Large-scale label-free quantitative mapping of the sputum proteome


J. Proteome Res., Just Accepted Manuscript • DOI: 10.1021/acs.jproteome.8b00018 • Publication Date (Web): 08 May 2018

Just Accepted

"Just Accepted" manuscripts have been peer-reviewed and accepted for publication. They are posted online prior to technical editing, formatting for publication and author proofing. The American Chemical Society provides “Just Accepted” as a service to the research community to expedite the dissemination of scientific material as soon as possible after acceptance. “Just Accepted” manuscripts appear in full in PDF format accompanied by an HTML abstract. “Just Accepted” manuscripts have been fully peer reviewed, but should not be considered the official version of record. They are citable by the Digital Object Identifier (DOI®). “Just Accepted” is an optional service offered to authors. Therefore, the “Just Accepted” Web site may not include all articles that will be published in the journal. After a manuscript is technically edited and formatted, it will be removed from the “Just Accepted” Web site and published as an ASAP article. Note that technical editing may introduce minor changes to the manuscript text and/or graphics which could affect content, and all legal disclaimers and ethical guidelines that apply to the journal pertain. ACS cannot be held responsible for errors or consequences arising from the use of information contained in these “Just Accepted” manuscripts.
Large-scale label-free quantitative mapping of the sputum proteome

Dominic Burg\(^1,2\)†, James P R Schofield\(^1,2\)*, Joost Brandsma\(^2\), Doroteya Staykova\(^1\), Caterina Folisi\(^1\), Aruna Bansal\(^3\), Ben Nicholas\(^2\), Yang Xian\(^4\), Anthony Rowe\(^5\), Julie Corfield\(^6\), Susan Wilson\(^2\), Jonathan Ward\(^3\), Rene Lutter\(^7,8\), Louise Fleming\(^9\), Dominick E Shaw\(^10\), Per S Bakke\(^1\), Massimo Caruso\(^12\), Sven-Erik Dahlen\(^13\), Stephen J. Fowler\(^14\), Simone Hashimoto\(^15\), Ildikó Horváth\(^16\), Peter Howarth\(^2\), Norbert Krug\(^17\), Paolo Montuschi\(^18\), Marek Sanak\(^19\), Thomas Sandström\(^20\), Florian Singer\(^21\), Kai Sun\(^4\), Ioannis Pandis\(^8\), Charles Auffray\(^22\), Ana R Sousa\(^3\), Ian M Adcock\(^24\), Kian Fan Chung\(^9\), Peter J Sterk\(^2\), Ratko Djurkovic\(^25,8\), Paul J Skipp\(^1,8\) and the U-BIOPRED Study Group\(^25\)

\(^1\)Centre for Proteomic Research, University of Southampton, UK
\(^2\)NIHR Southampton Respiratory Biomedical research unit, University Hospital Southampton, UK
\(^3\)Acclarogen Ltd, Cambridge, UK
\(^4\)Data Science Institute, Imperial College, London, UK
\(^5\)Janssen Research & Development, Buckinghamshire, UK
\(^6\)Areteva Ltd, Nottingham, UK
\(^7\)AMC, Department of Experimental Immunology, University of Amsterdam, Amsterdam, The Netherlands
\(^8\)AMC, Department of Respiratory Medicine, University of Amsterdam, Amsterdam, The Netherlands
\(^9\)Airways Disease, National Heart and Lung Institute, Imperial College, London & Royal Brompton NIHR Biomedical Research Unit, London, United Kingdom
\(^10\)Respiratory Research Unit, University of Nottingham, UK
\(^11\)Institute of Medicine, University of Bergen, Bergen, Norway
\(^12\)Dept. of Clinical and Experimental Medicine Hospital University, University of Catania, Catania, Italy.
\(^13\)The Centre for Allergy Research, The Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden.
\(^14\)Respiratory and Allergy Research Group, University of Manchester, Manchester, UK.
\(^15\)Dept. of Respiratory Medicine, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands.
\(^16\)Dept. of Pulmonology, Semmelweis University, Budapest, Hungary
\(^17\)Fraunhofer Institute for Toxicology and Experimental Medicine Hannover, Hannover, Germany.

Faculty of Medicine, Catholic University of the Sacred Heart, Rome, Italy.

Laboratory of Molecular Biology and Clinical Genetics, Medical College, Jagiellonian University, Krakow, Poland

Dept. of Medicine, Dept of Public Health and Clinical Medicine Respiratory Medicine Unit, Umeå University, Umeå, Sweden.

University Children’s Hospital Zurich, Zurich, Switzerland.

European Institute for Systems Biology and Medicine, CNRS-ENS-UCBL-INSERM, Université de Lyon, France

Respiratory Therapeutic Unit, GSK, Stockley Park, UK.

Cell and Molecular Biology Group, Airways Disease Section, National Heart and Lung Institute, Imperial College London, Dovehouse Street, London, UK

A full list of the U-BIOPRED Study Group members and their affiliations can be found in the acknowledgements.

*To whom correspondence should be addressed: James P R Schofield at J.P.R.Schofield@soton.ac.uk, +44 (0) 23 80594204

†Equally contributing joint first authors

# Joint senior authors

\(^\) Joint senior authors
Abbreviations:

U-BIOPRED - Unbiased BIOmarkers Predictive of RESpiratory Disease outcomes
HDMSE – High definition mass spectrometry
COPD – Chronic Obstructive Pulmonary Disease
CE - collision energy
BAL - bronchoalveolar lavage

Abstract: Analysis of induced sputum supernatant is a minimally invasive approach to study the epithelial lining fluid and, thereby, provide insight into normal lung biology and the pathobiology of lung diseases. We present here a novel proteomics approach to sputum analysis developed within the U-BIOPRED (Unbiased BIOmarkers Predictive of RESpiratory Disease outcomes) international project. We present practical and analytical techniques to optimise the detection of robust biomarkers in proteomic studies. The normal sputum proteome was derived using data-independent HDMS applied to 40 healthy non-smoking participants, which provides an essential baseline from which to compare modulation of protein expression in respiratory diseases. The “core” sputum proteome (proteins detected in ≥40 % of participants) was composed of 284 proteins and the extended proteome (proteins detected in ≥3 participants) contained 666 proteins. Quality control procedures were developed to optimise the accuracy and consistency of measurement of sputum proteins and analyse the distribution of sputum proteins in the healthy population. The analysis showed that quantitation of proteins by HDMS is influenced by several factors, with some proteins being measured in all participants’ samples and with low measurement variance between samples from the same patient. The measurement of some proteins is highly variable between repeat analyses, susceptible to sample processing effects, or difficult to accurately quantify by mass spectrometry. Other proteins show high inter-individual variance. We also highlight that the sputum proteome of healthy individuals is related to sputum neutrophil levels, but not gender or allergic sensitisation. We illustrate the importance of design and interpretation of disease biomarker studies considering such protein population and technical measurement variance.

Key words: asthma, proteomics, biomarkers, U-BIOPRED, sputum, HDMS, unbiased, variance, allergic, neutrophil

INTRODUCTION

Sputum induction is a widely applied method of sampling the epithelial lining fluid that lines the lower airways constituting the tracheobronchial tree. It enables assessment of both the cellular and extracellular environments in the lung and is particularly useful in the study of inflammatory respiratory diseases, providing insight into the immune and structural cell populations and their secreted products. Initial studies of induced sputum focused on inflammatory cell counts and targeted quantification of soluble proteins by enzyme-linked immunosorbent assay (ELISA). Such analyses identified several induced sputum biomarkers as valuable in the description of inflammation in common chronic airway diseases including asthma and chronic obstructive pulmonary disease (COPD), providing insight into determinants of disease severity and relevant pathophysiological abnormalities, such as airway hyperresponsiveness and changes in airway geometry. Combined with studies of cell function such as chemotactic activity; quantification of cytokines and chemokines in the sputum fluid phase has provided a better understanding of the extent to which individual mediators contribute to inflammation, thereby providing initial stratification of respiratory disease.

Methods for global, unbiased analysis, that do not select a priori which analytes are measured, including transcriptomics, proteomics and lipidomics, appear useful for stratifying disease. However, only a limited number of unbiased proteomic studies focusing on the lungs have been published to date. Ten years ago, we described the first sputum proteome, applying a shotgun method to an induced sputum sample from a female smoker with no detectable evidence of lung disease. Since then, there have been a number of reports of this approach in COPD including a study highlighting the utility of protein network analysis in sputum, and a large study combining proteomic and transcriptomic analyses. Likewise, limited studies of sputum have been performed to study asthma and some have been relatively low throughput. Apart from one study, by Titz et al., coverage of the sputum proteome remains low. Our previous study highlighted some overlap between sputum proteomes and proteomes of other sample types, namely bronchoalveolar lavage (BAL) and saliva. In recent years, attention has been drawn to the repeated failure of published biomarkers to translate to the clinic. Such failure is often attributable to study design and validation, insufficient sample size and inappropriate experimental methodology. Problems with sample size are beginning to be addressed in proteomics, particularly with the advent of data independent approaches such as MS which allow absolute comparison of samples without the inherent limitations associated with multiplexing, labelling or
spectral counting-based quantitation. Despite the utility of MS for large clinical studies, there is very little information on the effect of measuring samples over extended periods and resulting data variability. The approach to quality control in the analysis of human BAL samples using repeated measurements and pooled samples published by the Moseley group at Duke University is a standout example of the necessary approach required for clinical studies; however, sample sizes in these studies have been relatively small.

In the current study, we have applied state of the art quantitative HDMS analysis to a large set of sputum samples to advance on the sputum proteome previously reported. As part of the method evaluation, we explored the impact of granulocytic infiltration of the airways, participant gender and other common demographics on the sputum proteome. Given the high prevalence of allergic sensitisation to common airborne allergens (e.g. house dust mite and pollens) in the general population, we also examined how atopy, defined by sensitisation to at least one common aero-allergen, affects the sputum proteome. As a key component of the study, we assessed variability in proteomic measurements and considered the impact of such variability on biomarker discovery. Using repeated measurements, pooled samples, comparison between individuals and to serum samples from the same study participants, we assessed the likely source of variability in measurements on a protein by protein basis. We discuss the impact of variability on effective sample size and statistical power for comparative studies. Finally, we have performed an in-depth analysis of tissue and cellular origins of proteins from previous proteomic studies and defined the accessible functional proteomic space using functional enrichment analysis.

MATERIALS AND METHODS

Study design and participant characteristics

The U-BIOPRED study was performed in 14 European clinical centres with extensive experience in sputum induction and processing. The clinical study has been described elsewhere and the protocol was approved by all local Ethics Review Boards. Participants gave their written informed consent for extensive characterisation using routine clinical protocols, including, lung function tests, assessment of sensitization to common aero-allergens, and haematological and biochemistry blood tests (reported in detail in Shaw et al., 2015). Samples were stored in a central biobank (CIGMR Biobank, University of Manchester) where they were blinded. Identity of the samples were un-blinded only after all the mass spectrometric analyses and data pre-processing had been completed.

Table 1 Demographics and sputum cell characteristics

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (mean ± SEM)</td>
<td>40</td>
<td>35.4 ± 2.3</td>
</tr>
<tr>
<td>Weight (kg) (mean ± SEM)</td>
<td></td>
<td>85 ± 2.14</td>
</tr>
<tr>
<td>BMI (mean ± SEM)</td>
<td></td>
<td>25.62 ± 0.51</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td>29/11</td>
</tr>
<tr>
<td>Race (% white Caucasian)</td>
<td></td>
<td>90%</td>
</tr>
<tr>
<td>Smoking history</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ex smokers (count, % of total)</td>
<td></td>
<td>5 (12.5%)</td>
</tr>
<tr>
<td>Atopy positive (count, % of total)</td>
<td></td>
<td>13 (32.5%)</td>
</tr>
<tr>
<td>Positive IgE Assay (count, % of total)</td>
<td></td>
<td>7 (17.5%)</td>
</tr>
<tr>
<td>Positive Skin Prick Assay (count, % of total)</td>
<td></td>
<td>12 (30.0%)</td>
</tr>
<tr>
<td>FEV1/FVC predicted (% (mean ± SEM)</td>
<td></td>
<td>83.58 ± 0.38</td>
</tr>
<tr>
<td>Sputum % Neutrophil (median, range)</td>
<td></td>
<td>37.95 (2.71 - 88.34)</td>
</tr>
<tr>
<td>Sputum % Eosinophil (median, range)</td>
<td></td>
<td>0.00 (0.00 - 2.57)</td>
</tr>
<tr>
<td>Sputum % Lymphocytes (median, range)</td>
<td></td>
<td>1.22 (0.00 - 7.76)</td>
</tr>
<tr>
<td>Sputum % Macrophages (median, range)</td>
<td></td>
<td>60.30 (7.11 - 96.10)</td>
</tr>
<tr>
<td>Sputum % Squamous epithelial (median, range)</td>
<td></td>
<td>14.70 (0.00 - 39.20)</td>
</tr>
</tbody>
</table>

Forty healthy individuals (mean age 36.0 years, range 18 to 65, 70% male), provided sputum samples considered representative of the bronchial compartment, i.e. ≤ 40% contaminating squamous cells (Table 1). The frequency of atopy, demonstrated by positive skin or serum IgE specific for at least one common aero-allergen test, was 32.5%. The atopic and non-atopic participants did not differ in respect of sputum cell counts, including sputum eosinophils (<2% of total inflammatory cells in all participants), blood eosinophils, and lung function. As expected, total serum IgE concentrations were higher in atopic individuals (p<0.05), but, surprisingly, serum LDH was also slightly higher in the atopic participants (p=0.01) while alkaline phosphatase was lower (p=0.05).

Sputum induction and processing

Sputum induction with nebulised hypertonic saline (4.5% NaCl) and sample processing were performed in accordance with the recommendations of the European Respiratory Society Task Force on induced sputum methods. Uniformity of methods was ensured by all study sites using standard operating procedures (SOP) and centralised training. For consistency required for comparison with proteomes in patients with disease (asthma or COPD), all
participants were pre-medicated with the β2-agonist, salbutamol, given as standard, to prevent excessive bronchoconstriction in patients with airways disease.

Induced Mucoid portions of the induced sputum were selected with forceps to reduce salivary contamination, weighed and solubilised at room temperature with 6.8mM dithioerythritol (DTE) in HEPES buffered saline, added at a 4:1 w/v ratio. The solution was filtered through a 100-µm filter, centrifuged at 400 x g to remove the cell pellet, further centrifuged at 12,000 x g to remove cell debris, both at 4°C, and stored at -80°C. The cell pellets were processed for quantification of alive or dead respiratory cells, squamous cells and differential inflammatory cell counts (by Diff-Quick rapid Romanowsky stain); eosinophil, neutrophil, macrophage/monocyte, lymphocyte and mast cell/basophil counts were reported as percentages of total inflammatory cells, while squamous cells were reported as a percentage of total cell counts.

**Protein isolation and preparation for analysis**

Sputum samples were thawed to room temperature before taking 100-µl aliquots for extraction of lipids using a semi-automated Bligh-Dyer protocol (Bligh and Dyer, 1959) on a robotic liquid handling platform (Freedom EVO 100; TECAN, Männedorf, Switzerland). Briefly, each sample was made up to a volume of 800 µl with 0.9% saline solution before adding 2 ml of methanol (MeOH) and 1 ml of dichloromethane (DCM) and 10 µl of anti-oxidant (5 mg ml⁻¹ hydroxytoluene in MeOH). Samples were centrifuged at 1,000 x g for 10 min, and collected supernatants stored at -80°C. The cell pellets were processed for quantification of alive or dead respiratory cells, squamous cells and differential inflammatory cell counts (by Diff-Quick rapid Romanowsky stain); eosinophil, neutrophil, macrophage/monocyte, lymphocyte and mast cell/basophil counts were reported as percentages of total inflammatory cells, while squamous cells were reported as a percentage of total cell counts.

**Serum collection and processing**

Clotted venous blood samples were centrifuged at 1,000 x g for 10 min, and collected supernatants stored at -80°C. A pooled serum sample was created as for sputum. In order to increase the number of identifications, the 12 most abundant proteins were immunodepleted using disposable agarose columns (Pierce / Thermo-Fisher) and eluates reduced, alkylated, digested and lyophilised. Peptide extracts were then re-suspended in 3% ACN, 0.1% TFA and desalted using 96 well RP solid phase extraction plates (3M Empore). Eluates were transferred to separate microcentrifuge tubes, lyophilised and stored on ice until mass spectrometry.

**Mass spectrometry**

Peptide extracts were re-suspended in buffer A, (3% ACN, 0.1% Formic acid (v/v) and the concentration measured using a Direct Detect System (Millipore). An internal standard mixture of E. coli ClpB H3 standard (Waters), yeast enolase (ENO) and yeast alcohol dehydrogenase (ADH) was added to a final concentration in 20µl of 250ng/µl sputum peptide, 12.5 fmol/µl ClpB, 12.5 fmol/µl ENO, and 8.75 fmol/µl ADH (serum was 25% more concentrated).

Samples were analysed in duplicate, sequentially (not spread across batches), via HDMS® on a Waters Synapt G2S high definition mass spectrometer coupled to a nanoAcquity UPLC system. 4µl of peptide extract was injected onto a C18 BEH trapping column (Waters) and washed with buffer A for 5 min at 5µl/min. Peptides were separated using a 25cm T3 HSS C18 analytical column (Waters) with a linear gradient of 3-50% ACN + 0.1% formic acid over 50 minutes at a flow rate of 0.3µl/min. Eluted samples were sprayed directly into the mass spectrometer operating in MS² mode. Data were acquired from 50 to 2000 m/z with the quadrupole in RF mode using alternate low and elevated collision energy (CE) scans, resolution of 35,000. Low CE was 5V and elevated CE ramp from 15 to 40V. Ion mobility separation was implemented prior to fragmentation using a wave velocity of 650 m/s and wave height of 40V. The lock mass Glu-fibrinopeptide, (M+2H)⁺, m/z = 785.8426 was infused at a concentration of 100 fmol/µl at a flow rate of 250 nl/min and acquired every 60 sec.

**Database searching and curation**

Raw data were processed using a custom package (Regression tester) based upon executable files from ProteinLynx Global Server 3.0 (Waters). The optimal setting for peak detection across the dataset was determined using Threshold inspector (Waters) and these thresholds were applied: low energy = 100 counts; high energy = 30 (for serum this was set to 25) and a total energy count threshold of 750. Database searches were performed using regression tester and searched against the Uniprot human reference database (20/11/2014; 20,229 entries) with added sequence information for internal standards. A maximum of two missed cleavages was allowed for tryptic digestion and the variable modification was set to contain oxidation of methionine and carboxyamidomethylation of cysteine. Precursor and product ion mass tolerances were calculated
automatically during data processing and the false discovery rate (FDR) was set at 4%. We report only proteins identified in at least two patient samples, which results in a FDR below 1%. Only proteins identified in each technical replicate of at least two patient samples were considered; thus, the false positive rate is minimized, since chemical noise is random in nature and does not replicate across injections. Quantity was estimated in absolute amounts using the Top 3 method. The ion accounting output files were compiled and summary information generated from search log files using custom Python scripts. Information contained in ion accounting files were collated into a single .csv document using a custom Python script.

Data filtering and normalisation

Protein identifications collated from the ion accounting files were further quality filtered by allowing only identifications with the following criteria: identification in at least two separate samples (not including replicate injections), a process that required at least three high quality unmodified peptides using the Top 3 method, and 2 peptides with at least 4 fragment ions for each protein. All other protein identities were removed. Proteins were first ranked according to coverage across the samples, and then each protein entry was ranked according to the order in which they were run. QC information was added for each sample (batch information, protein concentration, ion counts). First, differences in run-to-run intensity (loading) were adjusted by normalising each run to the sum of top 3 intensities of the proteins up to the point where the sample set reached 10% missing data (we refer to this as 'top-90 normalisation'). ComBat was used to adjust for batch to batch variation.

InforSense software (ID Business Solutions, Guildford, UK) was applied to generated heat maps for the top 150 proteins using both 'top 3 peptide intensity sum' (a proxy for concentration) and peptide concentrations (expressed in fmol) on column calculated from internal standards. Sample-wise correlation plots were created using Inferno RDN (http://omics.pnl.gov/software/infenordn). Heat maps and correlation plots were inspected for poor samples or injections; those with very low or no ID's and/or poor correlation were removed from the dataset.

Samples were analysed in duplicate and the average intensity values used for analysis. For the purpose of quality control, several analyses were performed. Replicate injections were inspected for consistency in quantitation. To achieve this, an average of the two injections 'top 3 peptide intensity sum' was used and a distance matrix calculated by taking the Euclidian distance between the two injections as a function of the average of the injections. These values were visualised in a heat map, enabling rapid inspection of duplicates with high variance, which likely indicated a technical issue between injections (e.g. sprayer dropout, or failure to inject the correct volume). Data were corrected by applying the following universal rule (Rule 1): for samples with >2-fold between-injection difference in average intensity of all proteins: "report injection one intensity values for all proteins, unless a specific protein was only quantified in injection two, then include this value for increasing coverage". Injection 1 was selected for consistency as it is not possible to distinguish which run more accurately represents the true abundance.

While the above method was useful in identifying whole samples with poor repeatability between injections, there were cases where the concentrations of individual proteins were highly variable. To assess these cases, a log was created using a custom script, which highlighted those proteins where the ratio between injections was >1.5. Proteins with high frequency of poor measurement stability across all samples were processed according to 'Rule 2: "if the variation between injections is greater than 1.5-fold, take the quantity measured using injection one". The fraction of samples where the ratio between runs had to be >1.5 was 0.5 to apply rule 2. This rule was only applied to 11 proteins in the extended proteome and 0 proteins in the core proteome. This consistent approach to dealing with large variation in between-run protein measurements was useful in reducing technical variation in the dataset, whilst minimising reductions in proteome coverage. However, we recognise that there remains increased uncertainty in the measurement of proteins treated in this way, and the issue may be minimised in future studies by increased replicate measurements. Mean values were derived from replicate sample injections except for those cases where rule 1 and rule 2 were applied, and those cases where the protein was quantified in only one sample.

Data retrieval and conversion

Data from previous studies of relevant tissues were retrieved from the following sources: the Protein Atlas (www.proteinatlas.org), the HUPO plasma reference set, the reference plasma dataset from the laboratory of Matthias Mann, proteomes of BAL fluid, sputum proteomes, proteomes of BAL fluid, exhaled breath condensate, pure airway mucus, saliva, macrophage proteomes, eosinophil proteomes, whole neutrophils, investigations of neutrophil extracellular traps, neutrophil microparticles, and neutrophil granules. Data were tabulated and identifiers converted to Uniprot format using the Uniprot ID mapping service (http://www.uniprot.org/mapping/) or via DAVID. Redundant, discontinued, merged and incomplete entries (e.g. assignments to protein fragments, pseudogenes, or to non-human proteomes) were either disregarded or were assigned to Uniprot identifiers.

Statistical analyses and informatics
Statistical analyses were performed in R\textsuperscript{86}, Inferno RDN\textsuperscript{44}, Microsoft Excel and Minitab\textsuperscript{13}, using parametric or non-parametric tests as appropriate. Coefficient of variation was calculated for log-normal distributions. Visualisation was performed in Origin 9.1 (http://www.originlab.com/9), R, and InforSense (IDBS). Venn diagrams were generated using Venny (http://bioinfgp.cnb.csic.es/tools/venny/\textsuperscript{59}). Tree maps were drawn using Treemap 4.1.2 (http://www.cs.umd.edu/hcil/treemap/\textsuperscript{69}). Pathway analysis, functional enrichment analysis and biological inference were performed using Ingenuity Pathway Analysis software (IPA\textsuperscript{4}, QIAGEN Redwood City). The IPA analyses were performed against a background gene set restricted to Ingenuity Knowledge Base genes from expected sources of sputum proteins, tissues and cell types of the lung or near the lung in homo sapiens. Functional enrichment was calculated via FunRich \textsuperscript{83}, and the Secreted Protein Database \textsuperscript{84} was used to identify secreted proteins. Protein-protein interactions were explored using String (http://string-db.org/\textsuperscript{89}). Ontology was annotated from retrieval via UniProt\textsuperscript{94}, and retrieval and enrichment analysis was performed via GOrilla (http://cbl-gorilla.cs.technion.ac.il/\textsuperscript{93}) against the homo sapiens proteome and REVIGO (http://revigo.ibb.cnr.it/\textsuperscript{97}). The data were clustered by topological data analysis (TDA) \textsuperscript{88–93}. TDA provides geometric representation of the relationships between patient data and variables in high-dimensional data sets. TDA structures were generated using the Ayasdi Cure application (Ayasdi, Menlo Park, CA) with a norm correlation metric and two MDS lenses (resolution, 20 bins; gain, ×5.0; equalized).

Power calculations were performed using R-package size\textsuperscript{80}, assuming two equally sized groups, to achieve 80% power to detect a given fold change (FC), at the FDR-adjusted 5% significance threshold, using a two-sided, two-sample t-test, assuming that the percentage of true null hypotheses, \(p_0\), is 95% and 97.5% respectively. Equal variances were assumed for cases and controls. The open source software, variancPartition, was used to identify the drivers of protein measurement variation\textsuperscript{94}.

**RESULTS**

**Identified proteins**

A total of 4,182 proteins were identified in the sputum in ≥1 individual(s); 2,354 proteins in ≥2 individual(s), 284 proteins in ≥40% and 73 in ≥90% of individuals (Supplementary Figure 1). High abundance proteins were generally more frequently identified (Fig. 1), but many high abundance proteins were also identified at lower frequency. These abundant proteins would be expected to be observed across multiple time points in the same patient, although there may be a proportion that are not replicated because of biological variation or where they are near the limits of detection. With consideration for these effects, we have defined the sputum proteome in two ways, the ‘core’ and ‘extended’ sputum proteome (supplementary excel file 1). The 284 proteins identified in ≥40% of participants were defined as the ‘core sputum proteome’ and were used in the statistical analysis. The ‘core’ proteome represents the most commonly detected proteins within the sputum samples. The cut-off was defined at ≥40%, since at this frequency of identification, the frequency vs. protein rank curve was close to the point of inflection (Fig. 1), where even a slight increase in the frequency of identification ‘cut off’, significantly increased the sparsity of the dataset and, hence, the total number of missing values. We also defined an ‘extended healthy sputum proteome’ dataset consisting of 1,666 proteins identified in ≥3 individuals.

![Figure 1](image-url)  
**Figure 1.** Defining the ‘core’ sputum proteome from the relationship of abundance of identified proteins and frequency of identification across samples. A) illustrates the 40% frequency of detection used as the cut-off for the ‘core’ sputum proteome. The relationship between rank of frequency of detection and the number of proteins identified (healthy sputum protein rank) is approximately linear between 100% and 40%. This is similar to the relationship between protein abundance and rank of frequency of detection. Red lines indicate 40% cut off points for the 284 protein ‘core sputum proteome’ and also illustrate that this level is close to the point of inflection of the curve. The ‘core sputum proteome’ is shaded pink. At this point, increasing the coverage cut off point for analysis significantly increases the sparsity of the dataset and hence the total number of missing values. B) The intensity of protein measurement correlated weakly with the frequency of protein identification.

**Impact of gender, age, atopy and granulocyte counts**

No significant differences in proteomes were observed when comparing age, atopic and non-atopic individuals or males and females (Fig. S2). Furthermore, using an FDR-adjusted t-test, only CLIP-
associating protein 1, CLASP1_HUMAN, was found to be significantly different (q=0.04) between males and females. No proteins were significantly different between atopic and non-atopic individuals (q < 0.05).

The network shown in Fig. S2 is constructed using multidimensional scaling (MDS) lenses (similarity metric) projected onto a TDA network, representing the structure of the proteomic data. This is an advanced technique for clustering data according to similarity and was used to explore the shape of the data for impacts of potential covariates. There was a large range of sputum neutrophil counts, and a small number of individuals had counts (>80%) that would be classified as neutrophilia. Compared to the other participants, these individuals had elevated levels of Neutrophil Defensin (Mann Whitney; q=0.02) and borderline results (q=0.06) for neutrophil-associated proteins: leukocyte elastase inhibitor (Serpin B1), MMP9, and S100A8/9, and RHO protein GDP dissociation inhibitor. There was also a weak, but statistically significant, positive correlation between some of the major granule proteins and neutrophil counts as a % of total inflammatory cells (4 of the top 5 proteins with greatest R² correlation scores are shown in Fig. 2). The average R² for correlation of proteins with neutrophil count was 0.07 and 91% of proteins had an R² < 0.2.

![Figure 2. Neutrophil proteins and neutrophil counts. We observed relationships between neutrophil granule proteins and the proportion of neutrophils in sputum. Neutrophil proteins correlated with percentage sputum neutrophil cell counts.](image)

**Protein variability, intensity adjustment and measurement accuracy**

Protein measurements between samples were visualised in heat-maps before and after intensity adjustment and after batch effect correction (Fig. S3), allowing rapid assessment of fluctuations in instrument performance and systematic variation, e.g. sensitivity and column changes over time and between analytical batches. Effects that can be easily visualised in non-adjusted maps (Fig. S3A) were corrected by normalising the intensity of proteins in each sample to the sum of the measured top3 intensity of the proteins for that sample, up to the level of 10% missing data across all samples; a method which we termed the ‘top90’ method (Fig. S3B). The top90 adjustment corrected for protein intensity variation more effectively than normalisation to ‘total intensity’; avoiding effects of rarely measured protein abundances. ComBat was used to adjust for batch-to-batch variation (Fig. S3C).

High sample to sample variability was observed in the sputum proteomes. Pooled sputum sample replicates, which were processed and analysed at regular intervals throughout the acquisition of healthy sputum data, were compared to the participant sputum samples and the matched healthy serum dataset (Fig. 4). Compared to individual sputum samples, pooled samples contained higher numbers of proteins with a lower percentage Coefficient of Variation (CV%). In all datasets, the variability in protein measurements increased as the protein frequency of identification decreased. While the level of variability in sputum pools was low, it was even lower in matched serum samples, thus further indicating that the source of variability arises from the sample type rather than the instrumentation.

We can describe the variability in the sputum as emerging from heterogeneity in the population, assessed by comparing pool variance to population variance, and technical variability; sample processing and measurement in the mass spectrometer. Such technical variability can be assessed through analysis of pooled reference samples and replicate injections, respectively. Furthermore, we used the open source software, variancePartition, to identify the drivers of protein measurement variation. Heterogeneity in the population drives most variance in the measurement of the top 20 proteins with highest coverage across samples, except for IGHA1_human, Immunoglobulin heavy constant alpha 1, whose variance was most attributable to mass spectrometry running batch. Immunoglobulins are highly conserved proteins, difficult to distinguish; minor changes in the running conditions within the mass spectrometer may result in different identity assignment. Across the top 40 proteins with highest coverage across samples, 75% of variability could be attributed to heterogeneity in the population (Fig. 3B).
Sources of variability in protein measurement estimated by variance Partition. A) the fraction of total variation in measurement of the top 20 proteins by highest coverage in samples, attributable to individual, extraction batch, eosinophil cell count, age, neutrophil cell count, total measured serum IgE, mass spectrometry running batch, sex and residuals. B) proteome-wide violin plot of the distribution of variance explained by each variable across the top 40 proteins by highest coverage in samples.

Proteins with high inter-individual population variance are shown in Supplementary Table S1, which illustrates those proteins that had a high CV% in the healthy population but were relatively stable in the pooled samples and injection replicates. Many of the proteins with individual to individual variation are known to have roles in inflammation (S100A proteins, A1AT), or are likely the result of salivary contamination of samples (e.g. Amylase).

One of the features used for assessing measurement error was injection repeatability. We defined a poor injection repeat as any protein in a given sample with >1.5-fold difference in measurement between injections. Such variation occurred in ~6% of all quantified sputum proteins, and ~5% of all quantified serum proteins. Variability occurred less frequently in the ‘core proteome region’ of serum where there was only 1.2% variation across duplicate injections. However, in sputum this value increased to 8.7% (1632 of 20412 individual quantifications) of the identified proteins (~6% in pools). It should be noted that the majority of these poor replicators in sputum occurred where there was a lower frequency of identification (higher rank in Supplementary Tables S2 and S3).

Variability in protein abundance measurements across samples. Frequency histograms represent: on the x axis CV% in increments of bin size 20, and inset scatter plots show CV% vs protein rank (proteins were ranked by order of abundance and frequency of identification across samples). Heatmap B illustrates Healthy sputum is highly variable from participant to participant compared to pools. Plots A and C show variability in protein abundance across healthy sputum samples, and pooled samples respectively; with the pools showing least variability across the different samples. Inset graphs illustrate the variability increase as coverage and abundance decreases. The variability seen in sputum is likely due to sample heterogeneity, and this is contrasted to serum sample measurements in plots D, which illustrate the relative homogeneity of that fluid across study participants.
Figure 5. Distribution and variation in measurement of potential inflammatory biomarkers in the healthy population compared to pooled controls. Proteins showed varying levels of spread indicating that the sample size required for statistical power will vary significantly per analyte, with the top graph showing those proteins with lower variability and the bottom, those with higher variability. Note that the pooled samples were taken from asthmatics and non-asthmatics, and as such the means of a number of these proteins will be higher or lower in the pools depending on each protein’s role in inflammatory disease. Therefore, pool samples can be used to contrast not only the measurement error but also any potential sub-clinical inflammatory effects in healthy participants.

Proteins that showed variability due to sample handling were identified by their high measurement stability in replicate injections but high variability across the pools (Supplementary Table S2). We observed a number of proteins with poor repeatability of quantification.

Some proteins are difficult to measure with good repeatability by Mass Spectrometry. These poor MS quantifiers showed high CV% in pooled samples and poor replication of quantification across injections. This variability in measurement is likely due to the behaviour of their peptides in HDMS\(^E\) or errors in database searching and quantification (e.g. due to homologues or protein to protein ambiguity) (Supplementary Table S3).

The variability in quantitation was non-uniform throughout the dataset, i.e. it varied on a protein to protein basis. This observation in the samples from healthy participants, which would be used as a control group for comparison with samples from participants with disease, has far reaching implications. This is particularly so in terms of experimental design and statistical power for biomarker discovery using unbiased sampling techniques. In order to explore this phenomenon further, we performed a literature search and generated a database of potential respiratory biomarkers (supplementary excel file 2). The database was used to identify highly cited respiratory proteins associated with disease, and then, in a post-hoc manner and using the data on variability gathered from our experiments, we performed an in silico exploratory evaluation of sample size requirements for MS-based biomarker studies for these biomarkers.

Biomarker proteins identified in more than three published studies and also identified in our study were chosen for more detailed analysis. A sub-set of these proteins are presented in Fig. 5. Low variability in the pooled samples, but high variability in the population, indicated that the majority of variability arose from inter-individual variation, while diverse measurements in both the pools and the population indicated that there was a likely influence of experimental variation to consider for that protein.

Twenty-four of the proteins in the biomarker database were observed in ≥20 healthy participants in our study, which allowed a series of sample size calculations to be performed. The calculations quantified the relationship between variability of the samples, given by the standard deviation of the measurements, and the sample size required to achieve 80% power (Table 2 and supplementary Fig. S4). The sample size required for the given statistical power varied between proteins because of a combination of experimental and biological variation. This highlights that while statistical differences for some biomarkers can be reliably identified from sample sizes that are routinely used in proteomic analysis, others require very large sample sizes in order to confidently identify an effect. It is noted that in studies where patient allocation is unbalanced (e.g. 1:2 or 1:3 cases to controls), the total necessary sample size required is greater. A 1:1 allocation provides the most efficient design.
Table 2. Sample size per group, to detect 1.5 or 2-fold differences with 80% power.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Fold Change</th>
<th>SD(log2(x))</th>
<th>Po</th>
<th>1.5%</th>
<th>97.5%</th>
<th>2%</th>
<th>95%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMP1</td>
<td>0.303</td>
<td>0.352</td>
<td>145</td>
<td>15</td>
<td>74.5</td>
<td>28</td>
<td>80</td>
<td>6</td>
</tr>
<tr>
<td>GSTT1</td>
<td>0.324</td>
<td>0.398</td>
<td>146</td>
<td>15</td>
<td>74.5</td>
<td>28</td>
<td>80</td>
<td>6</td>
</tr>
<tr>
<td>ACTB</td>
<td>0.379</td>
<td>0.453</td>
<td>147</td>
<td>15</td>
<td>74.5</td>
<td>28</td>
<td>80</td>
<td>6</td>
</tr>
<tr>
<td>TRY</td>
<td>0.478</td>
<td>0.525</td>
<td>148</td>
<td>15</td>
<td>74.5</td>
<td>28</td>
<td>80</td>
<td>6</td>
</tr>
<tr>
<td>S10A1</td>
<td>0.525</td>
<td>0.609</td>
<td>149</td>
<td>15</td>
<td>74.5</td>
<td>28</td>
<td>80</td>
<td>6</td>
</tr>
<tr>
<td>CO3</td>
<td>0.609</td>
<td>0.683</td>
<td>150</td>
<td>15</td>
<td>74.5</td>
<td>28</td>
<td>80</td>
<td>6</td>
</tr>
<tr>
<td>S10A9</td>
<td>0.711</td>
<td>0.776</td>
<td>151</td>
<td>15</td>
<td>74.5</td>
<td>28</td>
<td>80</td>
<td>6</td>
</tr>
<tr>
<td>S10A8</td>
<td>0.776</td>
<td>0.840</td>
<td>152</td>
<td>15</td>
<td>74.5</td>
<td>28</td>
<td>80</td>
<td>6</td>
</tr>
<tr>
<td>TRFL</td>
<td>0.804</td>
<td>0.879</td>
<td>153</td>
<td>15</td>
<td>74.5</td>
<td>28</td>
<td>80</td>
<td>6</td>
</tr>
<tr>
<td>ANXA2</td>
<td>0.879</td>
<td>0.953</td>
<td>154</td>
<td>15</td>
<td>74.5</td>
<td>28</td>
<td>80</td>
<td>6</td>
</tr>
<tr>
<td>VTDB</td>
<td>0.953</td>
<td>1.021</td>
<td>155</td>
<td>15</td>
<td>74.5</td>
<td>28</td>
<td>80</td>
<td>6</td>
</tr>
<tr>
<td>A1AT</td>
<td>1.021</td>
<td>1.088</td>
<td>156</td>
<td>15</td>
<td>74.5</td>
<td>28</td>
<td>80</td>
<td>6</td>
</tr>
</tbody>
</table>

Figure 6. Comparison of coverage of U-BIOPRED sputum with other studies. Including other sputum studies (A-B), other proteomic studies of airway fluids and secretions (C), and respiratory tissues found in the protein atlas normal expression database (D). U-BIOPRED (Healthy) sputum proteome was compared to known proteomes of granulocytes and macrophages (E), and to the proteomes of neutrophil granules (F). Comparisons were also made with Saliva and Blood (G).

Salivary contamination

We compared the abundance of proteins reported to be salivary proteins against the squamous cell counts from the study participants (Supplementary Figure S5). Although salivary proteins tended to be high when the percentages of squamous cell counts were high, this was not consistent and many participants had low squamous counts but high levels of salivary proteins.

Tissue and cell origins of the sputum proteome

Proteins found in this study have been previously observed in studies of human sputum proteomes. Comparison with studies of induced sputum by Gharib et al., Nicholas et al., and Titz et al. (Fig. 6 A to B) showed an extensive overlap of measured proteins. Here, in the extended proteome, we found 63% of the 232 proteins identified by Gharib et al., 86% of the 171 proteins identified by Nicholas et al., but only 31% of the 278 proteins found in the study by Titz et al., reflecting the heterogeneity in sputum proteomes. Overlap of proteins found between two studies particularly with different methodologies increases confidence of the presence of this protein in sputum. Overlap of proteins between more than two studies further supports the identifications. The extended proteome in the current study showed the highest number of proteins shared with other studies, representing an improvement in protein identification, but also added confidence to the proteins identified in few samples by HDMS.

The overlap of proteins found between the extended sputum proteome and other tissues and cells indicates potential origins of proteins identified in this study of the sputum proteome. Sputum is a complex biofluid, consisting of proteins of multiple origins and can therefore reflect a complex biological picture. Despite sampling different airway compartments, similar patterns of proteome coverage were also observed when comparing airway mucus, exhaled breath condensate (EBC) and studies analysing bronchial alveolar lavage (BAL). However, the number of proteins identified here were only of similar magnitude to those in the BAL study. This may be partially due to improvements in sensitivity of protein measurement. Approximately 20% of proteins identified in the BAL study were also identified in the current study, which likely reflects an overlap in the high confidence detection of high abundance proteins, and a variability between studies in sampling of less abundant proteins. Results also showed extensive overlap with proteins identified from respiratory tissue analyses in the Protein Atlas (Fig. 6D). There was also overlap in the identifications of measured
proteins to those of proteomes measured for
eosinophils, macrophages and granulocyte (Fig. 6E)
and neutrophil granules (Fig. 6F). Identification of
proteins from all major tissues and cell types of the
airways highlights not only the complexity of sputum
as a clinical fluid, but also its utility for accessing lung
biology. Such results were also reflected in our core
sputum proteome, where extensive overlap was
observed across a variety of tissues and/or biofluids
(Supplementary Fig. S5). The interpretation of these
results is limited due to the differences in the
sensitivity and variability of the protein measurement
techniques. In most cases, the proteins identified in
the studies are not tissue or cell specific, for example,
the proteome of the compared macrophage data set
covers half of the genome, which suggests that most of
these proteins are not macrophage specific. With
increasing sensitivity in protein measurement
techniques, there is a corresponding increase in the
proportion of low abundance proteins, therefore,
overlap of proteins detected in studies reporting fewer
proteins may be more useful in finding the biological
origins of the proteins. In studies where fewer proteins
are reported, these likely reflect high abundance, easy
to detect proteins and do not include low abundance
proteins which are harder to detect.

Figure 7 Subcellular localisation prediction results from IPA
analyses of the proteins from the core and extended sputum
proteome against a background gene set restricted to
Ingenuity Knowledge Base genes from expected sources of
sputum proteins.

To further understand the protein composition of
sputum, we also investigated possible systemic or
non-respiratory origins of our identified proteins
through comparison to other relevant biofluids. We
identified a 50% overlap with proteomes of sputum,
the upper airway, saliva and blood, similar to that
reported previously in the upper airway, saliva and blood, similar to that
reported previously66 (Fig. 6G).

Functional analysis

Subcellular localisation of proteins was analysed using
Ingenuity Pathways Analysis (IPA) (Fig. 7). 27% of the
proteins in the core sputum proteome were predicted
to be extracellular or secreted, with a further 8% predicted to be integral membrane or cell surface
proteins. 50% of identities were predicted to be
cytoplasmic proteins.

The LXR/RXR pathway refers to liver X receptor (LXR)
which is activated by oxysterol ligands to bind retinoid X
receptors (RXRs). Resultant LXR-RXR heterodimers bind
LXR response elements and regulate expression of genes
involved in inflammation, metabolism and cholesterol
metabolism. The FXR/RXR pathway refers to the bile-acid
concentration-mediated farnesoid X receptor (FXR) and RXR
regulation of lipid metabolism. Integrin linked kinase (ILK)
signalling refers to the ILK-mediated control of cytoskeleton
remodelling. eNOS Signalling refers to the mechanism of
Nitric Oxide (NO) production by the endothelial NO
Synthase (eNOS).

Enrichment and pathway analyses for the core sputum
proteome and extended sputum proteome were
performed using Funrich, Go-Rilla and IPA. As gene
ontology mapping presents a high-level of redundant
terms, we utilised REVIGO to collapse and summarise
like terms as treemaps. The results of these GO
enrichment analyses for the core sputum proteome
are shown in Supplementary Fig. S7, A for cellular
components, B for molecular function, and C for
biological processes. Analyses of the extended sputum
proteome were also performed and shown in Fig. S7
D-E. The size of each individual square represented in
each treemap is proportional to the \(-\log_{10} p\) value for
the enrichment of that category, measured by Fisher’s
exact test. The ontologies are grouped by related
terms and defined by colour, providing a landscape
overview of induced sputum. Enrichment was observed for vesicle related components; GORilla
q=1.92E-7, with 41 proteins identified with the GO term
‘vesicle’, however, proteins have multiple GO terms
and these results only suggest an enrichment of
proteins originating from vesicles. Many granulocyte
functions involve vesicle formation, such as the
release of extracellular vesicles, some with antibacterial effect, released during spontaneous
death of neutrophils69. Enrichment was also observed
for extracellular proteins, (immune) receptor and
antigen binding functions and dominated by
processes involved in homeostasis, and mucosal and
innate immunity. Further analysis of enriched
biological pathways showed that the top IPA
canonical pathways are thematically similar (Table 3),
with cell migration and tissue organisation (integrin
linked kinase signalling, actin cytoskeleton signalling,
leukocyte extravasation signalling), innate immunity
(acute phase response, complement) and regulation of
cytoskeleton, extracellular matrix (ECM) remodelling
and inflammation (e.g. FXR/RXR (farnesoid X
receptor / retinoid X receptor) or LXR (liver X
receptor)/RXR, RhoA - Ras homolog A, RhoGDI - Rho
GDP-dissociation inhibitor) being the most enriched
pathways and, to a lesser extent, those involved in
energy metabolism (e.g. glycolysis). Similar trends
were observed when exploring the association of
proteins with specific functions and diseases, with
inflammatory responses and immune cell migration
dominating the functional enrichment categories
(Table 4). These enriched pathways and functions
were also conserved in the extended sputum proteome
dataset (Supplementary Tables S4 and S5).

Investigating protein interaction networks (http://string-db.org) in our core sputum proteome
dataset (Supplementary Fig. S8), we observed a
number of small groups of interacting protein partners and significantly one large interaction network, highlighting the large number of functional relationships that are experimentally accessible in the core sputum proteome.

Table 3. Top 20 Enriched canonical pathways in the core sputum proteome set.

<table>
<thead>
<tr>
<th>Ingenuity Canonical Pathways</th>
<th>Enrichment P-value</th>
<th>Total number matched proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>LXR/RXR Activation</td>
<td>3.16E-14</td>
<td>17</td>
</tr>
<tr>
<td>Glycolysis I</td>
<td>2.00E-11</td>
<td>9</td>
</tr>
<tr>
<td>Acute Phase Response Signalling</td>
<td>6.31E-11</td>
<td>17</td>
</tr>
<tr>
<td>FXR/RXR Activation</td>
<td>5.01E-10</td>
<td>15</td>
</tr>
<tr>
<td>Gluconeogenesis I</td>
<td>7.76E-10</td>
<td>8</td>
</tr>
<tr>
<td>Actin Cytoskeleton Signalling</td>
<td>2.69E-09</td>
<td>15</td>
</tr>
<tr>
<td>ILK Signalling</td>
<td>2.00E-09</td>
<td>13</td>
</tr>
<tr>
<td>Primary Immunodeficiency</td>
<td>3.89E-09</td>
<td>7</td>
</tr>
<tr>
<td>Pyruvate Fermentation to Lactate</td>
<td>1.45E-09</td>
<td>3</td>
</tr>
<tr>
<td>Complement System</td>
<td>1.70E-09</td>
<td>6</td>
</tr>
<tr>
<td>RhoGDI Signalling</td>
<td>2.95E-09</td>
<td>11</td>
</tr>
<tr>
<td>Hematopoiesis from Pluripotent</td>
<td>3.16E-09</td>
<td>6</td>
</tr>
<tr>
<td>Leukocyte Extravasation Signalling</td>
<td>3.89E-09</td>
<td>12</td>
</tr>
<tr>
<td>RhoA Signalling</td>
<td>3.89E-09</td>
<td>9</td>
</tr>
<tr>
<td>Clathrin-mediated Endocytosis</td>
<td>5.50E-09</td>
<td>11</td>
</tr>
<tr>
<td>Glucocorticoid Receptor Signalling</td>
<td>7.08E-09</td>
<td>14</td>
</tr>
<tr>
<td>Cell Cycle: G2/M DNA Damage</td>
<td>8.13E-09</td>
<td>6</td>
</tr>
<tr>
<td>Epithelial Adherens Junction</td>
<td>1.69E-09</td>
<td>9</td>
</tr>
<tr>
<td>Coagulation System</td>
<td>1.90E-09</td>
<td>5</td>
</tr>
<tr>
<td>eNOS Signalling</td>
<td>1.99E-09</td>
<td>9</td>
</tr>
</tbody>
</table>
We have assessed the biological relevance of the sputum using MS E. This study provides the first large-scale analysis of the sputum proteome, particularly in the context of enrichment of proteins unique to each study.

Our previous study of the sputum proteome we have previously reported 284 proteins in the core healthy sputum proteome that are reliably and repeatedly measured and 1666 proteins in the extended proteome, additionally detailing less repeatedly measured proteins that are usually also less abundant. We have assessed the biological relevance of the proteome, particularly in the context of enrichment against the wider human proteome and have highlighted that secreted proteins and vesicular components are highly enriched in this biofluid.

Proteomic studies of sputum to date have been small but, nevertheless, useful. Our previous study identified 191 proteins in the sputum of a single individual using 2-dimensional gel electrophoresis and mass spectrometry which is biased towards high abundance proteins and not amenable to high throughput analysis. Comparison of our results with sputum studies by other authors shows excellent overlap in detected proteins, but also a number of proteins unique to each study.

Comparison of our results with sputum studies by other authors shows excellent overlap in detected proteins, but also a number of proteins unique to each study. Gharib et al., applied shotgun mass spectrometry to assess 5 healthy and 10 asthmatic participants and identified 254 proteins in all participants' sputum. Using a non-parametric test developed by the authors, called the spectral index, they found 17 proteins whose concentrations were significantly different between asthma and health, including serpin peptidase inhibitor (SRPINA1) and secretoglobin (SCGB1A1, also known as Clara cell 10-kD protein) that were increased and decreased in concentration, respectively in asthma.

Titz et al., conducted an impressive study applying LC–MS/MS of tryptic peptides labelled with Tandem Mass Tags (TMT) to sputum samples of 216 participants equally composed of healthy non-smokers, healthy never smokers, COPD patients and current smokers.

### Table 4: Top 20 enriched functions and diseases found by IPA of the core sputum proteome.

<table>
<thead>
<tr>
<th>Disease and function category hierarchy</th>
<th>Specific annotation</th>
<th>Enrichment p value</th>
<th>Number of matched proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infectious Diseases, Inflammatory Disease, Respiratory Disease</td>
<td>Severe acute respiratory syndrome</td>
<td>1.92E-05</td>
<td>14</td>
</tr>
<tr>
<td>Infectious Diseases</td>
<td>Viral Infection</td>
<td>4.17E-05</td>
<td>18</td>
</tr>
<tr>
<td>Inflammatory Response</td>
<td>Inflammatory response</td>
<td>4.40E-05</td>
<td>18</td>
</tr>
<tr>
<td>Connective Tissue Disorders, Inflammatory Disease, Skeletal and Muscular Disorders</td>
<td>Rheumatic Disease</td>
<td>9.30E-05</td>
<td>19</td>
</tr>
<tr>
<td>Connective Tissue Disorders, Inflammatory Disease, Skeletal and Muscular Disorders</td>
<td>Arthritis</td>
<td>1.30E-07</td>
<td>18</td>
</tr>
<tr>
<td>Cellular Movement, Immune Cell Trafficcking</td>
<td>Leukocyte migration</td>
<td>2.22E-07</td>
<td>20</td>
</tr>
<tr>
<td>Cellular Movement, Haematological System Development and Function, Immune Cell Trafficcking, Inflammatory Response</td>
<td>Cell movement of phagocytes</td>
<td>2.87E-07</td>
<td>16</td>
</tr>
<tr>
<td>Immunological Disease</td>
<td>Systemic autoimmune syndrome</td>
<td>3.05E-07</td>
<td>16</td>
</tr>
<tr>
<td>Inflammatory Disease</td>
<td>Chronic inflammatory disorder</td>
<td>5.92E-07</td>
<td>16</td>
</tr>
<tr>
<td>Cellular Movement, Haematological System Development and Function, Immune Cell Trafficcking</td>
<td>Cell movement of leukocytes</td>
<td>6.79E-07</td>
<td>19</td>
</tr>
<tr>
<td>Cellular Movement, Haematological System Development and Function, Immune Cell Trafficcking, Inflammatory Response</td>
<td>Rheumatoid arthritis</td>
<td>1.02E-06</td>
<td>15</td>
</tr>
<tr>
<td>Cellular Movement, Haematological System Development and Function, Immune Cell Trafficcking</td>
<td>Cell movement of myeloid cells</td>
<td>1.73E-06</td>
<td>15</td>
</tr>
<tr>
<td>Cell-To-Cell Signalling and Interaction, Haematological System Development and Function, Immune Cell Trafficcking</td>
<td>Adhesion of immune cells</td>
<td>3.58E-06</td>
<td>13</td>
</tr>
<tr>
<td>Cellular Movement, Haematological System Development and Function, Immune Cell Trafficcking, Inflammatory Response</td>
<td>Chemotaxis of phagocytes</td>
<td>5.26E-06</td>
<td>12</td>
</tr>
<tr>
<td>Cellular Movement, Haematological System Development and Function, Immune Cell Trafficcking, Inflammatory Response</td>
<td>Chemotaxis of leukocytes</td>
<td>6.22E-06</td>
<td>14</td>
</tr>
<tr>
<td>Cellular Movement, Haematological System Development and Function, Immune Cell Trafficcking, Inflammatory Response</td>
<td>Chemotaxis of myeloid cells</td>
<td>8.82E-06</td>
<td>12</td>
</tr>
<tr>
<td>Cellular Movement, Haematological System Development and Function, Immune Cell Trafficcking</td>
<td>Cell movement of granulocytes</td>
<td>1.06E-05</td>
<td>11</td>
</tr>
<tr>
<td>Cellular Assembly and Organization</td>
<td>Formation of rosettes</td>
<td>1.45E-05</td>
<td>3</td>
</tr>
<tr>
<td>Organismal Survival</td>
<td>Organismal death</td>
<td>1.86E-05</td>
<td>7</td>
</tr>
<tr>
<td>Cellular Movement, Haematological System Development and Function, Immune Cell Trafficcking, Inflammatory Response</td>
<td>Cell movement of neutrophils</td>
<td>2.08E-05</td>
<td>9</td>
</tr>
</tbody>
</table>

DISCUSSION

This study provides the first large-scale analysis of sputum using MS E, a data-independent proteomics approach. To our knowledge, it provides the most comprehensive description to date of airway lining fluid proteins and, thus, significantly advances on the sputum proteome.
Relative quantification was achieved by sequentially measuring the proteome of demographically matched samples (1 from each cohort per run) against a pooled reference sample. They analysed differential protein expression only in proteins detected in at least 2/3 of samples per study group and reported proteins differentially abundant between groups but did not report the total proteomes. This approach pointed to 15 proteins differentially abundant when comparing patients with COPD and current smokers and many more when compared with non-smokers.

Differences in reported proteins between studies can be explained by differences between individual phenotypes; biological variability. Differences are also attributable to differences in sample preparation and analysis methods. Likewise, where there are similarities in techniques used, there are large overlaps in proteins reported. Highlighting this paradigm, mucins 5A and 5C and the IgGFc-binding protein, which are all major components of airway mucus, were not identified in our study, likely due to due to post-digestion filtration of samples to remove large indigestible substances e.g. mucopolysaccharides, thereby removing these proteins from the analysis.

**Protein detection and quality control**

The study design and attention to quality control has allowed us to investigate numerous experimental and measurement effects in our dataset: technical reproducibility in measurement of each protein, inter-individual variability of each protein and identification of salivary contamination of sputum samples. The semi-stochastic nature of peptide sampling in MS-based proteomic approaches, even when using data independent methods, often results in highly abundant proteins being measured more reproducibly, with proteins of lower abundance being identified less frequently, leading to increased sparsity within a dataset. This is most evident in larger studies, where the proportion of proteins identified across all samples is lower than in smaller studies. These effects were visible in both the serum and sputum datasets in the current study, and as such, influenced our approach to developing our ‘Top-90 intensity normalisation’ strategy. In a data matrix ranked by protein intensity and frequency of identification, the low frequency region was seen to be more variable between samples, both in the number of protein identities per sample and their intensity measurement. Such variation can influence normalisation strategies based upon total MS intensity. We have found that using the region with the greatest coverage for adjusting samples, the topplo method, is an effective method of normalisation for protein load and intensity.

Comparison of variability across injection replicates and pooled QC samples within the healthy sputum proteome dataset allowed the precision of measurement for specific proteins to be assessed. Notably, measurements of some proteins were reproducible across samples, whereas other proteins showed poor replicability across sample replicates and pooled QC samples, indicating that these proteins are intrinsically recalcitrant to precise measurement by HDMS. As expected, there was a trend for proteins with lower intensity to show higher variability since they are generally represented by fewer peptides and are more prone to interference due to noise. In this study we chose to define a core proteome using proteins detected in ≥40% of samples representing proteins reproducibly measured. However, we recognise that the proteome presented here will require verification in future studies.

By systematically examining the variability in pools, samples and replicate injections, we were able to identify potential sources of variation as either: sample heterogeneity, sample processing effects, and MS measurement or post processing effects. Proteins that display high CV of measurement across pools, but good replication between injections, are likely to be poor quantifiers due to sample preparation and are likely to be unstable or variably modified. The majority of these are at the lower end of the coverage spectrum. Many of the variable proteins were also identified as membrane or centriolar proteins, which often contain hydrophobic protein regions that can affect protein digestion or peptide extraction efficiencies, thereby influencing the MS measurement. In a similar fashion, those proteins that were stable in the pools, but were variable in injection replicates, are likely to be poor MS quantifiers or liable to post-processing errors. Many of these have potential homologs which could interfere with their in silico quantitation. The results of our QC analysis suggest that variation of proteins in pooled QC samples and sample replicate measurements need to be routinely assessed across sample populations for individual proteins and for different biofluids. This will be critical in developing targeted biomarker assays and designing large discovery projects.

**Protein measurement and concentration variability and effect on statistical power**

Determination of normal/healthy levels of proteins is essential for the identification of potential disease biomarkers. It is, therefore, crucial to understand both the technical and biological variation affecting measurement of this baseline proteome. In the current study, the measurements of a number of ‘core sputum proteome’ proteins were highly stable in repeat analyses of pooled sputum samples, pointing to the robustness of the MS analysis, but inter-individual variability was significant (Supplementary table S1), in stark contrast to the low inter-individual variability in serum from the same individuals.

Based upon respiratory protein biomarkers previously identified from other studies and the variability of the proteins observed in our healthy population study, we calculated the sample size required to observe a
specific effect against that baseline in the context of a disease setting. By parallel reference to our pooled sample measurements, we inferred how much of the variance can be attributed to processing and/or technical measurement and how much is a consequence of inter-individual heterogeneity. The greater the variability in the measurement, the greater the sample size required to achieve 80% power (Table 2). For example, quantitation of modest (1.5-fold) differences for a highly variable protein, such as haemoglobin B (HBB), requires >500 patients. By contrast, differential expression of proteins such as polymeric immunoglobulin receptor may be quantifiable with as few as 20 patients. In the case of HBB, high sample-to-sample variation was observed, although variation of HBB across the pooled samples was low. As highlighted above, saliva is often a contaminant in healthy individuals who have difficulty producing an adequate sputum sample. Since saliva is susceptible to contamination with blood as a result of gum disease (Djukanovic ERJ). However, in the current study, there were no significant differences between atopic and non-atopic patients, possibly because their allergic status was too mild or the allergic processes in the airways were not active at the time of sampling.

A small number of individuals displayed neutrophilia (>80% neutrophils) which was reflected by observed differences in neutrophil proteins measured in these study participants. Neutrophilic inflammation is often a hallmark of infection, disease exacerbations, or severity of chronic airway diseases like asthma and COPD (Djukanovic ERJ), but is often confounded by smoking status and steroid treatment, influencing neutrophil half-life, activity and migration (Djukanovic ERJ). A number of neutrophil proteins correlated with neutrophil cell counts (Fig. 2); neutrophil defensin, S100A8, S100A9 and MMP9. Concentrations of the major neutrophil granule protein, myeloperoxidase only weakly correlated with sputum % neutrophil cell counts. Although neutrophil cell counts correlate specifically with these neutrophil derived proteins, we do not observe a large effect across the rest of the core or extended proteome as shown in Figure 3. As such, these proteins are useful biomarkers of neutrophilia whilst not compromising the potential identification of other disease signatures.

Salivary contamination is an inevitable confounding factor in sputum analyses, particularly in healthy individuals and other participants where sputum induction is less successful. We compared squamous cell counts with the expression of proteins found in saliva in previous studies and found mixed correlations. Squamous cell count does not perfectly reflect salivary contamination of sputum samples, however, the salivary proteins studied in these comparisons may additionally originate from other tissues, not just saliva. Further investigation is required to determine better markers of salivary contamination in sputum.

Origins and biological context of the sputum proteome

We investigated the predicted subcellular localisation of the proteins from the core and extended sputum proteomes (Fig. 7 & S7). A large portion of the proteins measured in the sputum proteome were reported to be cytoplasmic by GOrilla analysis. Since our analysis is of the supernatant from induced sputum, where cells and cell debris are removed prior to analysis, this was expected. In addition, the supernatant would also include granulocytes, which contain granules of cytoplasmic proteins for release into the sputum.

When comparing the sputum proteome in the current study with those from studies of airway tissue and fluid samples, significant numbers of identified proteins were seen to overlap. The number of proteins for which a potential cell or tissue origin could not be assigned was low. Significant numbers of the identified proteins likely have origins in other tissues of the body: for example, acute phase proteins are produced largely in the liver and enter the lungs via capillaries together with other plasma proteins. Functional analysis of the healthy sputum proteome identified proteins associated with innate immune defence, inflammatory responses via complement and acute phase proteins, phagocytic cells (macrophages and neutrophils) and, to a lesser extent, eosinophils.

A number of signalling pathways were enriched in sputum, including LXR, a member of the nuclear receptor family of transcription factors that are closely related to nuclear receptors, such as the peroxisome proliferator-activated receptors (PPARs), FXR, also known as bile acid receptor, RXR, that is activated by retinoic acid. FXR/RXR are known to be important in...
macrophage lipid metabolism and inflammation. Such regulatory pathways have complex roles in inflammatory biology: FXR inhibits inflammation, while LXR agonists have been shown to increase airway reactivity and smooth muscle growth in an asthma model. LXR has also been implicated in counter-regulation of toll like receptor induced inflammatory responses, which are involved in inflammatory respiratory diseases. We also observed an enrichment of the RhoA and RhoGDI signalling pathways, known to be involved in hyper-responsiveness in asthma, and superoxide generation in macrophages, respectively.

An interesting observation was the significant enrichment of vesicles and vesicular components in the sputum (Fig. S7). These small membranated particles, either described as exosomes (nanovesicles) or ectosomes (microparticles), are released by multiple cell types, including immune cells, and have been reported in sputum and lung secretions. Secretory vesicles have recently become an area of interest for their potential pro and anti-inflammatory functions. In response to non-specific complement mediated inflammation neutrophils produce ectosomes that are coated in and loaded with proteins often associated with granules. In addition to neutrophil derived vesicles, antigen-loaded exosomes from mast cells, dendritic cells, epithelial cells and T lymphocytes have been highlighted as being potentially important for allergy, and eosinophils produce cytokine containing vesicles e.g. degranulation, that may also be important for asthma.

Secretory vesicles contain high levels of cytoplasmic proteins. For example, exosomes isolated in the BAL of asthmatics are enriched for inflammatory leukotriene production and may help explain the large numbers of cytosolic proteins we have identified in sputum. In addition to the protein loading capacity of these vesicles, functions involving the transport of nucleic acids have also been identified. For example, microRNA-loaded vesicles have been highlighted as potentially important in asthma and inflammatory lung disease signalling, and have been shown to be different between BAL samples of healthy and asthmatic individuals.

CONCLUSIONS

The mapping of the healthy human sputum proteome in the current study has considered experimental and technical variability, population variance driven by daily exposure of the lung to the external environment, and sample complexity due to multiple potential protein origins, effects of cellular composition and potential for contribution of vesicular components. Functionally, homeostasis and defence mechanisms dominate the measured sputum proteome. The specific experimental and technical variability of the applied methodology must not be underestimated and there are implications for minimum sample sizes required for determining differential abundance between groups with statistical power. The comprehensive approach we have used for this analysis of the healthy sputum proteome provides a good comparator dataset for proteomes from patients suffering from inflammatory lung disease.

SUPPORTING INFORMATION

The following supporting information is available for charge at ACS website http://pubs.acs.org

Supplementary figures; S1. Distribution of sputum proteins across samples. S2. Topological data analysis (TDA) network showing no clustering according to abundant proteins or clinical variables. S3. Heatmaps of the U-BIOPRED sputum proteomics data. S4. The sample size required per group relative to variability of the samples given by the standard deviation of the measurements. S5. Correlations between proteins and squamous cell counts. S6. Overlap of proteins in the U-BIOPRED core sputum proteome in comparison to other studies. S7. Ontology treemaps. S8. Interaction network of the Healthy core sputum proteome. Table S1. Proteins with high population variance compared to measurement variance. Table S2. Proteins with poor repeatability due to sample processing. Table S3. Poor quantifying proteins in the mass spectrometer. Table S4. Top Canonical pathways enriched in the extended Healthy sputum proteome. Table S5. Top function and disease representatives in the extended healthy sputum proteome. SupplementaryExcelFile1; proteins of the core and extended sputum proteomes. SupplementaryExcelFile2; a database of potential respiratory biomarkers.
ACKNOWLEDGMENT

This paper is presented on behalf of the U-BIOPRED Study Group with input from the U-BIOPRED Patient Input Platform, Ethics Board and Safety Management Board. We thank all the members of each recruiting centre for their dedicated effort, devotion, promptness and care in the recruitment and assessment of the participants in this study. U-BIOPRED is supported through an Innovative Medicines Initiative Joint Undertaking under grant agreement no. 115010, resources of which are composed of financial contribution from the European Union’s Seventh Framework Programme (FP7/2007–2013) and EFPIA companies’ in kind contribution (www.imi.europa.eu). We would also like to acknowledge help from the IMI funded eTRIKS project (EU Grant Code No. 115446).

The members of the U-BIOPRED Study Group are as follows: H. Ahmed, European Institute for Systems Biology and Medicine, University of Lyon, France; D. Allen, North West Severe Asthma Network; Pennine Acute Hospital NHS Trust; P. Badorek, Fraunhofer ITEM; S. BallerEA, European Institute for Systems Biology and Medicine, University of Lyon, France; F. Baribaud, Janssen R&D, USA; M.K. Batuwitage, Imperial Asthma Network; Pennine Acute Hospital NHS Trust; P. Badorrek, Fraunhofer ITEM; S. BallerEA, European Institute for Systems Biology and Medicine, University of Lyon, France; M. Behndig, Mannheim, Germany; A.F. Behndig, Umeå University; A. Berglind, Karolinska University Hospital and Karolinska Institutet; A. Berton, Boehringer Ingelheim Pharma GmbH & Co. KG; J. Bigler, Amgen Inc; M.J. Boedigheimer, Amgen Inc; K. Bønnelykke, Boehringer Ingelheim Pharma GmbH & Co. KG; J. Bigler, Amgen Inc; M.J. Boedigheimer, Amgen Inc; K. Bønnelykke, University of Copenhagen and Danish Pediatric Asthma Center, Gentofte Hospital, University of Copenhagen, Denmark; P. Brinkman, Academic Medical Centre, University of Amsterdam; A. Bush, Department of Paediatrics and National Heart and Lung Institute, Imperial College, London; Department of Respiratory Paediatrics, Royal Brompton Hospital, London, UK; D. Campagna, University of Catania; C. Casaulta, University Children’s Hospital Bern, Switzerland; A. Chaiboonechoe, European Institute for Systems Biology and Medicine, University of Lyon, France; T. Davison, Janssen R&D, USA; B. De Meulder, European Institute for Systems Biology and Medicine, University of Lyon, France; I. Delin, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden; P. Dennison, NIHR Southampton Respiratory Biomedical Research Unit and University of Southampton; P. Dodson, AstraZeneca, Mölndal, Sweden; L. El Hadjam, European Institute for Systems Biology and Medicine, University of Lyon, France; D. Erzen, Boehringer Ingelheim Pharma GmbH & Co. KG; C. Faulenbach, Fraunhofer ITEM; K. Fichtner, Boehringer Ingelheim Pharma GmbH & Co. KG; N. Fitch, BioSci Consulting, Belgium; E. Formaggio, PhD, Project manager, Verona Italy; M. Gahlmann, Boehringer Ingelheim (Schweiz) GmbH; G. Galfy, Semmelweis University, Budapest, Hungary; D. Garissi, Global Head Clinical Research Division, CROMSOURCE, Italy; T. Garret, BioSci Consulting, Belgium; J. Gent, Royal Brompton and Harefield NHS Foundation Trust; E. Guillmant-Farry, Royal Brompton Hospital, London, UK; E. Henriksson, Karolinska Institutet; U. Hoda, Imperial College; J.M. Hohlfeld, Fraunhofer ITEM; X. Hu, Amgen Inc; A. James, Karolinska Institutet; K. Johnson, Centre for respiratory medicine and allergy, Institute of Inflammation and repair, University Hospital of South Manchester, NHS Foundation Trust, Manchester, UK; N. Jullian, European Institute for Systems Biology and Medicine, University of Lyon, France; G. Kerry, Centre for respiratory medicine and allergy, Institute of Inflammation and repair, University Hospital of South Manchester, NHS Foundation Trust, Manchester, UK; M. Klüglich, Boehringer Ingelheim Pharma GmbH & Co. KG; R. Knowles, Arachos Pharma, Stevenage, UK; J.R. Konradsen, Karolinska University Hospital and Karolinska Institutet; K. Kretos, UCB, Slough, UK; L. Krueger, University Children’s Hospital Bern, Switzerland; A-S. Lantz, Karolinska University Hospital and Karolinska Institutet; C. Larminie, GSK, London, UK; P. Latzin, University Children’s Hospital Bern, 3010 Bern, Switzerland; D. Lefaudeux, European Institute for Systems Biology and Medicine, University of Lyon, France; N. Lemonnier, European Institute for Systems Biology and Medicine, University of Lyon, France; L.A. Lowe, Centre for respiratory medicine and allergy, Institute of Inflammation and repair, University Hospital of South Manchester, NHS Foundation Trust, Manchester, UK; R. Lutter, Academic Medical Centre, University of Amsterdam; A. Manta, Roche Diagnostics GmbH, Mannheim, Germany; A. Mazein, European Institute for Systems Biology and Medicine, University of Lyon, France; L. McEvoy, University Hospital, Department of Pulmonary Medicine, Bern, Switzerland; A. Menzies-Gow, Royal Brompton
and Harefield NHS Foundation Trust; N. Mores, Università Cattolica del Sacro Cuore; C.S. Murray, Centre for Respiratory Medicine and Allergy, The University of Manchester, Manchester Academic Health Science Centre, University Hospital of South Manchester NHS Foundation Trust, Manchester, UK; K. Nething, Boehringer Ingelheim Pharma GmbH & Co. KG; U. Nihlén, Department of Respiratory Medicine and Allergology, Skåne University Hospital, Lund, Sweden; AstraZeneca R&D, Mön达尔, Sweden; R. Niven, North West Severe Asthma Network, University Hospital South Manchester NHS Trust; B. Nordlund, Astrid Lindgren Children’s Hospital, Karolinska University Hospital, Stockholm, Sweden; Department of Women’s and Children’s Health, Karolinska Institutet, Stockholm, Sweden; S. Nsubuga, Royal Brompton Hospital, London, UK; J. Pellet, European Institute for Systems Biology and Medicine, University of Lyon, France; C. Pison, European Institute for Systems Biology and Medicine, University of Lyon, France; G. Praticò, CROMSOURCE, Verona, Italy; M. Puig Valls, CROMSOURCE, Barcelona, Spain; K. Riemann, Boehringer Ingelheim Pharma GmbH & Co. KG; J.P. Rocha, Royal Brompton and Harefield NHS Foundation Trust; C. Rosso, Imperial College; G. Santini, Università Cattolica del Sacro Cuore; M. Saqi, European Institute for Systems Biology and Medicine, University of Amsterdam, Amsterdam, The Netherlands; Antonios Aliprantis, Merck Research Laboratories, Boston, USA; Kjell Alving, Dept Women’s and Children’s Health, Uppsala University, Uppsala, Sweden; Per Bakke, Department of Clinical Science, University of Bergen, Bergen, Norway; David Balgoma, Centre for Allergy Research, Karolinska Institutet, Stockholm, Sweden; Clair Barber, NIHR Southampton Respiratory Biomedical Research Unit and Clinical and Experimental Sciences and Human Development in Health Academic Unit, University of Southampton Faculty of Medicine, Southampton, UK; The David Hide Asthma and Allergy Research Centre, St Mary’s Hospital, Isle of Wight, UK; A.H. Zwinderman, Academic Medical Centre, University of Amsterdam.

The U-BIOPRED consortium wishes to acknowledge the help and expertise of the following individuals and groups without whom, the study would not have been possible.

Investigators and contributors: Nora Adriaens, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands; Antonios Aliprantis, Merck Research Laboratories, Boston, USA; Kjell Alving, Dept Women’s and Children’s Health, Uppsala University, Uppsala, Sweden; Per Bakke, Department of Clinical Science, University of Bergen, Bergen, Norway; David Balgoma, Centre for Allergy Research, Karolinska Institutet, Stockholm, Sweden; Clair Barber, NIHR Southampton Respiratory Biomedical Research Unit and Clinical and Experimental Sciences, Southampton, UK; Frédéric Baribaud, Janssen R&D, USA; Stewart Bates, Respiratory Therapeutic Unit, GSK, London, UK; An Baumanns, MSD, Brussels, Belgium; Jorge Beleta, Almirall S.A., Barcelona, Spain; Grazyna Bochenek, II Department of Internal Medicine, Jagiellonian University Medical College, Krakow, Poland; Joost Brandsma, University of Southampton, Southampton, UK; Armin Braun, Fraunhofer Institute for Toxicology and Experimental Medicine, Hannover, Germany; Dominic Burg, Centre for Proteomic Research, Institute for Life Sciences, University of Southampton, Southampton, UK; Leon Carayannopoulos, previously at: MSD, USA; João Pedro Carvalho da Purificação Rocha, Royal Brompton and Harefield NHS Foundation Trust, London, UK; Romans Chaleckis, Centre of Allergy Research, Karolinska Institutet, Stockholm, Sweden; Arnaldo D’Amico, University of Rome ‘Tor Vergata’, Rome Italy; Jorge De Alba,
Asthma UK, London, UK; Maria Mikus, Science
Jane Martin, NIHR Southampton Respiratory
Annemiek Dijkhuis, Academic Medical Centre,
Amsterdam, Amsterdam, The Netherlands; Inge De Lepeleire, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands; Aleksandra Draper, BioSci Consulting, Maasmechelen, Belgium; Jessica Edwards, Asthma UK, London, UK; Rosalia Emma, Department of Clinical and Experimental Medicine, University of Catania, Catania, Italy; Magnus Ericsson, Karolinska University Hospital, Stockholm, Sweden; Breda Flood, European Federation of Allergy and Airways Diseases Patient’s Associations, Brussels, Belgium; Hector Gallart, Centre for Allergy Research, Karolinska Institutet, Stockholm, Sweden; Cristina Gomez, Centre for Allergy Research, Karolinska Institutet, Stockholm, Sweden; Kenny Gove, NIHR Southampton Respiratory Biomedical Research Unit and Clinical and Experimental Sciences, Southampton, UK; Neil Gozzard, UCB, Slough, UK; John Haughney, Imperial Primary Care Respiratory Group, Aberdeen, Scotland; Lorraine Hewitt, NIHR Southampton Respiratory Biomedical Research Unit, Southampton, UK; Jens Hohlfeld, Fraunhofer Institute for Toxicology and Experimental Medicine, Hannover, Germany; Cecile Holweg, Respiratory and Allergy Diseases, Genentech, San Francisco, USA; Richard Hu, Amgen Inc. Thousand Oaks, USA; Sile Hu, National Heart and Lung Institute, Imperial College, London, UK; Juliette Kamphuis, Longfonds, Amersfoort, The Netherlands; Erika J. Kennington, Asthma UK, London, UK; Dyson Kerry, CromSource, Stirling, UK; Hugo Knobel, Philips Research Laboratories, Eindhoven, The Netherlands; Johan Kolmert, Centre for Allergy Research, Karolinska Institutet, Stockholm, Sweden; Maxim Kots, Chiesi Pharmaceuticals, SPA, Parma, Italy; Scott Kuo, National Heart and Lung Institute, Imperial College, London, UK; Maciej Kupczyn, Centre for Allergy Research, Karolinska Institutet, Stockholm, Sweden; Bart Lambrech, University of Gent, Gent, Belgium; Saeeda Lone-Latif, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands; Matthew J. Loza, Janssen R&D, USA; Lisa Marouzet, NIHR Southampton Respiratory Biomedical Research Unit, Southampton, UK; Jane Martin, NIHR Southampton Respiratory Biomedical Research Unit, Southampton, UK; Sarah Masefield, European Lung Foundation, Sheffield, UK; Caroline Mathon, Centre of Allergy Research, Karolinska Institutet, Stockholm, Sweden; Sally Meah, National Heart and Lung Institute, Imperial College, London, UK; Andrea Meiser, Data Science Institute, Imperial College, London, UK; Leanne Metcalf, previously at: Asthma UK, London, UK; Maria Mikus, Science for Life Laboratory and The Royal Institute of Technology, Stockholm, Sweden; Montse Miralpeix, Almirall, Barcelona, Spain; Philip Monk, Synairgen Research Ltd, Southampton, UK; Shama Naz, Centre for Allergy Research, Karolinska Institutet, Stockholm, Sweden; Ben Nicholas, University of Southampton, Southampton, UK; Peter Nilsson, Science for Life Laboratory and The Royal Institute of Technology, Stockholm, Sweden; Jörgen Östling, AstraZeneca, Mölndal, Sweden; Antonio Pacino, Lega Italiano Anti Fumo, Catania, Italy; Susanna Palkonen, European Federation of Allergy and Airways Diseases Patient’s Associations, Brussels, Belgium; Stelios Pavlidis, National Heart and Lung Institute, Imperial College, London, UK; Giorgio Pennazza, University of Rome ‘Tor Vergata’, Rome Italy; Anne Petren, Centre for Allergy Research, Karolinska Institutet, Stockholm, Sweden; Sandy Pink, NIHR Southampton Respiratory Biomedical Research Unit, Southampton, UK; Anthony Postle, University of Southampton, UK; Pippa Powell, European Lung Foundation, Sheffield, UK; Malaya Rahman-Amin, Previously at: Asthma UK, London, UK; Navin Rao, Janssen R&D, USA; Lara Ravaneetti, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands; Emma Ray, NIHR Southampton Respiratory Biomedical Research Unit, Southampton, UK; Stacey Reinke, Centre for Allergy Research, Karolinska Institutet, Stockholm, Sweden; Leanne Reynolds, previously at: Asthma UK, London, UK; John Riley, Respiratory Therapeutic Unit, GSK, London, UK; Martine Robberechts, MSD, Brussels, Belgium; Amanda Roberts, Asthma UK, London, UK; Kirsty Russell, National Heart and Lung Institute, Imperial College, London, UK; Michael Rutgers, Longfonds, Amersfoort, The Netherlands; Marco Santoninco, University of Rome ‘Tor Vergata’, Rome Italy; Corinna Schoelch, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany; James P.R. Schofield, Centre for Proteomic Research, Institute for Life Sciences, University of Southampton, Southampton, UK; Marcus Sjödin, Centre for Allergy Research, Karolinska Institutet, Stockholm, Sweden; Paul J. Skipp, Centre for Proteomic Research, Institute for Life Sciences, University of Southampton, Southampton, UK; Barbara Smids, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands; Caroline Smith, NIHR Southampton Respiratory Biomedical Research Unit, Southampton, UK; Jessica Smith, Asthma UK, London, UK; Katherine M. Smith, University of Nottingham, UK; Doroteya Staykova, University of Southampton, Southampton, UK; Kai Sun, Data Science Institute, Imperial College, London, UK; John-Olof Thörngren, Karolinska University Hospital,
Stockholm, Sweden; Bob Thornton, MSD, USA; Jonathan Thorsen, COPSAC, Copenhagen
Prospective Studies on Asthma in Childhood,
Herlev and Gentofte Hospital, University of
Copenhagen, Copenhagen, Denmark; Marianne
van de Pol, Academic Medical Centre, University
of Amsterdam, Amsterdam, The Netherlands;
Marleen van Geest, AstraZeneca, Mölndal,
Sweden; Jenny Versnel, previously at: Asthma UK,
London, UK; Anton Vink, Philips Research
Laboratories, Eindhoven, The Netherlands; Frans
Wald, Boehringer Ingelheim Pharma GmbH & Co.
KG, Biberach, Germany; Samantha Walker,
Asthma UK, London, UK; Jonathan Ward,
Histochemistry Research Unit, Faculty of
Medicine, University of Southampton,
Southampton, UK; Zsoka Weiszhart, Semmelweis
University, Budapest, Hungary; Kristiane Wetzel,
Boehringer Ingelheim Pharma GmbH, Biberach,
Germany; Craig E. Wheelock, Centre for Allergy
Research, Karolinska Institutet, Stockholm,
Sweden; Coen Wiegman, National Heart and
Lung Institute, Imperial College, London, UK;
Siân Williams, International Primary Care
Respiratory Group, Aberdeen, Scotland; Susan J.
Wilson, Histochemistry Research Unit, Faculty of
Medicine, University of Southampton,
Southampton, UK; Ashley Woodcock, Centre for
Respiratory Medicine and Allergy, Institute of
Inflammation and Repair, University of
Manchester and University Hospital of South
Manchester, Manchester Academic Health
Sciences Centre, Manchester, UK; Xian Yang, Data
Science Institute, Imperial College, London, UK;
Elizabeth Yeyasingham, UK Clinical Operations,
GSK, Stockley Park, UK.

Partner organisations: Novartis Pharma AG;
University of Southampton, Southampton, UK;
Academic Medical Centre, University of
Amsterdam, Amsterdam, The Netherlands;
Imperial College London, London, UK; University
of Catania, Catania, Italy; University of Rome ‘Tor
Vergata’, Rome, Italy; Hvidore Hospital, Hvidore,
Denmark; Jagiellonian Univ. Medi.College,
Crakow, Poland; University Hospital, Inselspital,
Bern, Switzerland; Semmelweis University,
Budapest, Hungary; University of Manchester,
Manchester, UK; Université d’Aix-Marseille,
Marseille, France; Fraunhofer Institute, Hannover,
Germany; University Hospital, Umea, Sweden;
Ghent University, Ghent, Belgium; Ctr. Nat.
Recherche Scientifique, Villejuif, France;
Università Cattolica del Sacro Cuore, Rome, Italy;
University Hospital, Copenhagen, Denmark;
Karolinska Institutet, Stockholm, Sweden;
Nottingham University Hospital, Nottingham,
UK; University of Bergen, Bergen, Norway;
Netherlands Asthma Foundation, Leusden, NL;
European Lung Foundation, Sheffield, UK;
Asthma UK, London, UK; European Fed. of
Allergy and Airways Diseases Patients’
Associations, Brussels, Belgium; Lega Italiano Anti
Fumo, Catania, Italy; International Primary Care
Respiratory Group, Aberdeen, Scotland; Philips
Research Laboratories, Eindhoven, NL; Synairgen
Research Ltd, Southampton, UK; Aerocrine AB,
Stockholm, Sweden; BioSci Consulting,
Maasmechelen, Belgium; Almirall; AstraZeneca;
Boehringer Ingelheim; Chiesi; GlaxoSmithKline;
Roche; UCB; Janssen Biologics BV; Amgen NV;
Merck Sharp & Dohme Corp.

Third Parties to the project, contributing to the
clinical trial: Academic Medical Centre (AMC),
Amsterdam (In the U-BIOPRED consortium the
legal entity is AMC Medical Research BV (AMR);
AMR is a subsidiary of both AMC and the
University of Amsterdam; AMC contribute across
the U-BIOPRED project); University Hospital
Southampton NHS Trust (third party of the
University of Southampton and contributor to the
U-BIOPRED clinical trial); South Manchester
Healthcare Trust (third party to the University of
Manchester, South Manchester Healthcare Trust,
contributor to the U-BIOPRED clinical trial and to
the U-BIOPRED Biobank); Protisvalor
Méditerranée SAS (third party to University of the
Mediterranean; contributor to the U-BIOPRED
clinical trial); Karolinska University Hospital
(third party Karolinska Institutet (KI), contributor
to the U-BIOPRED clinical trial); Nottingham
University Hospital (third party to University of
Nottingham, contributor to the U-BIOPRED
clinical trial); NIHR-Wellcome Trust Clinical
Research Facility.

Members of the ethics board: Jan-Bas Prins,
biochemical research, LUMC, the Netherlands;
Martina Gahlemann, clinical care, BI, Germany;
Luigi Visintin, legal affairs, LIAF, Italy; Hazel
Evans, paediatric care, Southampton, UK; Martine
Puhl, patient representation (co-chair), NAF, the
Netherlands; Lina Buzermaniene, patient
representation, EFA, Lithuania; Val Hudson,
patient representation, Asthma UK; Laura Bond,
patient representation, Asthma UK; Pim de Boer,
patient representation and pathobiology, IND;
Guy Widdershoven, research ethics, VUMC, the
Netherlands; Ralf Sigmund, research methodology
and biostatistics, BI, Germany.

The patient input platform: Amanda Roberts, UK;
David Supple (chair), UK; Dominique
Hamerlijnck, The Netherlands; Jenny Negus, UK;
Juliette Kamphuis, The Netherlands; Lehanne
Sergison, UK; Luigi Visintin, Italy; Pim de Boer
(co-chair), The Netherlands; Susanne Onstein,
The Netherlands.
Members of the safety monitoring board: William MacNee, clinical care; Renato Bernardini, clinical pharmacology; Louis Bont, paediatric care and infectious diseases; Per-Ake Wecksell, patient representation; Pim de Boer, patient representation and pathology (chair); Martina Gahlemann, patient safety advice and clinical care (co-chair); Ralf Sigmund, bio-informatician.

This work was partially funded by the Engineering and Physical Sciences Research Council, UK (EP/N014189: Joining the Dots, from Data to Insight).

Instrumentation in the Centre for Proteomic Research is supported by the BBSRC (BM/M012387/1) and the Wessex Medical Trust.

We thank Ayasdi Inc. for use of, and support with, the Ayasdi TDA software.

Data Availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD005949 and 10.6019/PXD005949.

REFERENCES


Table of Content Graphic

HDMS

SPUTUM