M, Heidelberg, Germany) in a drop of OCT compound (Tissue-Tek, Miles, Elkhart, IN, USA) and cut in sections with a thickness of 4 mm. After drying in air at room temperature, the sections were wrapped in aluminum foil and stored at $-70^\circ C$ until they were stained using conventional Hematoxylin and Eosin staining [143].

6.4 Lung function tests

All study subjects included were evaluated by several respiratory function tests including forced expiratory volume in one second ($FEV_1$), diffusion capacity of the lung for carbon monoxide transfer ($DLCO$), total lung capacity (TLC), and vital capacity (VC). The individual group characteristics from the initial study (trial I) are shown in Table 6.1. Both the never smokers and smokers showed ventilatory function results which fell within the normal range. However, overall smokers showed significantly lower lung function values than the never smokers. The study cohort was at follow up (trial II) evaluated by spirometry, high-resolution computed tomography (HRCT), and clinical examination. Table 5.1 shows the individual subject characteristics of the 7 smokers who developed moderate COPD (trial II). Table 5.1 also shows the individual group characteristics of the asymptomatic smokers from trial II.
<table>
<thead>
<tr>
<th></th>
<th>Smokers (n=30)</th>
<th>Never-smokers (n=18)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lung function</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TLC %</td>
<td>97 ± 14</td>
<td>96 ± 13</td>
<td>n.s.</td>
</tr>
<tr>
<td>RV %</td>
<td>116 ± 32</td>
<td>94 ± 27</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>FEV₁ %</td>
<td>92 ± 14</td>
<td>107 ± 16</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>DLCO %</td>
<td>84 ± 16</td>
<td>95 ± 16</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td><strong>Protein conc. BAL</strong></td>
<td>77 ± 40</td>
<td>86 ± 38</td>
<td>n.s.</td>
</tr>
<tr>
<td>(µg/mL)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Recovery of BAL</strong></td>
<td>64 (30-110)</td>
<td>89 (50-120)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>(mL)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Smoker Group</th>
<th>Pack years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Light (n=15)</td>
<td>22</td>
</tr>
<tr>
<td>Heavy (n=15)</td>
<td>45</td>
</tr>
</tbody>
</table>

Table 6.1: Characteristics of the Study Subjects. Data are presented as mean ± standard deviation and mean and range values, P-value according to Mann-Whitney’s U-test.
7. Proteomic Methods

7.1 Sample preparation of bronchoalveolar lavage

In order to study a complex biological sample such as bronchoalveolar lavage (BAL) the sample preparation is crucial. After the BAL has been taken it is centrifuged to separate cells from the protein and further analysis performed on the protein fraction. Factors that complicate analysis of the BAL proteome are a highly variable dilution factor, a low protein content, a wide dynamic range of proteins and its high salt concentration that comes from the saline used for the BAL procedure. In this thesis, various protein concentration and desalting techniques, and albumin depletion were studied to address these complications, with the intent to sustain high yields and high protein recoveries.

7.1.1 Albumin depletion using an anti-HSA column

The highest abundant protein in BAL is albumin and in order to reveal lower abundant proteins on the two-dimensional gels, albumin was depleted. This was done by using an anti-human serum albumin column (anti-HSA column). High importance was placed on the goal that the albumin removal should maintain the integrity of the BAL constituents and allow high definition 2-D gel separation.

The anti-HSA column is an immunoaffinity based chemistry product comprised of a goat polyclonal antibody generated against HSA immobilized on a POROS perfusion chromatography support (Applied Biosystems, Framingham, MA, USA). Before BAL fluid was run through the anti-HSA column one blank run was performed in which no sample was injected. The equilibration followed by the elution cycle should remove any weakly bound antibody from the cartridge and increase the run-to-run reproducible. The BAL sample was then injected and the albumin free BAL sample was collected. After loading the BAL sample a wash step where PBS was passed through in order to remove any unbound protein from the cartridge. During the wash step fractions were collected and the protein containing fractions (measured by UV absorbance) pooled. The cartridge was reconditioned before reuse. The protein concentrations in the recovered eluate were determined with the Coomassie
plus protein assay reagent (Pierce) protein assay with BSA as standard.

7.1.2 Desalting and protein concentration techniques

The high salt concentration that comes from the saline used for the BAL procedure is incompatible with the 2-D gel analysis. To remove salt and concentrate the protein, we investigated the following techniques: size exclusion, TCA precipitation, precipitation with PlusOne 2-D Clean-Up kit, ultramembrane filtration and dialysis.

7.1.2.1 Size exclusion

BAL fluid was desalted by gel filtration through a PD-10 column (Amersham Biosciences, Uppsala, Sweden) and eluted with a 12 mM ammonium hydrogen carbonate buffer, pH 7.0. The BAL sample was then freeze-dried and resuspended in the sample solution and run on the first dimension.

7.1.2.2 TCA precipitation

BAL fluid proteins were precipitated with TCA (10% final concentration) in an ice bath for 20 min, and subsequently centrifuged at 15 300 rpm for 15 min at 4°C. The pellet was suspended in ice-cold acetone using a sonicator and centrifuged at 15 300 rpm for 15 min at 4°C. The pellet was air-dried and resuspended in the sample solution and run on the first dimension.

7.1.2.3 Precipitation with PlusOne 2-D Clean-Up kit

BAL fluid proteins were precipitated with PlusOne 2-D Clean-Up kit (Amersham Biosciences). The PlusOne 2-D Clean-Up kit procedure uses precipitant and coprecipitant in combination to precipitate proteins. Precipitated proteins are pelleted by centrifugation and the precipitate is further washed to remove non-protein contaminants. After a second centrifugation, the resultant pellet is resuspended in the sample solution and run on the first dimension.

7.1.2.4 Ultramembrane filtration

BAL fluid proteins were desalted using Ultrafree®-4 Centrifugal Filter Unit, cut-off 5 kDa (Millipore, Bedford, MA, USA). The BAL sample was then resolved in the sample solution and run on the first dimension.

7.1.2.5 Dialysis

BAL fluid proteins were desalted with a dialysis cassette (Slide-A-Lyzer®3.5K; Pierce) in a 12 mM ammonium hydrogen carbonate buffer, pH 7.0 for 24 h. The BAL sample was then volume reduced by SpeedVac and resuspended in the sample solution and run on the first dimension.
7.2 2-D PAGE of the BAL proteome

The most widely used BAL proteome profiling method has been two-dimensional polyacrylamide gel electrophoresis (2-D PAGE). 2-D PAGE enables the separation of complex mixtures of proteins according to isoelectric point (pl) and molecular weight (Mr). Once stained, the resulting two-dimensional gels deliver a map of intact proteins, which reflects changes in protein expression level, isoforms or post-translational modifications.

7.2.1 Sample loading

In order to obtain good protein resolution on 2-D gels, optimal protein loading levels are essential. Different BAL protein starting concentrations were loaded on the IPG strips to compare the spot resolution. 50 µg, 100 µg and ≥ 150 µg of BAL fluid protein starting concentration were loaded on IPG strips (0.5 × 3 × 240 mm), which separate in pH range between 3-10, 4-7 and 4.5-5.5. The protein samples were mixed with a standard (7 M urea, 2 M thiourea, 4% CHAPS, 2.5% DTT, 2% IPG buffer) or modified (7 M urea, 2 M thiourea, 4% CHAPS, 2.5% DTT, 10% isopropanol, 5% glycerol, 2% IPG buffer) sample loading buffer.

7.2.2 First dimension: isoelectric focusing (IEF)

In the first dimension, proteins are separated on the basis of their isoelectric point. Electroendosmotic flow (EOF) of water in the IPG strip during isoelectric focusing (IEF) and migration of DTT in alkaline pH gradients is a problem during protein separation. To overcome these problems we rehydrated the IPG strips overnight in the modified lysis buffer. And the BAL fluid proteins were also mixed with the modified sample loading buffer. Glycerol and isopropanol in the modified lysis/rehydration buffer was aimed to minimise electroendosmotic flow of water in the IPG strips with concomitant gel glueing which hinders protein transfer from the first dimension. The problem with migration of dithiothreitol (DTT) under alkaline conditions was alleviated in the Multiphor by adding “CleanGel” paper wick immersed in 3.5% DTT, at the cathodic side.

Using a sample prepared by ultramembrane filtration, we compared first dimensional isoelectric focusing (IEF) performed with an IPGphor 2-DE setup (Amersham Biosciences) and a Multiphor 2-DE setup (Pharmacia, Biotech, Uppsala, Sweden).

7.2.2.1 IEF with IPGphor

BAL fluid proteins, mixed with sample loading buffer were loaded directly on the IPG strips and focusing was performed overnight (75 kVh).
7.2.2.2 IEF on the Multiphor

BAL fluid proteins, mixed with the modified sample loading buffer were rehydrated over night with the IPG strips. The focusing was performed overnight (50kVh). At the cathodic side a “CleanGel” paper wick immersed in 3.5% DTT, was added. During IEF, electrode paper strips were replaced by fresh ones three times.

7.2.3 Second dimension: SDS-PAGE

In the second dimension, proteins are separated according to their molecular weight. Second dimension sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) was carried out by transferring the IPG strips to a 12% SDS gel (0.5*180*245 mm). The gels were run at 120-150 mA for about 17 h. Staining was made by post-gel fluorescence using Sypro Ruby (Molecular Probes, USA).

7.2.4 2-DE gel image analysis

Protein patterns in the gels were analyzed as digitalized images using a high resolution scanner. Quantitative analysis was performed by utilizing the image software PDQuest™ ver. 6.0 (BioRad, Richmond, CA, USA). Individual spots were assigned unique standard spot numbers (SSPs) and the amount of protein in a spot was assessed as integrated optical density (IOD). In order to normalize for differences in total staining intensity between different 2-DE images, the amount of different spots were expressed as the percentage of the individual spot IOD per total IOD of all the spots (% IOD).

7.3 MALDI-TOF and MALDI-TOF/TOF Protein Analysis

Protein spots excised from the 2-D gels were washed with ammonium bicarbonate buffer, reduced and alkylated followed by tryptic digestion over night. Micro-affinity capillary extraction was used for sample enrichment and improved protein identification. Three cm capillaries were used with 4-6 mm packed bed sizes using POROS-50 beads. Sample deposition was performed by spotting onto stainless steel MALDI-target plates using α-cyano-4-hydroxycinnamic acid as matrix. Mass spectrometry analysis was performed by MALDI-TOF using a Voyager DE-PRO MALDI time-of-flight mass spectrometer (Applied Biosystems, Framingham, MA, USA) equipped with a 337 nm nitrogen laser with built-in delayed extraction and by MALDI-TOF/TOF using an ABI 4700 (Applied Biosystems). The
7.4 LC-MS/MS

The 2-dimensional gel electrophoresis system can identify proteins present in medium to high abundance in BAL samples. To provide a more comprehensive BAL proteome database, one-dimensional reversed phase liquid chromatography coupled with tandem mass spectrometry was used to analyze the protein composition in BAL fluid.

7.4.1 Sample preparation

Bronchoalveolar lavage proteins were taken from each subject. As an internal standard 9.5 pmol fetuin (Sigma-Aldrich, Saint Louis, Missouri) per 100 μg, BAL protein was added to each sample. 6 M GuCl in 0.1 M ammonium carbonate (pH 8) was then added to the samples to a final concentration of 5 mol/L followed by the addition of DTT to a final concentration of 5 mM. The mixtures were incubated at 75°C for 30 min then cooled to ambient temperature. After incubation, iodoacetamide solution was added to a final concentration of 20 mM, and the samples were incubated for 1.5 hours in darkness. After the second incubation, the samples were solvent exchanged using 0.1 M ammonium bicarbonate buffer (pH 8) with a 10kDa Amicon filter (0.5 mL capacity; Millipore, MA, USA), and adjusted to a final volume of 114 μL using 0.1 M ammonium carbonate. Then trypsin (Promega, Madison, WI) was added to a protein/trypsin ratio of 100:1. The samples were incubated at 37°C for 12 hours. For complete digestion, trypsin was added a second time followed by incubation at 37°C for an additional 8 hours. Before injection on the LC MS/MS system, the digested samples were acidified to pH 4 with 10% formic acid. As a standard, 5 pmol of angiotensin I and II was added to each sample.

7.5 LC-LTQ

BAL digests were separated on a C18 nano-capillary column (Magic C18; 150 × 0.075 mm (i.d.); packed in house) on an Ettan multidimensional liquid chromatography (LC) system. The flow rate was maintained at 400 nL/min. Buffer A was water containing 1 mL/L formic acid. Buffer B was ACN containing 1 mL/L formic acid. The gradient was started at 2% buffer B, increased to 35% buffer B in 60 min, then increased to 60% buffer B in 15 min, and finally, increased to 90% buffer B in 10 min. Of each sample, 2 μg protein
in 20 μl was injected on the column by the auto sampler. The resolved peptides were detected on an LTQ mass spectrometer (ThermoElectron) with a nanoelectrospray ionization ion source. To provide consistency, as proposed by Washburn et al. [144], each pooled sample was analyzed in triplicate.

### 7.6 LTQ-FT

BAL digests were analyzed by LC MS/MS using a LTQ FT mass spectrometer (Thermo, San Jose, CA, USA) equipped with a TriVersa NanoMate (Advion BioSciences, Ithaca, NY). Peptides were captured and concentrated for on-line RP-HPLC using a C18 capillary column (2.1 mm i.d. × 50 cm), with 1.9 μm packed particles (Higgins Analytical Inc., Mountain View, CA). Buffer A was water containing 1 mL/L formic acid. Buffer B was ACN containing 1 mL/L formic acid. The gradient was started at 100% buffer A, increased to 50% buffer B in 60 min, then increased to 90% buffer B in 10 min, and finally, increased to 100% buffer B in 4 min. 20 μl of each sample containing 20 μg of protein was injected. The flow was split 1:1000 (NanoMate LTQ FT to sample plate) using a MicroTee (Upchurch Scientific, Oak Harbor; WA). The low-flow branch of the split was directed into the mass spectrometer and full scan (MS) data was collected in the FT-ICR MS in profile mode between m/z 400-2000. Simultaneous data dependent ion selection was monitored to select the seven most abundant ions from an MS scan for MS/MS analysis on the ion trap mass spectrometer. Dynamic exclusion was continued for a duration of 2 min. The temperature of the ion transfer tube was controlled at 185°C and the collision energy was set at 35% for MS/MS. The high-flow branch was collected into a 96-well plate using a time-based fractionation of 36 sec/fraction for offline investigative work. To provide consistent data, each sample was analyzed in duplicate.

### 7.7 Quantitative Proteomics

Multiple strategies can be used in order to obtain quantitative data on protein expression. Relative quantitation was obtained from 2-D gels by comparing presence, absence, and optical intensity scores of the individual stained protein spots.

For label-free protein quantitation, to estimate relative protein abundances from the LC-LTQ study, we considered the number of peptides leading to identification and the semiquantitative Sequest score parameter in conjunction with peak-area measurements. Peak-area measurements were performed on abundant peptides. We extracted the peak area of the m/z signal of a selected
Data Analysis and Interpretation

For protein differential analysis of the LC-LTQ and LC-LTQ-FT generated data, DeCyder™ MS, which is a new software program, was also used. DeCyder MS is a tool for visualization, detection, comparison, and label-free relative quantitation of LC-MS/MS data [145]. The software displays LC-MS analysis as two-dimensional intensity maps. Detection, profile comparison, background subtraction, and quantitation were done on the full scan mode. The PepDetect module of DeCyder MS was used for automated peptide detection, charge-state assignments, deconvolution, and quantitation based on MS signal intensities of individual LC-MS analysis. The PepMatch module was used to match peptides in different intensity maps from the different runs, which resulted in a quantitative comparison including statistical evaluation. The DeCyder MS algorithm is based on accurate mass and reproducible retention times. Peptides were identified using intact masses by exporting the MS/MS files into TurboSEQUEST protein identification software and subsequently importing TurboSEQUEST search results into the PepMatch module. The peptide matches were filtered based on cross-correlation scores of 1.9, 2.5 and 3.75 for charge states 1+, 2+, and 3+, respectively.

7.8 Data Analysis and Interpretation

7.8.1 2-D gel generated data

PCA and PLS-DA multivariate analysis was performed using the Simca-P software v. 10.0.4 (Umetrics AB, Umeå, Sweden). Principal Components Analysis (PCA) was used for unsupervised views of data. Partial least squares discriminant analysis (PLS-DA) was used to build classifiers of clinical category [106, 108]. Both methods provide one measure, R², of model fit, and another, Q², of the predictability of the model. In the hierarchical clustering method, data were normalized by subtracting the mean intensity and dividing by the standard deviation separately for each SSP.

7.8.2 LC-MS/MS generated data

For annotation of BAL peptides and proteins we applied as closely as possible proposed publication guidelines [146, 147]. The peptide sequences generated by LTQ and LTQ-FT MS were identified by correlation with
the peptide sequences present in the nonredundant National Center for Biotechnology Information protein database (TaxonomyID=9606, available at www.ncbi.nlm.nih.gov) [148], which contains Swiss-Prot protein entries, using the Sequest algorithm, BioWorks™ (Ver. 3.1 SR and 3.2; Thermoelectron). The following settings were used: trypsin was used as enzyme, a maximum of one missed cleavages was allowed, precursor-ion mass accuracy 1.4 amu, peptide mass accuracy 2.0 amu, fragment-ion mass accuracy 1.0, and carbamidomethylcysteine modification was allowed. The Sequest results were further filtered by Xcorr versus charge state. The Xcorr function measures the similarity between the mass-to-charge ratios (m/z) for the fragment ions predicted from amino acid sequences obtained through the database, and the fragment ions observed in the MS/MS spectrum. Xcorr was used with a match of 1.5/1.9 for singly charged ions, 2.0/2.5 for doubly charged ions, and 2.5/3.75 for triply charged ions [149, 150]. The peptide identifications generated from triplicate analysis of a sample were all combined in the final protein identity. Proteins identified with two or more unique peptides or the same peptide in two or more samples were considered as positive identifications, while proteins identified with only a single peptide sequence were further investigated by inspecting the MS/MS spectrum manually. Protein IDs from positive single peptides were typically those with high signal to noise ratio and with correct multiple peptide fragment assignments.
Despite the fact that COPD is a common disease, our understandings of the underlying biological mechanisms remain limited. Early and specific detection of differentially expressed proteins in pre-COPD and COPD patients is of outstanding interest to enable accurate diagnosis, follow-up, prognosis and treatment. This thesis therefore studies the expression of proteins present in baseline bronchoalveolar lavage (BAL) samples, taken from asymptomatic sixty-year-old lifelong current smokers or healthy never smokers. These subjects were later re-evaluated to determine clinical outcome in terms of COPD and compared to a COPD diagnosed study population. This comparison could determine whether relative qualitative and quantitative differences in protein expression could be related to smoke exposure (paper I-III).

The key component for pursuing clinical proteomics studies is the careful selection of a well-characterized clinical material that can be normalized for study, both in terms of being associated with the biology of interest, but also for sampling variability. The study material used in this thesis consists of age matched men all born in 1933, living in one city, differing by lifelong smoking history, and compared by clinical function measurements and histological assessment at the same relative time points. BAL fluid was analyzed because of its ability to address secreted and extracellular proteins present within the central and descending airways.

In order to study a complex biological sample such as bronchoalveolar lavage, sample preparation is crucial. After bronchoscopy the BAL sample was centrifuged to separate cells from protein, and further analysis was performed on the protein fraction. Factors that complicate the analysis of the BAL proteome are the highly variable dilution factor, the low protein content, the wide dynamic range of proteins and its high salt concentration that comes from the saline used for the BAL procedure. To address these factors, this thesis studied various BAL sample preparation methods including protein concentration and desalting techniques, and albumin depletion (see section “Proteomic Methods”).
8.1 Methodology development for analysis of the BAL proteome

8.1.1 Albumin depletion

The highest abundant protein in BAL is albumin and in order to reveal less abundant proteins on the two-dimensional gels, albumin was depleted. This was done by using an anti-human serum albumin column. High importance was placed on the goal that the albumin removal should maintain the integrity of the BAL constituents and allow high definition 2-D gel separation.

As shown in Figure 8.1, the albumin depletion changed the pattern of BAL proteins resolved on the 2-D gels. Areas which were previously masked by human serum albumin (HSA) showed additional protein spots. In further experiments, we were able to determine statistically that this was due to the co precipitation and trapping of these proteins within the first dimension matrix rather than that these proteins were hidden behind the high abundant albumin isomers (data not shown).

![Figure 8.1: 2-D gels showing albumin depletion compared to no albumin depletion. (A) BAL fluids containing 100 μg of BAL protein not run through an anti-HSA column; (B) BAL fluids containing 100 μg of BAL protein run through an anti-HSA column; and (C) eluate from anti-HSA column, through which BAL fluid containing 150 μg of protein was run. (Adapted from paper I)](image)

HSA bound by the anti-HSA column was washed more stringently eluting proteins bound to the albumin matrix. These results showed that specific proteins in BAL were bound to albumin in complexes and specific proteins adsorbed nonspecifically to the anti-HSA column. By comparisons of processed and unprocessed BAL, it was confirmed that the majority of HSA bound proteins were present in the eluate from the anti-HSA assay and were detectable by both 1-D, and 2-D gel separation. In conclusion, the benefit in spot detection must be compared with the drawback of the albumin depletion technique.
8.1.2 Desalting and protein concentration techniques

The high salt concentration that comes from the saline used for the BAL procedure is incompatible with the 2-D gel analysis. In Figure 8.2 the isoelectric focusing did not succeed because of the salt content leading to a poor focusing effect and protein precipitation, with only 164 spots distinguished in this gel. The following desalting techniques were investigated in this thesis: size exclusion, TCA precipitation, precipitation with PlusOne 2-D Clean-Up kit, ultramembrane filtration and dialysis. Desalting by size exclusion, TCA precipitation and precipitation with PlusOne 2-D Clean-Up kit resulted in improved first dimension focusing and hence improved 2-D gel expression. Both high as well as low Mr regional protein spots could be detected. However, protein losses were up to 40% and resolution in the pI region above about 6.5 was not optimal (Figure 8.3). We conclude that these desalting techniques as such are difficult to reproduce quantitatively in samples of varying protein content.

![Figure 8.2: 2-D gel showing BAL fluid pooled from the smoking groups containing 2-D gel showing 100 μg of BAL fluid protein that had not been desalted. The first dimension was performed on a horizontal Multiphor 2-DE setup. (Adapted from paper I)](image)

Desalting by ultramembrane filtration was an effective technique for salt removal compared to desalting by size exclusion, TCA precipitation and precipitation with PlusOne 2-D Clean-Up kit. There was a 50% improvement in
content and resolution in these particular 2-D gels. A typical gel image is depicted in Figure 8.1a with a high number of protein spots throughout the Mr, as well as the pI range. Additionally, this is an easy and fast sample processing technique. Although protein loss occurs with this method, protein adsorption onto the membrane surface can easily be circumvented by repeated washing steps with sample solution. However, approximately 30% of the starting protein content was lost during the salt removal.

![Figure 8.3: 2-D gels showing various sample preparation techniques to address selective salt elimination in the BAL samples. The first dimension was performed on a horizontal Multiphor 2-DE setup. BAL fluid, containing 100 μg of BAL protein was desalted by (A) precipitation with TCA; (B) ultramembrane filtration; (C) dialysis, interfaced to SpeedVac. (Adapted from paper I)](image)

Desalting by dialysis was found to be equivalent to ultramembrane filtration with regards to protein yield, but it often results in sample dilution which requires a further concentration step. The process is also more time-consuming than ultramembrane filtration. Further BAL sample concentration through SpeedVac was found to be an easier method and generated less protein loss than freeze-drying.

In conclusion, we found that desalting and sample concentration by ultramembrane filtration was an effective technique for salt removal which achieved the highest level of annotation of the expression map of this proteome. And this method was used to run 2-D gels on all the 48 subjects in the “Men born in Göteborg in 1933” study.

8.1.3 First dimension: isoelectric focusing (IEF)

Electroendosmotic flow (EOF) of water in the IPG strip during IEF and migration of DTT in alkaline pH gradients is a great problem during protein separation. As reported by Görg et al. and Hoving et al. [151, 152] these problems were solved using a modified lysis buffer (7 M urea, 2 M thiourea, 4% CHAPS, 2.5% DTT, 10% isopropanol, 5% glycerol, 2% IPG buffer 3-10 and 4-7) in which the IPG strips were rehydrated overnight. We also came to the same conclusions. Using the modified sample buffer resulted in better separa-
tion of basic proteins with no smearing effects found (paper I).

The same sample prepared by the ultramembrane filtration technique, was compared using the Multiphor 2-DE and IPGphor 2-D first dimension IEF systems. The Multiphor gave the best results using a modified lysis/rehydration buffer with an increased number of proteins found on the gels, especially above pH 7. The glycerol and isopropanol in the modified lysis/rehydration buffer minimized the electroendosmotic flow (EOF) of water in the IPG strips with concomitant gel glueing and hindered protein transfer from the first dimension. The problem with migration of dithiothreitol (DTT) under alkaline conditions was alleviated in the Multiphor by adding “CleanGel” paper wick soaked in 3.5% DTT, at the cathodic side.

8.1.3.1 Sample loading for optimal resolution

In order to obtain good protein resolution on 2-D gels, optimal loading levels are essential. The differences seen in spot resolution caused by varying the protein concentration of the sample was measured. 50 μg, 100 μg and ≥ 150μg of BAL fluid protein starting concentration were loaded on the IPG strips (0.5 × 3 × 240nm), with pI 3-10, 4-7 and 4.5-5.5. By loading 50 μg on the IPG strips with pI 3-10, high density was reached with good resolution expressed throughout the Mr, as well as in the pl range. This was compared with loading of 100 μg and 150 μg. As protein load increased, the resolution decreased for high abundance protein but increased for low abundance protein. High abundance proteins showed a 30% loss in resolution between gels loaded with 50 or 100 μg. A further decrease in the number of spots occurred loading 150 μg onto the gel. The same comparison was made using narrow pl 4-7 strips which allow approximately double fold loading capacity within this pH range. A high degree of resolution throughout the Mr, as well as the pl range was reached by loading 100 μg. This is to be compared with levels of 50 μg and 150 μg, where fewer proteins were identified. Having achieved these results, we choose to load 100 μg of BAL starting concentration on pl 4-7 IPG strips, for 2-D gel analysis of all the 48 subjects in the “Men born in Göteborg in 1933” study. When higher BAL protein concentrations ≤ 1.5 mg were loaded on the zoom IPG strip, pl 4.5-5.5 less proteins were identified, especially among the high abundance proteins.

8.2 BAL proteome profiling in never smokers and smokers

8.2.1 2-D gels

Loading of 100 μg protein on IPG strips pl 4-7 following ultrafiltration, was used to run 2-D gels on all the 48 subjects from the “Men born in Göteborg in
Results

1933” study.

Representative 2-D gels from the never smokers and the chronic smokers are shown in Figure 8.4. Reference gels for the never smoking group, the light smoking group and the heavy smoking group (Figure 8.4 C, F, I) were synthesized by hierarchic clustering of images from the individual member gels of each group. All gels from all individuals were combined to create a synthetic master gel of all protein spots detected (Figure 8.4 J). Together, the master 2-D gel of all the subjects contained 944 standard spot numbers (SSPs), each of which was observed in one or more of the samples.

In group comparisons, the distributions of BAL proteins observed on the 2-D master gel from never smokers (Figure 8.4 A, B, C) varied qualitatively and quantitatively from the protein patterns observed for light smokers (Figure 8.4 D, E, F), and heavy smokers (Figure 8.4 G, H, I). Both the never smoking group, and the smoking groups, contained proteins in their BAL samples which were not frequently observed or existent in the other respective group gels. For example, the zoomed image of a representative gel region (Figure 8.4 J) displaying medium to low abundance protein spot expression levels, reveal a high degree of annotation differences between smokers (Figure 8.4 K)

Figure 8.4: Representative 2-D gels of the bronchoalveolar lavage proteome from never smokers and smokers. The distribution of BAL proteins on the raw gels from never smokers (A, B) varies qualitatively and quantitatively from the protein patterns for light smokers (D, E) and heavy smokers (G, H). Reference gels for each group (C never smokers, F light smokers, I heavy smokers) are synthesized by hierarchic clustering of images from the individual member gels of each group. (J) All individual gels are combined to create a synthetic master gel of all protein spots detected. The zoomed images of the gel region (J) reveals a high degree of annotation differences between smokers (K, M) and never smokers (L, N). (Adapted from paper III)
and never smokers (Figure 8.4 L). Software rendering of pixel density, into 3-D image profiles provides improved visualization of the relative presence-absence (Figure 8.4 M, N).

In order to associate SSP attributes between the never smoking and smoking groups, the global protein set was filtered to exclude proteins which were not frequently expressed by either group. Each subject gel image was scored for each SSP of the global set of 944 protein spots in terms of presence or absence, and for aggregate pixel density. When a stringency cut off demanding at least 30% presence of a given SSP in any single group was used, a total number of 818 SSP identifications remained for further analysis. Mean scores of SSP presence/absence and aggregate density for the never smoking group and the smoking group was calculated. These means were further used to compare the relative differences in expression levels between the groups by dividing the mean SSP density of the smoking group, by the SSP density of the never smoking group.

Figure 8.5: Constructed model to compare differences between the smoking and never smoking bronchoalveolar lavage proteome expression patterns. Protein spot count (Y-axis), the relative fold differences in individual SSP expression levels between smokers and never smokers (X-axis), and the individual SSP distribution at various points of presence between 30% (n= 818 SSPs) and ≥ 90% presence (n= 189 SSPs), (Z-axis). (Insert) 85 % of the SSP identities show similar expression patterns in both never smokers and smokers, 2 % of the SSP identities are down regulated in smokers (density ratio < -4) and 13 % are highly induced in the smokers (density ratio > 4). (Adapted from paper III)
8.2.1.1 A model for BAL protein differential expression comparison

A model was constructed to compare differences between the smoking and never smoking protein SSP expression patterns (Figure 8.5). Several questions were posed to the model:

<table>
<thead>
<tr>
<th>SSP</th>
<th>Protein</th>
<th>Regulation Factor</th>
<th>Accession number</th>
</tr>
</thead>
<tbody>
<tr>
<td>4806</td>
<td>Secretory IgA</td>
<td>-5.2</td>
<td>P01833</td>
</tr>
<tr>
<td>6601</td>
<td>Serum albumin fragment</td>
<td>-2.2</td>
<td>Q16167</td>
</tr>
<tr>
<td>504</td>
<td>Pulmonary surfactant-associated protein A1</td>
<td>-2.1</td>
<td>Q8IWL2</td>
</tr>
<tr>
<td>2802</td>
<td>Alpha-1B-glycoprotein</td>
<td>-2.1</td>
<td>P04217</td>
</tr>
<tr>
<td>3702</td>
<td>Ig alpha-1 chain C region</td>
<td>-2.1</td>
<td>P01876</td>
</tr>
<tr>
<td>2205</td>
<td>Chloride intracellular channel protein 1</td>
<td>2.0</td>
<td>Q00299</td>
</tr>
<tr>
<td>3004</td>
<td>Coactosin-like protein</td>
<td>2.0</td>
<td>Q14019</td>
</tr>
<tr>
<td>6119</td>
<td>Serum albumin fragment</td>
<td>2.2</td>
<td>Q16167</td>
</tr>
<tr>
<td>1002</td>
<td>Calcyphosine</td>
<td>2.3</td>
<td>Q13938</td>
</tr>
<tr>
<td>5305</td>
<td>Annexin III</td>
<td>2.4</td>
<td>P12429</td>
</tr>
<tr>
<td>1204</td>
<td>14-3-3 protein beta/alpha</td>
<td>2.6</td>
<td>P31946</td>
</tr>
<tr>
<td>2008</td>
<td>Rho GDP-dissociation inhibitor 2</td>
<td>3.0</td>
<td>P52566</td>
</tr>
<tr>
<td>6603</td>
<td>Serum albumin fragment</td>
<td>3.2</td>
<td>Q16167</td>
</tr>
<tr>
<td>4003</td>
<td>Ferritin light chain</td>
<td>3.3</td>
<td>P02792</td>
</tr>
<tr>
<td>8405</td>
<td>Fructose-1,6-bisphosphatase 1</td>
<td>3.4</td>
<td>P09467</td>
</tr>
<tr>
<td>2215</td>
<td>Annexin V</td>
<td>3.5</td>
<td>P08758</td>
</tr>
<tr>
<td>7710</td>
<td>Transferrin</td>
<td>3.6</td>
<td>P02787</td>
</tr>
<tr>
<td>3006</td>
<td>Serum albumin fragment</td>
<td>4.0</td>
<td>Q16167</td>
</tr>
<tr>
<td>2502</td>
<td>Actin, cytoplasmic 1</td>
<td>4.1</td>
<td>P60709</td>
</tr>
<tr>
<td>4507</td>
<td>Serum albumin fragment</td>
<td>4.6</td>
<td>Q16167</td>
</tr>
<tr>
<td>8601</td>
<td>Serum albumin fragment</td>
<td>4.8</td>
<td>P02787</td>
</tr>
<tr>
<td>4008</td>
<td>Albumin-thrombin-alpha-MSH</td>
<td>4.9</td>
<td>P02788</td>
</tr>
<tr>
<td>3107</td>
<td>Glutathione S-Transferase</td>
<td>5.5</td>
<td>P09211</td>
</tr>
<tr>
<td>5506</td>
<td>Serum albumin fragment</td>
<td>5.8</td>
<td>Q16167</td>
</tr>
<tr>
<td>3212</td>
<td>Glutathione S-Transferase</td>
<td>6.3</td>
<td>P09211</td>
</tr>
<tr>
<td>6508</td>
<td>Serum albumin fragment</td>
<td>7.4</td>
<td>Q16167</td>
</tr>
<tr>
<td>2206</td>
<td>Rho GDP-dissociation inhibitor 2</td>
<td>8.5</td>
<td>P52566</td>
</tr>
<tr>
<td>3203</td>
<td>Serum albumin fragment</td>
<td>9.2</td>
<td>Q16167</td>
</tr>
<tr>
<td>5204</td>
<td>Cathepsin D precursor</td>
<td>9.3</td>
<td>P07339</td>
</tr>
<tr>
<td>4110</td>
<td>Glutathione S-Transferase</td>
<td>20.0</td>
<td>P09211</td>
</tr>
</tbody>
</table>

1 Standard spot (SSP) numbers are assigned to each protein spot by PDQuest™ software.
2 Mean density ratios observed for smokers compared to never smokers.
3 Accession numbers according to entries in UniProtKB/Swiss-Prot.

Table 8.1: Representative examples of regulated and non regulated proteins identified in bronchoalveolar lavage found at presence rates ≥ 70% in either the never smoking or smoking groups. (Adapted from paper III)

1. What proportion of SSP identities are regulated in smokers compared to never smokers;
2. What is the relationship between presence rate and fold expression factor;
3. Are certain sets of SSPs associated with smoking?

By using the model, we could easily define modal distributions of protein expression at the level of individual SSPs. This enabled us to categorize the expressions as either up regulated, down regulated or not regulated, when comparing the smoking groups to the never smoking group. Irrespective of the presence rate, approximately 85% of the SSP annotation identities showed similar expression patterns between -4 and +4 fold differences of the mean density ratios in both never smokers and smokers (Figure 8.5 Insert).

Reflecting clear biological meaning regulation is therefore defined as $X < -4$, $X > 4$, where $X$ is density ratio. Approximately 2% of the SSP identities were down regulated in smokers. Among the remaining 13% of the protein identities, a bimodal highly induced expression pattern in the smoking group was observed. This relative trend was maintained over the range of low to high presence rates. Further analysis of spot identification was performed using a conservative set point that required $\geq 70\%$ presence rate (Figure 8.5, blue bars) in either the never smoking or the smoking cohorts. Representative spots from 406 SSPs annotations were identified by MALDI-TOF MS and MALDI-TOF/TOF MS utilizing the MASCOT search engine. In total, MS and MS/MS identified 200 proteins in the BAL proteome, of which about 15% were exclusively present, or regulated significantly within the smoking subject groups. Table 8.1 shows the regulated and non regulated protein identities that were found at presence rates $\geq 70\%$, in either the never smoking or smoking groups. These included a number of well described proteins associated with re-do reactions, immune reactivity, and inflammation.

### 8.2.1.2 Statistical distributions and separations of SSP profiles of subjects

Statistical analysis on all annotated protein spots on the 2-D gels were performed using multivariate analysis. In the principle component analysis (PCA) and the partial least squares discriminant analysis (PLS-DA), not only singular SSP characteristics (presence/relative expression level vs group behavior) are related, but they are also related to the same parameters in all the other SSPs in the dataset. Using PCA analysis of the 406 protein identities expressed by at least 70% of either the never smoking or smoking cohorts, the smokers were unambiguously ($p<0.001$) and accurately ($R^2=0.84$) separated from the never smokers based on composite protein expression phenotype (Figure 8.6 A). There is a progression of dimensional features which obliquely separate the never smokers (green) from light smokers (blue) to heavy smokers (red). In this analysis there is some degree of spontaneous separation. In particular the never smokers stand out as a group. Having established the group dynamics of the dataset of 406 SSPs, the level of predictive accuracy that could assign any given subject to a specific group designation based solely on their individual
Results

Figure 8.6: Multivariate analysis of the bronchoalveolar lavage protein identities expressed by the never smoking or smoking cohorts. (A) PCA analysis of the 406 bronchoalveolar lavage protein identities expressed by at least 70% of either the never smoking or smoking cohorts (never smokers, green; light smokers, blue; heavy smokers, red). The encircled patients were found to have developed moderate COPD in the follow up study. (B) Successive rounds of PLS-DA analysis of the 29 smokers, utilizing either 200, 100, 50, 25, or 10 SSP components. The Q2 scores for predicting the COPD clinical outcome rises significantly from a negative value using all 406 components to a high predictability of 0.8, applying 100 SSPs. (C) PLS-DA analysis of the set of 100 SSPs (COPD, black; smokers, red). The ellipse in represents the 95% confidence region based on Hotelling T2. (Adapted from paper III)

SSP profile was tested using PLS-DA. The Q2 measures of group predictability were 0.78, 0.54 and 0.69, respectively for never, light, and heavy smokers. In this analysis, 1.00 defines perfect prediction. These results validated that the qualifiers used as dimensions in these comparisons of SSP datasets were significantly related to a specific group character.

8.2.1.3 Clinical outcome at 6-7 years follow up

At a time 6-7 years after the original BAL sampling, the study cohort was asked to participate in a follow up evaluation by spirometry, high-resolution computed tomography (HRCT), and clinical examination. The subjects at this time were 67-68 years old. A total of 7 of the 29 of both the light and heavy smokers studied earlier, were found to have developed moderate COPD (GOLD stage 2) during the follow up period (Figure 8.6 A, encircled). None of the never smoking developed either mild or moderate COPD. As shown in Table 5.1, the patients with developed COPD had developed significant airway obstruction as seen in the decline in FEV1 (% predicted) scores, compared to the other smokers.

Using the set of 406 SSPs expressed at a 70% presence rate, it was not possible to separate the protein expression profiles of the seven COPD patients from the other 22 heavy or light smokers using the Q2 predictability measure (Figure 8.6 B). By re-analyzing the 29 smokers by successive rounds of PLS-DA analysis, utilizing either 200, 100, 50, 25, or 10 SSP components it was tested whether a subset of the 406 SSPs might be more significantly
<table>
<thead>
<tr>
<th>Protein</th>
<th>SSP</th>
<th>Protein spot (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta-1B-glycoprotein</td>
<td>5702, 4802, 4805</td>
<td>3</td>
</tr>
<tr>
<td>Serum albumin fragments</td>
<td>3408, 5509, 5610, 4608, 4412, 4607, 5401, 5507, 5510, 7018, 5512, 7605, 6608, 5612, 5613, 4411, 6601, 5502, 7604, 6115, 6613, 5405, 3610, 8603</td>
<td>24</td>
</tr>
<tr>
<td>Tropomyosin (Isoform 6)</td>
<td>1319</td>
<td>1</td>
</tr>
<tr>
<td>Ig (hypothetical 53,4 kDa protein)</td>
<td>6709</td>
<td>1</td>
</tr>
<tr>
<td>Human albumin-thrombin-alpha-MSH</td>
<td>4008</td>
<td>1</td>
</tr>
<tr>
<td>Actin</td>
<td>3502, 6617</td>
<td>2</td>
</tr>
<tr>
<td>Transthyretin</td>
<td>4004</td>
<td>1</td>
</tr>
<tr>
<td>Secretory IgA</td>
<td>4808, 4801</td>
<td>2</td>
</tr>
<tr>
<td>Mutant alpha-1-antitrypsin</td>
<td>701</td>
<td>1</td>
</tr>
<tr>
<td>Alpha-1-antitrypsin</td>
<td>1703, 1704</td>
<td>2</td>
</tr>
<tr>
<td>Actin, cytoplasmic 1</td>
<td>3512</td>
<td>1</td>
</tr>
<tr>
<td>Glutathione S-Transferase</td>
<td>4112, 3107</td>
<td>2</td>
</tr>
<tr>
<td>Alpha-1-antichymotrypsin</td>
<td>704</td>
<td>1</td>
</tr>
<tr>
<td>Cathepsin D</td>
<td>8801</td>
<td>1</td>
</tr>
<tr>
<td>Annexin III</td>
<td>5305</td>
<td>1</td>
</tr>
<tr>
<td>Alpha-1B-glycoprot.</td>
<td>3801</td>
<td>1</td>
</tr>
<tr>
<td>Complement component 1 Q subcomponent-binding protein, mitochondrial</td>
<td>3602</td>
<td>1</td>
</tr>
<tr>
<td>Heat shock-related 70 kDa protein 2</td>
<td>4804</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 8.2: Proteins found in the PLS-DA analysis that separate COPD patients (GOLD stage 2) from smokers free of COPD during the follow-up.
Results

associated with clinical outcome. As shown in Figure 8.6 B, the Q2 scores for predicting the COPD clinical outcome rises significantly from a negative value using all 406 components to a high predictability of 0.8, applying 100 SSPs. The PLS-DA analysis on this set of 100 SSP identities showed that all seven COPD patients could be well separated from the other smokers by both T1 and T2 dimensions (Figure 8.6 C). About half of this set of 100 SSP identities were identified using MALDI-TOF. Theses protein identities are shown in Table 8.2.

This set of 100 differentially expressed SSP identities was further analyzed using 2D hierarchal clustering. The data were normalized by subtracting the

![Figure 8.7: Heat map of the intensity scores of the 100 differentially expressed SSP identities from the 29 smokers. (Adapted from paper III)](image-url)
mean intensity and dividing by the standard deviation separately for each SSP. In Figure 8.7 is a heat map of the intensity scores of all 29 smokers aligned for similarity in X axis by individual to individual and on the Y axis by the expression behavior characteristics of the individual SSP identities. The smokers who developed COPD segregated together on the left side of the dendrogram except for one subject (Pt 855) whose clinical characteristics were not unsimilar from the others. The color scale describing low to high expressions (green to red). Thus, the set of 100 SSP identities associated with COPD outcome were shown to be composed of statistically related proteins that occurred at either much lower levels or much higher levels than seen in the other smokers. It can be concluded that certain patterns of protein expressions in BAL as defined by presence, absence, and intensity, can be associated with different rates of disease progression in subjects sharing the risk behavior for eventual disease development, such as age, smoking, and geographical location.

8.3 HPLC-MS/MS BAL proteome analysis

Due to the huge dynamic range of concentrations of BAL proteins and the inherent limitation of the 2-D gel electrophoresis-based technique, the approach has only been able to identify proteins present in medium to high abundance in BAL samples. In 2004, about 100 different human BAL proteins had been identified using 2-D PAGE. To provide a more comprehensive BAL proteome database, one-dimensional reversed phase liquid chromatography coupled with tandem mass spectrometry was used to analyze the protein composition in BAL fluid. In contrast to 2-D gel electrophoresis where intact
proteins are separated, liquid chromatography separates peptides of digested protein mixtures.

![Graph showing the number of proteins identified in BAL fluid, with different assigned GO biological processes.](image)

**Figure 8.9:** Diagram showing the number of proteins identified in BAL fluid, with different assigned GO biological processes. (Adapted from paper II)

The 90-min LC/LTQ platform identified 268 proteins in the pooled BAL samples from never smokers, 309 proteins in the pooled samples from light smokers, and 314 proteins in the pooled samples from heavy smokers (paper II). Approximately one third (n = 130) of all proteins identified were identified in all three groups. However, a substantial number of proteins were identified in either the samples from smokers (n = 137) or never smokers (n = 63). These groups of unique proteins included both high-abundance and medium-abundance proteins. The majority of these proteins have not been reported previously in BAL fluid. The five most abundant proteins corresponded to generally recognized major components of BAL fluid: albumin, transferrin, α1-antitrypsin, IgA, and IgG. In the case of albumin, the protein was identified with 83% sequence coverage, whereas for transferrin and α1-antitrypsin, the proteins were identified with 69% and 56% sequence coverage, respectively. These proteins were present in samples from all three groups. IgA and IgG were identified by their heavy and light chains, where each individual chain had a sequence coverage between 30% and 90%. Typical relative standard deviation (RSD) values for the high-abundance proteins albumin, transferrin, and IgG, based on triplicate runs, were 6.7%, 15%, and 26% for the never smokers and 3.1%, 17%, and 22% for the heavy smokers. The majority
of proteins identified on the 2-D gels were also identified by LC-MS/MS.

The entire dataset of annotated proteins identified by at least 2 peptides in the BAL fluid of the pooled BAL samples is shown in Table 1 online Data Supplement (paper II). The proteins were annotated according to Gene Ontology (GO). It must be taken into consideration that proteins are commonly assigned to more than a single molecular function and biological process. Accordingly, Figure 8.8 and figure 8.9 show bar diagrams of annotated proteins identified by at least 2 peptides in the BAL fluid of the pooled samples. For graphical reporting, the proteins were categorized into molecular function and biological process. It is clear from the GO cataloging that there is a broad distribution of both protein molecular functions and biological processes within the human BAL proteome.

8.3.0.4 BAL protein differential expression comparison by mass spectrometry

<table>
<thead>
<tr>
<th>ID</th>
<th>Protein name</th>
<th>Peptide</th>
<th>m/z</th>
<th>Hits</th>
<th>Scores</th>
<th>Peak areas</th>
</tr>
</thead>
<tbody>
<tr>
<td>KPY M</td>
<td>Pyruvate kinase, isozymes</td>
<td>LDIDSPITTAR</td>
<td>599.3</td>
<td>1</td>
<td>10.2</td>
<td>4.4</td>
</tr>
<tr>
<td>CAT D</td>
<td>Cathepsin D</td>
<td>LLDIACWIIHK</td>
<td>703.8</td>
<td>2</td>
<td>20.1</td>
<td>0.4</td>
</tr>
<tr>
<td>GDI R</td>
<td>Rho GDP-dissociation</td>
<td>SIQEJELDDDESRLR</td>
<td>682.9</td>
<td>1</td>
<td>10.1</td>
<td>0.2</td>
</tr>
<tr>
<td>KCY</td>
<td>UMP-CMP kinase</td>
<td>KNPDSQYGELIEK</td>
<td>760.8</td>
<td>1</td>
<td>10.1</td>
<td>0.5</td>
</tr>
<tr>
<td>GST A2</td>
<td>Glutathione S-transferase A2</td>
<td>NDGYLMFQVPMVEIDGMK</td>
<td>855.3</td>
<td>5</td>
<td>50.2</td>
<td>0.4</td>
</tr>
<tr>
<td>AKAP 9</td>
<td>A kinase anchor protein 9</td>
<td>KAYINTISSLKDLITK</td>
<td>905.3</td>
<td>4</td>
<td>40.1</td>
<td>0.2</td>
</tr>
<tr>
<td>SCN 1 A</td>
<td>Sodium channel protein, brain</td>
<td>FMASNP5K</td>
<td>883.0</td>
<td>4</td>
<td>40.1</td>
<td>0.3</td>
</tr>
<tr>
<td>1 alpha subunit</td>
<td></td>
<td>FDKDAFLAGTFISTVGIDFR</td>
<td>1081.8</td>
<td>4</td>
<td>40.1</td>
<td>0.5</td>
</tr>
</tbody>
</table>

a The protein ID and name is based on assignment from the non-redundant protein database human_rapido.fasta which contain Genbank and Swiss-Prot protein entries using Sequest. The nomenclature has been simplified to reflect the id. of the protein. b The number of redundant peptide identifications of the given protein. c Sequest protein score. d Peptide peak area.

Table 8.3: Proteins showing regulatory differences among the never smokers and the smokers according to the number of peptide hits, the Sequests protein score and according to peak areas. (Adapted from paper II)

The aim of the LC-MS/MS study was to explore the human BAL proteome qualitatively. However, to estimate relative changes in protein abundances corresponding to chronic exposure to cigarette smoke, the number of peptides leading to identification and the Sequest score in conjunction with total peptide ion intensity of a selected peptide fragment were compared in never smokers and in chronic smokers (paper II) (see section “Proteomic Methods”). As an example of quantitative regulation, the Sequest score and the number of peptide sequencing events showed an up regulation of UMP-CMP kinase among the heavy smokers. This up regulation was confirmed by comparison
of peak areas of the peptide KNPDSQYGELIEK (m/z 760.8) for UMP-CMP kinase.

Peak-area measurements from triplicate analysis gave an RSD of 12%. The peak areas of selected peptides used for preliminary quantification of the proteins are shown in Table 8.3. As an example of significant regulation in terms of presence-absence, the Sequest score and the number of peptide sequencing events showed up regulation of Cathepsin D among the smokers and of glutathione S-transferase A2 among the heavy smokers. Cathepsin D was below the detection limit in samples from the never smokers, and glutathione S-transferase A2 was below the detection limit in samples from light and never smokers. These cases of presence-absence were confirmed by comparison of peak areas of the peptide LLDIACWIHHK (m/z 703.8) for cathepsin D and the peptide NDGYLMFQQVPMEIDGMK (m/z 855.3) for glutathione S-transferase A2.

For protein differential analysis of the profile mode generated LC-LTQ and
LC-LTQ-FT generated data, *DeCyder™* MS, was used. This is a new software program, for visualization, detection, comparison, and label-free relative quantitation of LC-MS/MS data [145] (see section “Proteomic Methods”).

<table>
<thead>
<tr>
<th>Protein</th>
<th>Accession Number</th>
<th>Regulation</th>
<th>Presence COPD pat</th>
<th>Presence Smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha-1-acid glycoprotein 2 precursor</td>
<td>P19652</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Alpha-1-acid glycoprotein 1</td>
<td>P02763</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Alpha-1-antitrypsin precursor</td>
<td>P01009</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Annexin A2</td>
<td>P07355</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Annexin A5</td>
<td>P08758</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Apolipoprotein A-I precursor</td>
<td>P02647</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Apolipoprotein A-II precursor</td>
<td>P02652</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Calgranulin B</td>
<td>P06702</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Calmodulin</td>
<td>P62158</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Complement C3 precursor</td>
<td>P01024</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Cystatin B</td>
<td>P04080</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Epididymal secretory protein E1 precursor</td>
<td>P61916</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Fructose-1,6-bisphosphatase</td>
<td>P09467</td>
<td>↑ ↓</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Fructose-1,6-bisphosphatase isoyme 2</td>
<td>O00757</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Glutathione S-transferase P</td>
<td>P09211</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Haptoglobin precursor</td>
<td>P00738</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Haptoglobin-related protein precursor</td>
<td>P00739</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Haptoglobin precursor</td>
<td>P00738</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Hemoglobin alpha chain</td>
<td>P69905</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Hemoglobin beta chain</td>
<td>P68871</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Hemoglobin delta chain</td>
<td>P02042</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Hemoglobin gamma-2 chain</td>
<td>P69892</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Hemoglobin subunit beta</td>
<td>P68871</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Hemoglobin subunit delta</td>
<td>P02042</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Hemoglobin gamma-1 chain</td>
<td>P69891</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Hemoglobin epsilon chain</td>
<td>P02100</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Hemopexin precursor</td>
<td>P02790</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Ig alpha-1 chain C region</td>
<td>P01876</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Ig gamma-3 chain C region</td>
<td>P01860</td>
<td>↑</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Ig gamma-4 chain C region</td>
<td>P01861</td>
<td>↑</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Ig gamma-2 chain C region</td>
<td>P01859</td>
<td>↑</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Ig gamma-1 chain C region</td>
<td>P01857</td>
<td>↑</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Ig kappa chain C region</td>
<td>P01834</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 8.4 show the results of proteins found regulated between the COPD study cohort and the seven smokers who developed moderate COPD (GOLD stage 2) at follow-up.
<table>
<thead>
<tr>
<th>Protein</th>
<th>Accession Number</th>
<th>Regulation</th>
<th>Presence COPD pat</th>
<th>Presence Smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insulin-like growth factor binding protein 2 precursor</td>
<td>P18065</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Lactotransferrin precursor</td>
<td>P02788</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Lysozyme C precursor</td>
<td>P61626</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Neutrophil gelatinase-associated lipocalin</td>
<td>P80188</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Peptidyl-prolyl cis-trans isomerase A</td>
<td>P62937</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Peroxiredoxin 2</td>
<td>P32119</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Peroxiredoxin 1</td>
<td>Q06830</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Peroxiredoxin-5, mitochondrial</td>
<td>P30044</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Polymeric-immunoglobulin receptor precursor</td>
<td>P01833</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Saposin D</td>
<td>P07602</td>
<td>↑</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Profilin-1</td>
<td>P07737</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Proline-rich protein 4 precursor</td>
<td>Q16378</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Rho GDP-dissociation inhibitor 2</td>
<td>P52566</td>
<td>↑</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Transferrin</td>
<td>P02787</td>
<td>↑ ↓</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Serum albumin</td>
<td>P02768</td>
<td>↑ ↓</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Superoxide dismutase [Cu-Zn]</td>
<td>P00441</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Thioredoxin (ATL-derived factor) (ADF) (Surface)</td>
<td>P10599</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Transthyretin [Precursor]</td>
<td>P02766</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Trefoil factor 3 precursor</td>
<td>Q07654</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>U1 small nuclear ribonucleoprotein 70 kDa</td>
<td>P08621</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Clara cell 10 kDa protein</td>
<td>P11684</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>WAP four-disulfide core domain protein 2 [Precursor]</td>
<td>Q14508</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Zinc-alpha-2-glycoprotein precursor</td>
<td>P25311</td>
<td>↑</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Alpha-1-acid glycoprotein 1</td>
<td>P02763</td>
<td>↓</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 8.4: Table showing proteins regulated \((p=0.05)\) between the COPD study cohort and the seven smokers who developed moderate COPD at follow-up, using DeCyder™ MS (the results come from the LC-LTQ generated data) Accession numbers according to entries in UniProtKB/Swiss-Prot.
8.4 Validation of LC-MS/MS with individual samples and 2-D gels

We were interested in determining how well the protein profiles obtained with pooled samples compared with the possible ranges of expression present in individual samples. To determine the relative abundances and presence rates of the BAL proteins identified in the pooled samples, we ran BAL sample from each of the 48 study individuals separately on the LC-MS/MS platform. A comparison of the number of protein identities found in pooled or individual samples in never smokers and heavy smokers is shown in Figure 8.10.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Group</th>
<th>Presence rate*</th>
<th>Sequest Score (mean ± SD)</th>
<th>Peptides identified range</th>
<th>Peptide Presence In pool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annexin II</td>
<td>never-smokers</td>
<td>18/18</td>
<td>56 ± 32</td>
<td>1-7</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>smokers</td>
<td>29/30</td>
<td>68 ± 26</td>
<td>2-8</td>
<td>6</td>
</tr>
<tr>
<td>Serotransferrin</td>
<td>never-smokers</td>
<td>18/18</td>
<td>580 ± 184</td>
<td>4-36</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>smokers</td>
<td>30/30</td>
<td>402 ± 196</td>
<td>3-38</td>
<td>44</td>
</tr>
<tr>
<td>Alpha-1 Antitrypsin</td>
<td>never-smokers</td>
<td>18/18</td>
<td>249 ± 79</td>
<td>5-19</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>smokers</td>
<td>30/30</td>
<td>164 ± 86</td>
<td>2-20</td>
<td>18</td>
</tr>
<tr>
<td>Cathepsin D</td>
<td>never-smokers</td>
<td>1/18</td>
<td>10</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>smokers</td>
<td>15/30</td>
<td>36 ± 28</td>
<td>1-8</td>
<td>2</td>
</tr>
<tr>
<td>Calmodulin</td>
<td>never-smokers</td>
<td>18/18</td>
<td>35 ± 18</td>
<td>1-4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>smokers</td>
<td>25/30</td>
<td>50 ± 20</td>
<td>1-4</td>
<td>3</td>
</tr>
<tr>
<td>Insulin-like growth factor</td>
<td>never-smokers</td>
<td>3/18</td>
<td>10 ± 0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>binding protein</td>
<td>smokers</td>
<td>19/30</td>
<td>18 ± 6</td>
<td>1-2</td>
<td>1</td>
</tr>
</tbody>
</table>

* The presence rate among the never-smokers (n=18) and smokers (n=30).

Table 8.5: Comparison of the number of peptides identified in the pooled BAL samples with the range of peptides found in individual BAL samples. (Adapted from paper II)

We observed a wide variation in the total numbers of proteins that could be identified in a given sample (range, 48-314), irrespective of smoking history. The pooled samples achieved higher numbers of protein annotations than any of the individual samples. This is likely the result of an additive effect from several of the group members because of the lack of obvious outliers and the tight clustering seen in the distributions of the number of identities detected in both groups. In general, the patterns of common or unique protein identities observed in each of the group pools were also maintained by the individuals within the group. Analysis of individual samples allowed us to annotate the
relative presence rates of the individual identifications in the different samples.

<table>
<thead>
<tr>
<th></th>
<th>Never-smokers</th>
<th>Heavy smokers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>presence rate</td>
<td>example 1</td>
</tr>
<tr>
<td>Rho GDP-dissociation Inhibitor 1</td>
<td>14/18</td>
<td>15/15</td>
</tr>
<tr>
<td>Cathepsin D</td>
<td>10/18</td>
<td>14/15</td>
</tr>
<tr>
<td>Alpha-1 Antitrypsin</td>
<td>18/18</td>
<td>15/15</td>
</tr>
</tbody>
</table>

*The presence rate among the never smokers (n=18) and heavy smokers (n=15). (Adapted from paper II)

Figure 8.11: Representative examples of 2-DE spot patterns of proteins in BAL of individual never smoking or smoking subjects also identified in the pooled BAL samples using the LC-MS/MS platform. As an example, Table 8.5 shows representative examples of BAL proteins commonly identified in samples from both smokers and never smokers as well as proteins found only in samples from smokers. We compared the number of peptides identified in the pooled samples with the range of peptides found in individuals and found close agreement. We found no evidence that the pooled protein profiles were skewed because of the contribution of singular individuals who contributed a dominant selection of proteins to the pool. Altogether, these data indicate that the profiles for the pooled samples generally mirrored the entire group of individuals.

Using separate aliquots of the same samples, the protein expression profiles of a set of high- to medium-abundance proteins, identified by both 2-D PAGE and LC-MS/MS, were compared. Figure 8.11 presents representative examples of 2-dimensional electrophoresis spot patterns of proteins in BAL fluid from individual never smokers or smokers also identified in the pooled BAL samples by the LC-MS/MS platform. The Rho-GDP dissociation inhibitor protein, which was identified only in the smokers and not in the never smokers by LC, showed a strong spot in the smokers and only a weak spot in the never smokers. A similar concordant result was found for Cathepsin D, which showed consistency in the relative detectable concentrations of protein on the two platforms. The last example shown, α1-antitrypsin, showed a high relative abundance on the gels of both smokers and never smokers. To-
together, the results of the direct comparison of the protein expression profiles of individuals obtained by both 2-D PAGE and LC-MS/MS showed remarkable similarity to the relative distributions of proteins detected in the pooled samples.
9. Discussion

Proteomic studies provide orders of magnitude of quantitative and qualitative information. Proteomic techniques have the advantages of being able to simultaneously study a subset of all proteins as opposed to a single protein. However, so far protein marker studies have mainly been focused on single “candidate” protein approaches [62]. Patterns of disease-related changes, including novel disease marker proteins, would provide substantially more useful clinical information than a single marker. Today novel technologies from proteomics and functional genomics allow unbiased assessment of how patterns of proteins differ in various disease states [153]. This clinical proteomics approach provide systematic, comprehensive, large-scale identification of protein patterns (“fingerprints”) of disease that can provide clinically useful information about susceptibility to disease, diagnosis, prognosis, and guided therapy. This has changed the knowledge of the pathogenesis of diseases and will thereby influence management in the future [62].

In this thesis we present a strategy for relating the measurement of sets of proteins with phenotypes of clinical presentation. We have chosen to analyze bronchoalveolar lavage (BAL) because of its ability to address secreted and extracellular proteins present within the central and descending airways of the smokers and never smokers studied. In order to accomplish this we have developed and utilized an interdisciplinary toolbox that includes protein separation (two-dimensional gel electrophoresis and liquid chromatography), mass spectrometry identification platforms and statistical methods for multivariate analysis. This allows the grouping of study subjects by individual presence/absence and intensity scores for each of the differentiated protein spot annotations.

The key component for pursuing such a study was the careful selection of a well characterized clinical material, both in terms of being associated with the biology of interest, but also for a low sampling variability. The study material used in this thesis consists of age matched men all born in 1933, living in one city, differing by lifelong smoking history, and compared by clinical function measurements and histological assessment at the same relative time points. By reducing the variability of the input samples we could concentrate on refining the technology and analysis platforms to standardize the quantitation and normalization methods in order to support our need for an unbiased as-
essment. The technology we adopted could have a general applicability to a wide variety of clinical samples, including plasma, serum, urine, sputum and other solubilized tissue components such as targeted cells obtained by laser capture microscopy.

Protein expression profiles can be used as diagnostic datasets composed of groups of signals representing the various states of numerous singular proteins. Due to the broad dynamic range of protein concentrations present in any given sample volume, it is attractive to take advantage of whatever associations of regularity that exists between individual protein identities, sets of proteins within given clinical samples, or given clinical presentations. In order to accomplish this level of segregation, it is necessary to link a specific phenotype of the clinical presentation to that exclusive protein set. However, validation of models based on associating proteins to disease can best be achieved by relating these values to the clinical outcome.

After 6-7 years we evaluated the clinical outcome and a group of subjects was identified who had developed moderate chronic obstructive pulmonary disease (COPD) GOLD stage 2. Seven of the 29 light and heavy smokers studied earlier were found to have developed moderate COPD. None of the never smokers developed neither mild nor moderate COPD. These eventual COPD patients shared a common distinct protein expression profile at the baseline BAL sample that could be identified using partial least squares discriminant analysis. This pattern was not observed in BAL samples of asymptomatic smokers free of COPD during the follow-up. The proteomic analysis of the seven COPD patients also showed a differential expression compared to the seven smokers who were found to have developed COPD.

Our results showed that the protein composition of BAL samples from smokers as a group was significantly different from the BAL samples of the age matched never smoking control group. Further, the smoking subjects were found to have lost the expression of a small subset of identities present in never smokers. They had however, acquired a high expression level of SSP identifies that were absent in the never smokers. We thus demonstrated that lifelong smokers develop significantly different protein expression patterns in their respiratory tract than age matched never smokers.

Thorough understanding of early protein abnormalities associated to disease, and knowledge of the molecular processes and mechanisms underlying the disease will enhance our ability to identify patients at risk, and those with preclinical disease. In our material, protein up regulation was the mainly observed feature but down regulation is also important to consider as it may be important in the pathogenesis of diseases. As an example we found secretory Immunoglobulin A (IgA) down regulated among smokers compared to never smokers. This regulation has also been shown by others [119, 154, 155]. Secretory IgA was found to be predictive for the development of COPD among
smokers.

Immunoglobulin A, is classified into serum and secretory IgA. Serum IgA is a monomer, and secretory IgA consists of two molecules of IgA joined by the peptide J chain and one secretory component [154]. Secretory IgA has been suggested to play an important role in the protection of the lung mucous membrane [156]. Cigarette smoke may suppress bronchial IgA production and hence disturbs the local immunity of the lung [154].

Cathepsin D, Glutathione S-Transferase, HSA fragments and Annexin A5 are examples of proteins that were found up regulated among the smokers compared to the never smokers. The first three proteins also predicted the development of COPD among the smokers, and all were found up regulated in the COPD population. Cathepsin D is an acidic, low molecular weight cystein protease, involved in tissue repair and cell proliferation. The presence of Cathepsin D in BAL correlates with reports of its expression by alveolar macrophages, bronchial epithelial cells and type I pneumocytes [157]. Cathepsin D has been shown to increase in BAL from subjects with idiopathic pulmonary fibrosis [158], and this increase has been found consistent with its potential role in remodeling processes occurring during fibrogenesis. Studies also show the involvement of cathepsins in lung emphysema [159, 160]. Thus there are several theoretical possibilities where Cathepsin D can be involved in the pathogenesis of COPD.

Human glutathione S-transferases are a functionally diverse family of soluble detoxification enzymes that use reduced glutathione in conjugation and reduction reactions to eliminate many different toxic electrophiles and products of oxidative stress [161]. Oxidative stress contributes to the development of both lung cancer and chronic obstructive pulmonary disease [31, 162, 163]. Therefore it might be of certain interest that we have found elevated levels of a detoxifying agent such as Glutathione-S transferase in smokers. A hypothesis could be that the protein is up regulated among the smokers to protect them from the oxidants in the cigarette smoke.

The oxidants in cigarette smoke cause lung injury through a number of mechanisms including:

- the depletion of glutathione and other antioxidants;
- the initiation of redox cycling mechanisms;
- enhancement of the respiratory burst in neutrophils and macrophages;
- inactivation of protease inhibitors such as α1-antitrypsin inhibitor;
- direct damage to lipids, nucleic acids and proteins [164].

The oxidation of proteins may play an important role in the pathogenesis of chronic inflammatory lung disease, as higher levels are measured in diseases such as cystic fibrosis, asbestosis, and idiopathic pulmonary fibrosis compared with healthy control measurements. Studies have also shown that the oxidative burden in lungs is increased in patients with COPD and may be involved in
the pathogenetic processes in the lung, and in the systemic manifestations of weight loss and muscle dysfunction [10]. A recent paper by Nagai et al [165] has also shown evidence of lung damage from oxidative stress due to cigarette smoking. They show that older smokers with long term smoking histories had excessive protein carbonyls in BAL fluid and that the oxidation of albumin has been shown to account for the excessive total protein carbonylation. The protein oxidation can lead to protein degradation. Our results showed that there was an increase of albumin fragments among smokers and COPD patients. Among the smokers who developed COPD, the albumin fragments were important predictors for the disease. A hypothesis could be that the oxidative burden is increased in the lungs of patients who will later develop COPD and it is this protein oxidation that leads to the degradation of albumin.

Annexin A5 belongs to the annexin superfamily of calcium and phospholipid-binding proteins [166]. The Annexins were first identified as intracellular proteins and was attributed intracellular functions. However, some Annexins (Annexin A1, Annexin A2 and Annexin A5) have also been found in several tissues as both a soluble and a membrane-bound protein [167], although it lacks a signal peptide and its mechanism of secretion is unknown.

It is not yet clear why this intracellular protein was present in high abundance in BAL. The possible causes may include:
1. cellular proteins shed into lung airway as a result of increased cell apoptosis and lysis;
2. increased leakage of those proteins circulating in the blood, and
3. protein secretions by pathways that remain largely unknown [76].

Annexin A5 has been found to be induced by hypoxic stress [168]. An explanation for the up regulation of Annexin A5 among the smokers compared to the never smokers could be to protect smokers from the increase in oxidative-stress caused by cigarette smoke.

The Annexins have also received attention as a putative mediator of the anti-inflammatory effects of glucocorticoids [169]. Another explanation for the up regulation of Annexin A5 among the smokers might be its anti-inflammatory effect. Increased levels of anti-Anx V antibodies have been found in the sera of patients with lupus erythematosus [170] and rheumatoid arthritis [171].

VEGF is fundamental in the development and maintenance of the vasculature [172]. Annexin A5 has been suggested to function as a signaling protein for VEGFR-2 by directly interacting with the intracellular domain of the receptor and appears to be involved in regulation of vascular endothelial cell proliferation mediated by VEGFR-2 [173]. Inhibition of VEGFR-2 in rat lungs has also been shown to cause endothelial and epithelial apoptosis followed by airspace enlargement [174]. Another explanation for the up regulation of Annexin A5 among smokers might therefore be that it has a
protective role against emphysema.

The traditional approach for characterizing proteomes is to use reductional methods that identify and characterize the individual components by MS and then differentially relate that proteome to a biological question or end point. Individual MS identities act as important annotation landmarks for comparison within or between studies but also for establishing a biological context to the model. Our results indicate that the information combined in large datasets of individual SSP presence/absence and abundance scores has sufficient power to accurately predict, identify and assign a stable phenotype. By using the multivariate algorithms, we proved that a high level of overall predictability of group association was maintained by both study cohorts. This result is important as it implies that stable predictive phenotypes of protein expression can be determined and used to segregate subjects even in clinical cohorts that consist of asymptomatic and clinically healthy subjects. This segregation can be accomplished despite the requirement of assigning actual protein identities to each of the components within the clinical proteome.
10. Future perspectives

Pathologic processes are often associated with changes in the expression and the modification of proteins. Through identification of protein expression patterns associated with disease, clinical proteomics has a great potential to broaden both our understanding as well as the treatment of lung diseases (295, 369). A crucial step in the treatment of COPD is to define disease markers that allow an early diagnosis. Studies presented in recent years have demonstrated that approaching the lung proteome is possible using clinical samples such as BAL fluid, pulmonary edema fluid, or breath condensates [175, 74, 133, 176, 72, 111, 177]. Improvements in BAL protein identification have increased the clinical relevance of such studies and have given further insights into the underlying pathological processes.

So far there have not been many studies published on the detection of post-translational modifications of BAL proteins. There is however a growing interest in identifying proteins in BAL which undergo post-translational modifications. It appears obvious that such unidentified proteins, present amongst the wide variety proteins in BAL, could serve as useful and potentially novel biomarkers for diagnosing lung diseases. Today’s proteomic approaches provide a promising tool for this. Future research projects in clinical proteomics, where large scale identification of disease protein patterns is made, will serve as an important bridge for connecting our growing appreciation of biological complexity with our ultimate goal of understanding and treating a disease [62].

While a smaller number of well-characterized clinical samples are critical for protein marker discovery, a larger number, often thousands, of patient samples are required to validate the discovered proteomic fingerprints in order to progress them toward routine clinical use [62].

To be more efficient, BAL proteome research needs an increased international cooperation with the aim of promoting better rationalization of BAL proteomic studies. A better rationalization seems to be necessary for future differential-display proteome analysis in the sample preparation and in the characterization of patients (age, sex, smoking or not smoking, status of lung pathologies .etc) [72, 62].

Proteomics is not the only way to broaden our understanding of lung diseases; several other methods could also be used. Lung imaging methods may
Future perspectives

become important tools for measuring the outcome of COPD through time- and age-dependent lung structural changes [178]. Promising imaging methods are high resolution computed tomography (HRCT), micro-CT and Magnetic Resonance Imaging (MRI). In particular, MR imaging with hyperpolarized noble gases appears very promising for lung imaging [179, 180].

The possibility to localize and follow changes in organisms at the molecular level by imaging component distributions throughout specific tissues is of prime importance. Changes in the lung of COPD patients have the potential to be evaluated by direct tissue profiling and imaging mass spectrometry. Matrix-assisted laser desorption/ionization mass spectrometric imaging (MALDI MSI) takes full advantage of the high sensitivity of mass spectrometry instrumentation [181, 182, 183]. It also uses the ability of the latter to simultaneously detect a wide range of compounds, almost regardless from their nature and mass. In MALDI MSI, protein patterns can be directly correlated to known histological regions within the tissue. Images can then be made of proteins specific for the morphological regions within a sample tissue. This methodology will almost certainly continue to be an expanding area of research.

In addition to these promising methodologies it may also become important to further develop pulmonary physiological measurements in COPD in order to enhance phenotypic profiling.

For a future clinical diagnostic test of COPD based on potential disease markers to be successful, it should be based on plasma. The advantage of using plasma based proteomic analysis is that blood samples are easily accessible. In theory, plasma should contain a large part of, if not all human proteins [184] and should therefore be an ideal target for a lung proteome approach. However, the dynamic range of protein concentrations in plasma is even greater than in other body fluids such as the epithelium lining fluid and BAL (\(\sim 10^{10} - 10^{12}\)) [184], and the concentration of pulmonary proteins in plasma is usually relatively low [185, 186, 187]. Consequently, this approach may only be useful to evaluate changes in a small subset of the lung proteome but might still be enough for a diagnostic test to be developed, which can be regarded as a promising clinical implementation [72].

The need to process minute quantities of large numbers of samples on which thousands of measurements can be made simultaneously is motivating the use of nanotechnology and microfluidic devices [188]. A rapidly maturing technology is protein microarrays, where a multitude of different protein molecules have been affixed on a chip surface at separate locations in an ordered manner, in a miniature array [189, 190]. These are used to identify protein-protein interactions, to identify the substrates of protein kinases, or to identify the targets of biologically active small molecules. The preferred method of detection currently is fluorescence detection. Fluorescent detection
is safe, sensitive, and can have a high resolution [191, 192]. This technology paves the way forward in high-throughput proteomic exploration [193].

Medicine is based on defining disease by patient history, physical examination, and various clinical laboratory parameters that enable guidance for effective treatment. The use of more sophisticated markers of disease, both in imaging and in the laboratory, will enable more targeted therapy. However, expansion of existing information technology infrastructure (bioinformatics and statistics) and training efforts (among clinicians and scientists) will be required to support effective progress toward these goals [62, 194].

Systems biology is a computational approach to study interactions between components of a biological system, and how these interactions give rise to the function and behavior of that system [188]. The understanding of large interrelated components of a system promises to transform how biology is done; DNA, RNA, proteins, macromolecular complexes, signaling networks, cells, organs, organisms and species within their environmental context. Predictive, preventive, personalized and participatory medicine will be the most obvious impact of systems biology, with the potential to transform medicine by decreasing morbidity and mortality due to chronic diseases [195]. The strong trend of technological development and improvements in all scientific fields will continue bringing new perspectives, ideas and insights into clinical proteomic, leading to improved patient treatment.

Denna avhandling syftar till att studera och karakterisera proteinerna i bronkskoljvätskan från personer som aldrig har rökt och kroniska rökare. Hypotesen var att proteinerna i bronkskoljvätskan speglar en persons rökvanor och att det hos de rökare som kommer att utveckla KOL finns karaktéristiska proteinmönster. Vi valde att studera bronkskoljvätska eftersom proteiner från de centrala luftvägarna och de nedre luftvägarna återfinns i denna biologiska vätska. En stor fördel med att studera bronkskoljvätska är att denna biologiska vätska erhålls från platsen för inflammationen, d.v.s. från luftvägarna.

För att studera proteinerna i bronkskoljvätskan har vi utvecklat analytiska metoder för tvådimensionella geler och vätskechromatografi. Därefter har proteiner identifierats med hjälp av masspektrometri. För att kunna dra slutsatser om proteinerna i bronkskoljvätskan speglar en persons rökvanor och om det hos de personer som kommer att utveckla KOL finns karaktéristiska proteinmönster utvecklades statistiska modeller.

Personerna vars bronkskoljvätska studerades, var alla födda i Göteborg 1933 och skiljde sig bara väsentligen åt i avseende på rökvanor. Samtliga personers lungfunktion undersöktes. Vid uppföljning 6-7 år undersöktes personerna igen och det visade sig då att en del av personerna utvecklat KOL. Hos personerna som utvecklat KOL kunde vi med hjälp tidigare utvecklade tekniker för proteinseparation finna att det fanns karaktéristiska proteinmönster. Dessa proteinmönster återfanns inte bland rökare som inte utvecklat KOL vid uppföljningen. Sammanfattningsvis pekar resultaten av denna avhandling på att proteinerna i bronkskoljvätskan speglar en persons rökvanor och att det hos de personer som kommer att utveckla KOL finns KOL karakteristiska proteinmönster.
Acknowledgements

I wish to express my sincere gratitude to all people who have been involved in my work and I would especially like to acknowledge:

Professor Claes-Göran Löfdahl, my supervisor and head of Division of Respiratory Medicine and Allergology, Lund University, for your encouragement, constant support and scientific guidance in respiratory medicine. György Marko-Varga my co-supervisor, and responsible for the Proteomics group at Biological Sciences, AstraZeneca R&D Lund, for your enthusiasm and guidance in the world of proteomics. Magnus Dahlbäck my co-supervisor at Discovery Medicine & Epidemiology, AstraZeneca R&D Lund, for your coordination and encouragement.

Ann Ekberg-Jansson, collaborator, for guiding and providing me with samples from her thesis. And also Kristina Andelid who is involved in the continued KOL-33 study.

All the people at AstraZeneca in Lund who have helped me from my master thesis work through my PhD. My co-authors Thomas Fehniger, Henrik Lindberg and Per Broberg. Tom, for all your relevant ideas and comments. Henrik, for sharing your great knowledge in the field of gel electrophoresis and proteomics. Per, for your expertise in statistics and bioinformatics. Claes, for your expert advice in the field of mass spectrometry. Sven and Jon for your valuable and encouraging scientific advice. Jon, also for proof-reading my thesis. My friend Susanne for all your tips and insights on being a PhD student. Karin and Eva for lighten up the workplace and making sure that everything runs properly. Karin also for handling the logistics. Krzysztof, Ib, Peder and Bosse for expert advice on bioinformatics. John, Sara, Carita and Robert, for always being there. Henrik Vissing and Kristina for showing great interest in my research work.

Pernilla Glader and Johan Malmström, former PhD students at AstraZeneca for making my PhD studies much more fun and for your encouragement and inspiring discussions. Pernilla for your friendship and Johan for being an inspiring scientific model.

Egidijus Machtejevas, for great company and discussions.

IPC, AstraZeneca, Lund for printing this thesis and my many posters during the years.

The studies were supported by AstraZeneca and the Swedish Heart Lung
Foundation.

Katarina Turesson, Division of Respiratory Medicine and Allergology, Lund University, who have helped me with all kinds of administrative things, in a very efficient way.

My friends and collaborators at Barnett institute, Northeastern University, Boston. My co-authors: Professor Bill Hancock, Marina Hincapie, Ziping Yang and Tomas Rejtar. Professor Barry Karger, Billy Wu and Jana Volf. PhD students Haven Baker, Xiaoyang Zheng, Christine Orazine, Majlinda Kulloli and Tatiana Plavina. Thank you all for creating a great scientific atmosphere.

My friends and collaborators at Thermo BRIMS Center, Cambridge, MA. Leo Bonilla, Jennifer Sutton and Tori Richmond-Chew, for use of mass spectrometers, and for your great support and generosity.

Anders Kaplan, GE Healthcare, Uppsala, for providing me with the latest versions of DeCyder™ MS, and important technical support.

Former students of mine, Cecilia Mårtensson and Anna Jonsson. I enjoyed very much working with you.

All of you in LURN. Gunilla Westergren-Thorsson, Leif Bjermer, Jonas Erjefält and Ellen Tufvesson, thank you for great meetings, conferences and discussions. Gunilla, for being a great scientific mentor and advisor. Former and present PhD students in LURN; Pernilla, Kristoffer, Kristian, Annika, Lizbet, David, Monika, Lena, Kristina, Mette, Lisa and Cecilia. Thank you for interesting and stimulating journal club discussions, and for making my PhD life even more fun.

Professor Anders Malmström and Cecilia Emanuelsson for recruiting me to Lund Graduate School of Biomedical Research and for your continued support. Anders, for being a great scientific mentor and advisor.

Professor John Yates lab at Scripps Research Institute for hosting me during the summer of 2006.

Professor Hakon Leffler, for being a great master thesis supervisor and for introducing me to the field of mass spectrometry and the U.S.

Professor Stig Steen and Trygve Sjöberg, for being great supervisors and for introducing me to thoracic surgery.

Simon, Sofia, Maria, Åsa, Niklas and Kerstin from Dep. of Electrical Measurements, for all the great fun on various conferences.

All my friends outside the lab. Anna-Karin, Louise, Cecilia and Magdalena, I value our friendship very much.

All my other friends, relatives and supporters.

Finally, I want to acknowledge my caring family; Anders, Margareta, Björn, Ellica and Louis, for your great support. This book is dedicated to you for all the joy and happiness you bring into my life.

Amelie Plymoth, Lund, November 7, 2006
References


