Protein Expression Patterns Associated with Progression of Chronic Obstructive Pulmonary Disease in Bronchoalveolar Lavage of Smokers

Amelie Plymoth,1* Claes-Göran Löfdahl,1 Ann Ekberg-Jansson,2 Magnus Dahlbäck,3 Per Broberg,4 Martyn Foster,5 Thomas E. Fehniger,4 and György Marko-Varga4

Background: We modeled the expression of proteins in baseline bronchoalveolar lavage (BAL) samples from asymptomatic 60-year-old lifelong current smokers or healthy never-smokers, who were reevaluated after 6 to 7 years to record clinical outcome.

Methods: Applying a technology toolbox consisting of replicate 2-dimensional gel separations, image annotation, and mass spectrometry identification, we catalogued a global set of proteins that were differentially expressed in individuals by presence, absence, and intensity scores.

Results: By use of multivariate analysis, we selected a subset of proteins that accurately separated smokers from never-smokers based on composite scoring. Follow-up after 6 to 7 years identified a group of individuals who had progressed to chronic obstructive pulmonary disease (COPD), Global Initiative for Chronic Obstructive Lung Disease stage 2. The baseline BAL samples of these eventual COPD patients shared a distinct protein expression profile that could be identified using partial least-squares discriminant analysis. This pattern was not observed in BAL samples of asymptomatic smokers free of COPD at 6- to 7-year follow-up.

Conclusions: Our model suggests 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, before disease is clinically evident.

© 2007 American Association for Clinical Chemistry

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 4000 substances, many of which have a toxic effect on cells (1). Within the respiratory tract, in addition to the xenobiotic effects associated with oncogene expression and malignant transformation, smoking drives a chronic inflammatory state resulting in the release of mediators such as proteases that cause the destruction of ground matrix and tissue components (2).

Little information is available 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 nonsmokers. To date, 2-dimensional polyacrylamide gel electrophoresis (2-D PAGE) and liquid chromatography platforms have been used to study clinical samples including bronchoalveolar lavage (BAL), nasal lavage, and whole lung tissue (3–6). These studies have shown that smokers typically show increased concentrations of inflammatory and redox proteins, as well as decreases in surfactants and Clara cell secretory proteins, compared with nonsmokers. These observations are in line with previous studies showing that gene expression patterns

1 Department of Respiratory Medicine and Allergology, Lund University Hospital, Lund.
2 Sahlgrenska University Hospital, Gothenburg, Sweden.
3 Discovery Medicine and Epi.
4 Department of Biological Sciences, AstraZeneca R&D, Lund, Sweden.
5 Department of Safety Assessment, AstraZeneca R&D, Loughborough, United Kingdom.

* Address correspondence to this author at: Fax +46-46-146793; e-mail amelie.plymoth@med.lu.se.
Received July 7, 2006; accepted January 18, 2007.
Previously published online at DOI: 10.1373/clinchem.2006.076075

© 2007 by The American Association for Clinical Chemistry
are permanently altered in bronchial epithelial cells and in the lung tissue of smokers who develop COPD (7, 8). In a recent study of BAL using liquid chromatography coupled with on-line linear ion trap quadrupole mass spectrometry (MS), we identified 481 high-to-low abundant proteins (6). These proteins were differentially expressed in BAL samples obtained from age-matched smokers compared with never-smokers.

In this study, we further characterized the global proteome of the BAL samples from the smokers and never-smokers studied above.

**Materials and Methods**

**STUDY POPULATION**

We recruited 60-year-old men from the randomized population study “Men born in Göteborg in 1933” [see Fig. 1 in the Data Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol53/issue4 (9–11), a subset of the study population of 879 individuals volunteered to undergo fiber-optic bronchoscopy at the age of 60. This group included 47 men: 29 asymptomatic chronic smokers (14 light and 15 heavy smokers); and 18 healthy never-smokers. At the time of BAL sampling, all men in the study were evaluated clinically as being healthy, as they had not sought medical attention for respiratory disease. Total lung capacity values did not differ significantly between smokers and never-smokers (Fig. 1); however, there was a trend for decreased lung function values in both the light and heavy smokers according to forced expiratory volume in 1 s (FEV1) and carbon monoxide lung diffusion capacity (DLCO).

At 6 to 7 years after the original BAL sampling, the study cohort was asked to participate in a follow-up. Seven of the 29 smokers had developed moderate COPD [Global Initiative for Chronic Obstructive Lung Disease (GOLD) stage 2] (12, 13). None of the never-smokers had developed either mild or moderate COPD. Three participants had died of nonrespiratory complications. As shown in Table 1, the patients who developed COPD had significant airway obstruction as seen in the decline in FEV1/forced vital capacity quotients and FEV1 (% predicted) scores, compared with the other smokers.

**BAL SAMPLING**

The studies were approved by the ethics committees of the Sahlgrenska Hospital (Gbg M 117-01) and Lund University Hospital (LU 689–01) in Sweden. Participants provided informed consent. BAL sampling was carried out under typical conditions (10), and the BAL fluid was immediately transported on ice to the laboratory for processing and storage at −80 °C (10). The protein concentrations in the recovered BAL were measured by use of Coomassie plus protein assay reagent (Pierce) with bovine serum albumin as a reference (14). The total protein concentration of the recovered BAL samples was not significantly different between individuals, but recovery of BAL was lower in the smoking cohort (6).

![Fig. 1. Lung function measurements of the never-smokers and smokers from the initial study.](image)

The lung function measurements show no significant difference in total lung capacity (TLC) values between smokers and never-smokers; however, there is a trend for decreased lung function values in both the light and heavy smokers in the measurements of FEV1 and DLco.

The definition of light smokers was 1–15 cigarettes smoked/day with a median number of 22 pack-years (range 9–45); heavy smokers, ≥15 cigarettes smoked/day with a median number of 45 pack-years (range 31–79). Pack-years are calculated by multiplying the number of packs of cigarettes smoked per day by the number of years the person has smoked.

![Table 1. Lung function measurements of the study participants.](table)

<table>
<thead>
<tr>
<th>COPD patient no.</th>
<th>Visit 1</th>
<th>Visit 2</th>
<th>Visit 1</th>
<th>Visit 2</th>
<th>Δ*</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>89</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>−11.0</td>
</tr>
<tr>
<td>Mean</td>
<td>65</td>
<td>56</td>
<td>82</td>
<td>67</td>
<td>−15.1</td>
</tr>
</tbody>
</table>

Asymptomatic smokers

<table>
<thead>
<tr>
<th>COPD patient no.</th>
<th>Visit 1</th>
<th>Visit 2</th>
<th>Visit 1</th>
<th>Visit 2</th>
<th>Δ*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>76</td>
<td>73</td>
<td>98</td>
<td>94</td>
<td>−3.4</td>
</tr>
</tbody>
</table>

The initial study (visit 1) took place during February 1994 to July 1995. The follow-up examinations (visit 2) took place in 2000 and 2001. Seven of the smokers developed moderate COPD and 13 of the smokers remained asymptomatic at follow-up. * Numerical difference: visit 2–visit 1. FVC, forced vital capacity.
Tissue Sampling and Histology
We took peripheral bronchial biopsy specimens from each of the groups with an alligator forceps from the main carina between the right and left lung as has been described (15). The sections were stained using conventional hematoxylin and eosin staining.

2-D Gel Electrophoresis
We generated 2-D PAGE protein maps according to our previously developed procedure (16). BAL fluid containing 100 μg protein was desalted by ultramembrane filtration and loaded using a pH 4–7 immobilized pH gradient strip. We performed quantitative analysis with the image software PDQuest™ version 6.0 (Bio-Rad) (16, 17).

Reference gels for each group were synthesized by hierarchic clustering of images from the individual member gels of each group and combined to create a synthetic master gel of all protein spots. For further description, see text in the online Data Supplement.

Standard Spot Number Presence Rates
To associate standard spot number (SSP) attributes between the never-smoking and smoking groups, we first filtered the global protein set to exclude proteins that were not frequently expressed by either group. We scored each gel image for each SSP of the global set of 944 identifications in terms of presence, absence, and aggregate pixel density.

Using a stringency cutoff demanding at least 30% presence of a given SSP in any single group, we analyzed 818 SSP identifications in detail. We calculated mean scores of SSP presence/absence and aggregate density for each group. 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 cohort.

Statistical Calculations
We performed principal components analysis (PCA) and partial least-squares discriminant analysis (PLS-DA) with Simca-P version 10.0.4 (Umetrics AB). PCA was used for unsupervised views of the data. PLS-DA was used to build classifiers of clinical category (18). The 2 methods provide 1 measure, R2, of model fit, and another, Q2, of the predictability of the model.

In the hierarchic clustering method, data were normalized by subtracting the mean intensity and dividing by the SD separately for each SSP. For further description, see text in the online Data Supplement.

Results

Strategy for Sample Standardization and Clinical Characterization
To identify and model individual variations in the protein expression profiles present in complex clinical samples such as bronchoalveolar lavage, we adopted a strategy to (a) reduce variability in sample selection, collection, and preparation and (b) address the range of biological variations in clinical presentation for comparison to protein expression patterns.

We were able to evaluate approximately two-thirds of the bronchial tissue biopsies. Two prominent histological features, the condition of the epithelium and the degree of inflammation, distinguished the mucosal compartment of never-smokers from smokers. As an example, in comparison to never-smokers [see Fig. 2A in the online Data Supplement], the bronchial epithelium of smokers as a group showed a broad phenotypic variation ranging from hyperplasia, metaplasia, and hypertrophy [see Fig. 2B in the online Data Supplement] to sloughing and regeneration [see Fig. 2C in the online Data Supplement]. Current smokers generally showed increased signs of diapedesis and inflammation throughout the submucosa [see Fig. 2D in the online Data Supplement] and the elastic lamina [see Fig. 2C in the online Data Supplement], as reported (15).

Both lung function measurements and histology evaluations therefore indicated that the range of features observed in the current-smoker groups were in variance with the never-smokers. This suggests ongoing episodes of pathogenesis despite the lack of clinical complaints or symptoms.

Global Protein Expression Maps and Annotations
We measured global protein expression patterns in BAL samples separated by 2-D PAGE. In-group comparisons, distributions of BAL proteins observed on raw gels from never-smokers (Fig. 2A, B) varied qualitatively and quantitatively from protein patterns observed for light smokers (Fig. 2C, D) and heavy smokers (Fig. 2E, F). Both the never-smoking group and the smoking groups expressed proteins in their BAL samples that were absent or not frequently observed in the respective group gels. We synthesized reference gels for each group (Fig. 2G never-smokers, Fig. 2H heavy smokers) by hierarchic clustering of images from the individual member gels of each group. All individual gels were combined to create a synthetic master gel of all protein spots detected (Fig. 2I). The synthetic master gel of all 47 samples contained a global set of 944 SSP identifications. The zoomed image of a representative gel region [see Fig. 3A in the online Data Supplement], displaying medium-to-low abundant protein spot expression levels, reveals a high degree of annotation differences between smokers [see Fig. 3, B and C in the online Data Supplement] and never-smokers [see Fig. 3, D and E in the online Data Supplement]. Software rendering of pixel density (z-axis) into 3-D image profiles improves visualization of the relative presence/absence.

We constructed a model to compare differences between the smoking and never-smoking SSP expression patterns, applying differing set points that spanned be-
Several questions were posed: What proportion of SSP identities is regulated in smokers compared with never-smokers? What is the relationship between presence rate and fold expression factor? Are certain sets of SSPs associated with smoking?

Fig. 3A displays a distribution plot that maps SSP count (y-axis), relative fold differences in individual SSP expression levels between smokers and never-smokers (x-axis), and individual SSP distribution at various points of presence between 30% (n = 818 SSPs) and ≥90% (n = 189 SSPs) (z-axis). The fold-change interval in the model was -20 ≤ X ≤ 20, where X < -20 is set to -20 and X > 20 is set to 20 to capture the observed variances in mean SSP expression in both groups. Very few SSP identities (n = 69) were present in 100% of individuals in either of the groups.

By use of the model, we could easily define modal distributions of protein expression at the level of individual SSPs. This enabled us to categorize expressions as up-regulated, down-regulated, or not regulated when we compared smoking groups to the never-smoking group. Irrespective of the presence rate, ~85% of the SSP annotation identities showed similar expression patterns in both never-smokers and smokers. The expression densities decreased to between -4- and +4-fold differences of the mean density ratios for smokers compared with never-smokers (Fig. 3B). Approximately 2% of the SSP identities were down-regulated in smokers (density ratio < -4). Among the remaining 13% protein identities, we observed a bimodal, highly induced expression pattern in the smoking group. This relative trend was maintained over the interval of low-to-high presence rates.

**Statistical Distributions and Separations of SSP Profiles**

We found that each individual expressed a partial subset of the global SSP set, in part because of the intrinsic biological variations observed between groups as described by our independent measurements, such as histological examination of bronchial biopsies (see Results). Conventional nonparametric statistical comparisons of the stringency cutoff set (70%, 406 SSPs), using parameter presence, absence, and abundance scoring, showed that significant differences in protein expression patterns occurred between never-smokers and smokers (Mann-Whitney P < 0.001). Separation of heavy smokers from
light smokers based on individual SSP scores required the application of further statistical tools.

We applied PCA and PLS-DA to the SSP dataset parameters. Using PCA analysis of 406 protein identities expressed by at least 70% of the never-smoking or smoking cohorts, we unambiguously \( P < 0.001 \) and accurately \( R^2 = 0.84 \) separated smokers from never-smokers based on composite protein expression phenotype (Fig. 4A). A progression of dimensional features obliquely separates 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, we tested at what level of predictive accuracy we could assign any given individual to a specific group designation based solely on their individual SSP profile. 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
Fig. 4. Unsupervised and supervised analysis of the BAL protein identities expressed by the never-smoking or smoking cohorts.

(A), Principal component analysis of the 406 BAL protein identities expressed by at least 70% of either the never-smoking or smoking cohorts separates the never-smokers (green) from the light smokers (blue) and the heavy smokers (red). The encircled patients were found to have developed moderate COPD in the follow-up study. (B), by reanalyzing the smokers using successive rounds of PLS-DA analysis, the Q2 scores for predicting the COPD clinical outcome rise significantly from a negative value using all 406 protein identities to a predictability of 0.8 applying 100 protein identities. (C), PLS-DA on this set of 100 SSP identities separates all 7 patients who developed moderate COPD (black) from the other smokers (red) by both T1 and T2 dimensions. The ellipse represents the 95% confidence region based on Hotelling T².
results validated that the qualifiers used as dimensions in these comparisons of SSP datasets were significantly related to a specific group character.

**CLINICAL OUTCOME AT 6 TO 7 YEARS’ FOLLOW-UP**

Seven of the 29 light and heavy smokers were found to have developed moderate COPD (GOLD stage 2, Fig. 4A, encircled) in the follow-up study. None of the never-smokers developed either mild or moderate COPD. Using the set of 406 SSPs expressed at a 70% presence rate, we were unable to separate the protein expression profiles of the 7 COPD patients from the other 22 heavy or light smokers using the Q2 predictability measure ($R^2 < 0.8$). We then tested whether a subset of the 406 SSPs might be more significantly associated with clinical outcome by reanalyzing the 29 smokers by successive rounds of PLS-DA analysis, utilizing 200, 100, 50, 25, or 10 SSP components. As shown in Fig. 4B, the Q2 scores for predicting the COPD clinical outcome rise significantly from a negative value using all 406 components to a high predictability of 0.8 applying 100 SSPs. We analyzed the expression profiles of all 29 smokers using PLS-DA analysis on this set of 100 SSP identities. We found that 7 COPD progressors could be well separated from the other smokers by both T1 and T2 dimensions (Fig. 4C).

We further analyzed the set of 100 differentially expressed SSP identities using 2-D hierarchal clustering (Fig. 5). Our application reveals order among the samples and within the SSPs. Progressing COPD patients segregated together on the left side of the dendrogram except for 1 individual (patient 855) whose clinical characteristics were not dissimilar from those of the other patients (Table 1).

**Discussion**

In this report, we present a strategy for relating the measurement of sets of proteins within a clinical sample with phenotypes of clinical presentation. To accomplish this, we used an interdisciplinary toolbox that included protein separation, identification platforms, and statistical methods for multivariate analysis. These methods allowed us to group the men in the study 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 clinical material that could be normalized for study, in terms of being associated with the biology of interest and also for sampling variability. Our study material is unique: age-matched men all born in 1933, living in 1 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 quantification and normalization methods in support of our need for an unbiased assessment. We have chosen to analyze BAL because of its utility to address secreted and extracellular proteins present within the central and descending airways of smokers and never-smokers. The technology we adopted has 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.

Using standardized 2-D gel technology, we observed that the protein components present in BAL are variably expressed in individuals. The set of SSP identities found in each BAL sample was quantitatively scored to establish hierarchal relationships between different SSPs for each of the 2 groups. Multivariate analysis provided us with the means to relate not only singular SSP characteristics such as presence and relative expression level values in comparison to group behavior, but also the same parameters to all of the other SSPs in the dataset.

We then asked what independent predictors among all SSPs characterize a particular group. By using presence rate stringency rules that could be applied independently to one or another of the study groups, we allowed for extreme changes in absolute factors of expression. We found that smoking individuals lost the expression of a small subset of identities present in never-smokers and acquired a high level of expression of SSP identities absent in never-smokers. We thus demonstrated that lifelong smokers develop significantly different protein expression patterns in their respiratory tract from those of age-matched never-smokers.

Diagnostic datasets are often composed of groups of signals representing the various transition states of numerous singular proteins. Because of the broad dynamic range of protein concentrations present in any given sample volume, it is beneficial to take advantage of whatever associations of regularity exist between individual protein identities, sets of proteins within given clinical samples, or given clinical presentations. To accomplish this degree 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 clinical outcome.

Because the study material was collected in 1993, when the asymptomatic men were 60 years old (and many of the participants agreed to be clinically evaluated at further time points), we were able to pose questions relating eventual clinical outcome to the status of BAL protein expression 6–7 years earlier. Over this time period, 7 of the smokers had clinically progressed to moderate COPD (GOLD stage 2). We were unable to segregate this COPD subset from the other smokers using the PLS-DA model based on the presence/absence and intensity scores of the global set of 406 proteins expressed at the stringency rate of 70% presence. As a cross-validation for further analysis, we applied repetitive cycles of filtering on this global set to obtain a limited protein set shared by the eventual COPD patients, yet not observed by the other
lifelong smokers. In particular, the hierarchal clustering analysis of 100 identities by presence/absence and intensity scores independently clustered 6 of the 7 participants in 1 quadrant, demonstrating the commonality in protein expression for this subset of eventual COPD patients compared with the other smokers. Thus, the set of 100 SSP identities associated with COPD outcome was shown to be composed of statistically related proteins that occurred at either much lower or much higher concentrations than in the other smokers. We conclude 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 individuals sharing risk behavior for eventual disease development, such as age, smoking, and geographical location.

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 endpoint. Individual MS identities act as important an-
notation 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 use of multivariate algorithms, we showed that both study cohorts maintained a high degree of overall predictability of group association. This result is important, as it implies that stable predictive phenotypes of protein expression can be determined and used to segregate individuals even in clinical cohorts that consist of asymptomatic and clinically healthy participants. This segregation can be accomplished despite the requirement of assigning actual protein identities to each of the components within the clinical proteome.

The study was supported by the Swedish Heart Lung Foundation.

References