Journal of proteome research

Article

Cite This: J. Proteome Res. 2018, 17, 2072−2091

pubs.acs.org/jpr

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,[2] Rene Lutter,[7,8] Louise Fleming,[9] Dominick E. Shaw,[10] Per S. Bakke,[11] 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,[4] Charles Auffray,[22] Ana R. Sousa,[23] Ian M. Adcock,[24] Kian Fan Chung,[9] Peter J. Sterk,[7] Ratko Djukanović,[2,‡] Paul J. Skipp,[1,‡] and the U-BIOPRED Study Group[§]

[1]Centre for Proteomic Research, Biological Sciences, University of Southampton, Southampton SO17 1BJ, U.K.

[2]NIHR Southampton Biomedical Research Centre, Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Southampton SO16 6YD, U.K.

[3]Acclarogen Ltd., Cambridge CB4 0WS, U.K.

[4]Data Science Institute, Imperial College London, London SW7 2AZ, U.K.

[5]Janssen Research & Development, Buckinghamshire HP12 4DP, U.K.

[6]Areteva Ltd., Nottingham NG7 6LB, U.K.

[7]AMC, Department of Experimental Immunology, University of Amsterdam, 1012 WX Amsterdam, The Netherlands

[8]AMC, Department of Respiratory Medicine, University of Amsterdam, 1012 WX Amsterdam, The Netherlands

[9]Airways Disease, National Heart and Lung Institute, Imperial College, London & Royal Brompton NIHR Biomedical Research Unit, London SW7 2AZ, United Kingdom

[10]Respiratory Research Unit, University of Nottingham, Nottingham NG7 2RD, U.K.

[11]Institute of Medicine, University of Bergen, 5007 Bergen, Norway

[12]Department of Clinical and Experimental Medicine Hospital University, University of Catania, 95124 Catania, Italy

[13]The Centre for Allergy Research, The Institute of Environmental Medicine, Karolinska Institutet, SE-171 77 Stockholm, Sweden

[14]Respiratory and Allergy Research Group, University of Manchester, Manchester M13 9PL, U.K.

[15]Department of Respiratory Medicine, Academic Medical Centre, University of Amsterdam, 1012 WX Amsterdam, The Netherlands

[16]Department of Pulmonology, Semmelweis University, Budapest 1085, Hungary

[17]Fraunhofer Institute for Toxicology and Experimental Medicine Hannover, 30625 Hannover, Germany

[18]Faculty of Medicine, Catholic University of the Sacred Heart, 00168 Rome, Italy

[19]Laboratory of Molecular Biology and Clinical Genetics, Medical College, Jagiellonian University, 31-007 Krakow, Poland

[20]Department of Medicine, Department of Public Health and Clinical Medicine Respiratory Medicine Unit, Umeå University, 901 87 Umeå, Sweden

[21]University Children's Hospital Zurich, 8032 Zurich, Switzerland

[22]European Institute for Systems Biology and Medicine, CNRS-ENS-UCBL-INSERM, Université de Lyon, 69007 Lyon, France

[23]Respiratory Therapeutic Unit, GSK, Stockley Park, Uxbridge UB11 1BT, U.K.

[24]Cell and Molecular Biology Group, Airways Disease Section, National Heart and Lung Institute, Imperial College London, Dovehouse Street, London SW3 6LR, U.K.

Supporting Information

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)

continued...

Received: January 10, 2018
Published: May 8, 2018

DOI: 10.1021/acs.jproteome.8b00018
J. Proteome Res. 2018, 17, 2072−2091

international project. We present practical and analytical techniques to optimize the detection of robust biomarkers in proteomic studies. The normal sputum proteome was derived using data-independent HDMS^E applied to 40 healthy nonsmoking 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 1666 proteins. Quality control procedures were developed to optimize the accuracy and consistency of measurement of sputum proteins and analyze the distribution of sputum proteins in the healthy population. The analysis showed that quantitation of proteins by HDMS^E 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 interindividual variance. We also highlight that the sputum proteome of healthy individuals is related to sputum neutrophil levels, but not gender or allergic sensitization. We illustrate the importance of design and interpretation of disease biomarker studies considering such protein population and technical measurement variance.

KEYWORDS: *asthma, proteomics, biomarkers, U-BIOPRED, sputum, HDMS^E, 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[1−7] 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),[8,9] providing insight into determinants of disease severity[9] and relevant pathophysiological abnormalities, such as airway hyperresponsiveness[10] and changes in airway geometry.[11] 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.[12−17] 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.[2] Since then, there have been a number of reports of this approach in COPD[18,19] including a study highlighting the utility of protein network analysis in sputum,[20] and a large study combining proteomic and transcriptomic analyses.[21] Likewise, limited studies of sputum have been performed to study asthma, and some have been relatively low throughput.[22,23] Apart from one study, by Titz et al.,[21] 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.[25−29] 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,[30] particularly with the advent of data independent approaches such as MS^E,[31,32] which allow absolute comparison of samples without the inherent limitations associated with multiplexing, labeling or spectral counting-based quantitation,.[33−35] Despite the utility of MS^E 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[36,37] 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^E analysis to a large set of sputum samples to advance on the sputum proteome previously reported.[2] 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 sensitization to common airborne allergens (e.g., house dust mite and pollens) in the general population, we also examined how atopy, defined by sensitization 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 centers with extensive experience in sputum induction and processing. The clinical study has been described elsewhere[39] and the protocol was approved by all local Ethics Review Boards. Participants gave their written informed consent for extensive characterization using routine clinical protocols, including lung function tests, assessment of sensitization to common aero-allergens, and hematological and biochemistry blood tests (reported in detail in Shaw et al., 2015[38]). Samples were stored in a central biobank (CIGMR Biobank, University of Manchester)

DOI: 10.1021/acs.jproteome.8b00018
J. Proteome Res. 2018, 17, 2072−2091

where they were blinded. Identity of the samples were unblinded only after all the mass spectrometric analyses and data preprocessing had been completed.

Forty healthy individuals (mean age 36.9 years, range 18−65, 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 nonatopic 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 ($p < 5 \times 10^{-5}$), but, surprisingly, serum LDH was also slightly higher in the atopic participants ($p = 0.01$) while alkaline phosphatase was lower ($p = 0.01$).

Table 1. Demographics and Sputum Cell Characteristics

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

Sputum Induction and Processing

Sputum induction with nebulized 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.[39] Uniformity of methods was ensured by all study sites using standard operating procedures (SOP) and centralized training. For consistency required for comparison with proteomes in patients with disease (asthma or COPD), all participants were premedicated 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 solubilized at room temperature with 6.8 mM 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*g* to remove the cell pellet, further centrifuged at 12 000*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 semiautomated 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 antioxidant (5 mg mL^{-1} butylated hydroxytoluene in MeOH). Samples were centrifuged at 1000*g* for 10 min at 10 °C to produce protein pellets which were snap-frozen in liquid nitrogen and stored at −80 °C.

In preparation for analysis, the frozen protein pellets were thawed to room temperature, dissolved in 150 μL of 50% trifluoroethanol, 50 mM ammonium bicarbonate and heated at 60 °C for 30 min. A pool for quality control was prepared with equal protein amounts from 40 different sputum samples (including healthy participants and participants with a diagnosis of asthma). Pool samples were processed and analyzed in parallel in batches containing 11 analytical samples and one pool. Dissolved protein pellets were reduced, alkylated and digested with trypsin. Peptide samples were filtered using a 10 KDa cutoff ultrafiltration device (Millipore) and the filtrate lyophilized in vacuo. Samples were dissolved in 3% acetonitrile (ACN), 0.1% trifluoroacetic acid (TFA) in preparation for reverse phase cleanup, performed according to the manufacturer's instructions using C18 spin tips (Protea Biosciences). Following elution, peptides were lyophilized and stored at −80 °C prior to analysis.

Serum Collection and Processing

Clotted venous blood samples were centrifuged at 1000*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 lyophilized. Peptide extracts were then resuspended in 3% ACN, 0.1% TFA and desalted using 96 well RP solid phase extraction plates (3 M Empore). Eluates were transferred to separate microcentrifuge tubes, lyophilized and stored on ice until mass spectrometry.

Mass Spectrometry

Peptide extracts were resuspended 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 Hi3 standard (Waters), yeast enolase (ENO) and yeast alcohol dehydrogenase (ADH) was added to a final concentration in 20 μL of 250 ng/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 analyzed in duplicate, sequentially (not spread across batches), via HDMS^E on a Waters Synapt G2S high definition mass spectrometer coupled to a nanoAcquity UPLC system. 4 μL of peptide extract was injected onto a C18 BEH trapping column (Waters) and washed with buffer A for 5 min at 5 μL/min. Peptides were separated using a 25 cm T3 HSS C18 analytical column (Waters) with a linear gradient of 3−50% ACN + 0.1% formic acid over 50 min at a flow rate of 0.3 μL/min. Eluted samples were sprayed directly into the mass spectrometer operating in MS^E 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 5 V and elevated CE ramp from 15 to 40 V. Ion mobility separation was implemented prior to fragmentation using a wave velocity of 650 m/s and wave height of 40 V. The lock mass Glu-fibrinopeptide, $(M + 2H)^{+2}$, m/z = 785.8426) was infused at a

DOI: 10.1021/acs.jproteome.8b00018
J. Proteome Res. 2018, 17, 2072−2091

concentration of 100 fmol/μL at a flow rate of 250 nL/min and acquired every 60 s.

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 data set 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%.[40] 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.[32,41] The ion accounting output files[42] 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 Normalization

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 normalizing 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 normalization"). ComBat was used to adjust for batch to batch variation.[43]

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/infernordn).[44] 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 data set.

Samples were analyzed 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 visualized 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 to 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 data set, while minimizing reductions in proteome coverage. However, we recognize that there remains increased uncertainty in the measurement of proteins treated in this way, and the issue may be minimized 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),[45] the HUPO[46,47] plasma reference set, the reference plasma data set from the laboratory of Matthias Mann,[48] sputum proteomes,[2,21,24] proteomes of BAL fluid,[36,37,49−52] exhaled breath condensate,[53,54] pure airway mucus,[55] saliva,[56−61] macrophage proteomes,[62,63] eosinophil proteomes,[64−68] whole neutrophils,[69−72] investigations of neutrophil extracellular traps,[16,73,74] neutrophil microparticles,[75] and neutrophil granules.[76,77] Data were tabulated and identifiers converted to Uniprot format using the Uniprot ID mapping service (http://www.uniprot.org/mapping/) or via DAVID.[78,79] Redundant, discontinued, merged and incomplete entries (e.g., assignments to protein fragments, pseudogenes, or to nonhuman proteomes) were either disregarded or were assigned to Uniprot identifiers.

Statistical Analyses and Informatics

Statistical analyses were performed in R,[80] Inferno RDN,[44] Microsoft Excel and Minitab,[13] using parametric or nonparametric tests as appropriate. Coefficient of variation was calculated for log-normal distributions. Visualization was performed in Origin 9.1 (http://www.originlab.com/91), R, and Inforsense (IDBS). Venn diagrams were generated using Venny (http://bioinfogp.cnb.csic.es/tools/venny/).[81] Tree maps were drawn using Treemap 4.1.2 (http://www.cs.umd.edu/hcil/treemap/).[82] Pathway analysis, functional enrichment analysis and biological inference were performed using Ingenuity Pathway Analysis software (IPA, QIAGEN Redwood City). The IPA analyses were

DOI: 10.1021/acs.jproteome.8b00018
J. Proteome Res. 2018, 17, 2072−2091

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,[83] and the Secreted Protein Database[84] was used to identify secreted proteins. Protein−protein interactions were explored using String (http://string-db.org/).[85] Ontology was annotated from retrieval via Uniprot,[86] and retrieval and enrichment analysis was performed via GOrilla (http://cbl-gorilla.cs.technion.ac.il/)[87] against the *Homo sapiens* proteome and REVIGO (http://revigo.irb.hr/).[88] The data were clustered by topological data analysis (TDA).[89−93] 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,[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, variancePartition, was used to identify the drivers of protein measurement variation.[94−96]

RESULTS

Identified Proteins

A total of 4182 proteins were identified in the sputum in ≥1 individual(s): 2354 proteins in ≥2 individual(s), 284 proteins in ≥40% and 73 in ≥90% of individuals (Supplementary Figure S1). High abundance proteins were generally more frequently identified (Figure 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 S1). 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 cutoff was defined at ≥40%, since at this frequency of identification, the frequency vs protein rank curve was close to the point of inflection (Figure 1), where even a slight increase in the frequency of identification "cut off", significantly increased the sparsity of the data set and, hence, the total number of missing values. We also defined an "extended healthy sputum proteome" data set consisting of 1666 proteins identified in ≥3 individuals.

Impact of Gender, Age, Atopy and Granulocyte Counts

No significant differences in proteomes were observed when comparing age, atopic and nonatopic individuals or males and females (Figure 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 nonatopic individuals (q < 0.05). The network shown in Figure S2 is constructed using multidimensional scaling (MDS) lenses (similarity metric) projected onto a TDA network,

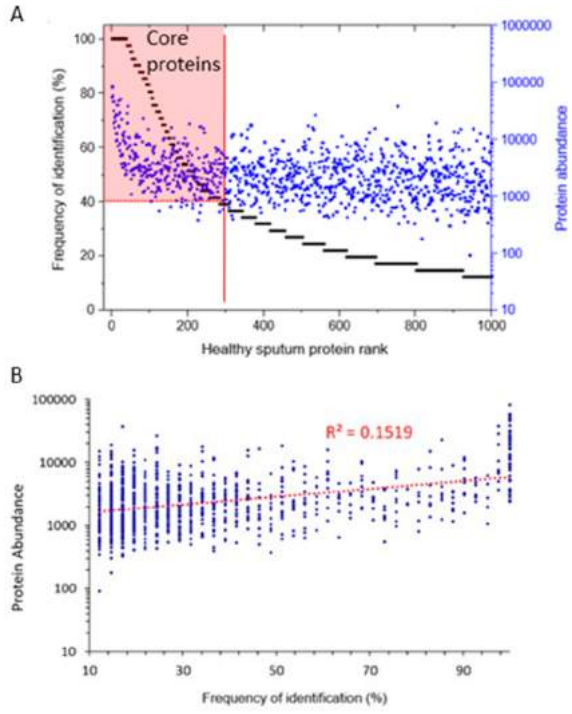


Figure 1. Defining the "core" sputum proteome from the relationship of abundance of identified proteins and frequency of identification across samples. (A) The 40% frequency of detection used as the cutoff for the "core" sputum proteome. The relationship between rank of frequency of detection and the 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 data set and hence the total number of missing values. (B) The intensity of protein measurement correlated weakly with the frequency of protein identification.

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 R^2 correlation scores are shown in Figure 2). The average R^2 for correlation of proteins with neutrophil count was 0.07 and 91% of proteins had an R^2 < 0.2.

Protein Variability, Intensity Adjustment and Measurement Accuracy

Protein measurements between samples were visualized in heatmaps before and after intensity adjustment and after batch effect correction (Figure 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 visualized in nonadjusted maps

DOI: 10.1021/acs.jproteome.8b00018
J. Proteome Res. 2018, 17, 2072−2091

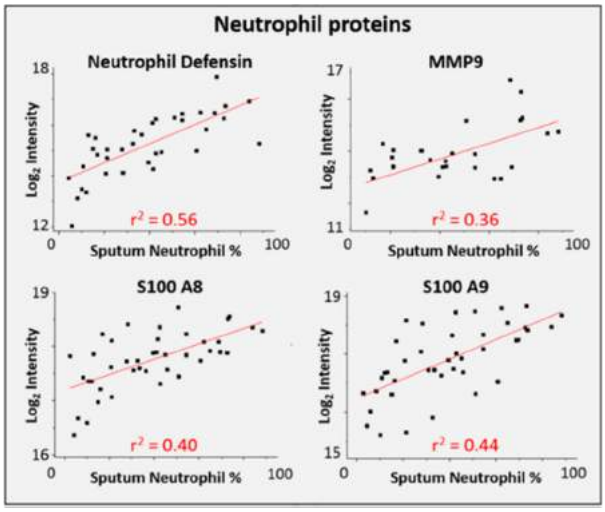


Figure 2. Neutrophil proteins and neutrophil counts. We observed relationships between neutrophil granule proteins and the proportion of neutrophils in sputum. Neutrophil proteins correlated with percentage sputum neutrophil cell counts.

(Figure S3A) were corrected by normalizing 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 (Figure S3B). The top90 adjustment corrected for protein intensity variation more effectively than normalization to "total intensity"; avoiding effects of rarely measured protein abundances. ComBat was used to adjust for batch to batch variation (Figure S3C).

High sample to sample variability was observed in the sputum proteomes. Pooled sputum sample replicates, which were processed and analyzed at regular intervals throughout the acquisition of healthy sputum data, were compared to the participant sputum samples and the matched healthy serum data set (Figure 4). Compared to individual sputum samples, pooled samples contained higher numbers of proteins with a lower percentage Coefficient of Variation (CV %). In all data sets, 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 (Figure 3B).

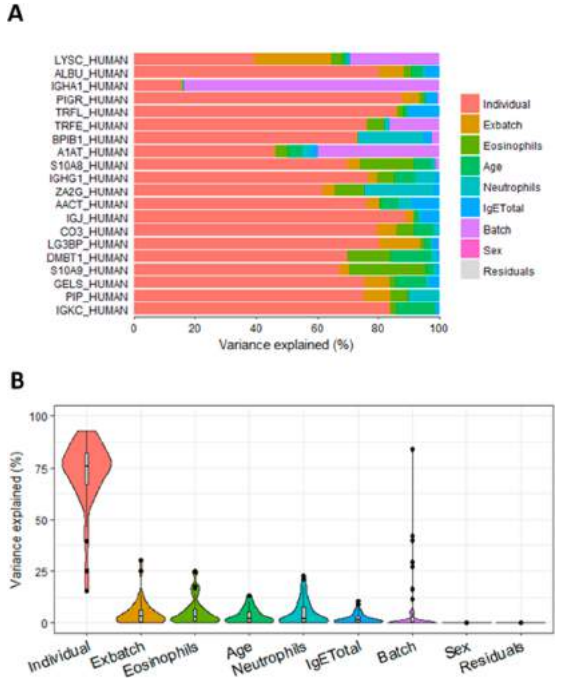


Figure 3. Sources of variability in protein measurement estimated by variancePartition. (A) The fraction of total variation in measurement of the top 20 proteins by highest coverage in samples, attributable to individual, extraction batch, eosinophil cell count, age, neutrophil cell count, total measured 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 interindividual 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,[97] A1AT[98]), or are likely the result of salivary contamination of samples (e.g., Amylase[99−123]).

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 8.1% (1632 of 20 412 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).

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 behavior of their peptides in HDMS[E] or errors in database

DOI: 10.1021/acs.jproteome.8b00018
J. Proteome Res. 2018, 17, 2072−2091

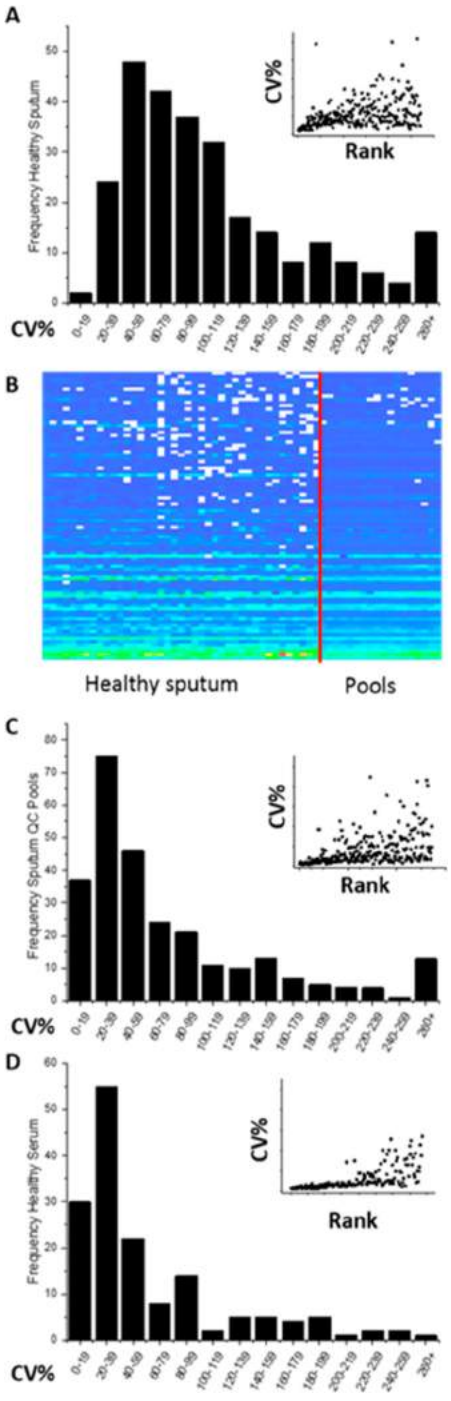


Figure 4. Variability in protein abundance measurements across samples. Frequency histograms represent: on the *x* axis CV % in increments of bin size 20, and inset scatter plots show CV % vs protein rank (proteins were ranked by order of abundance and frequency of identification across samples). Heatmap (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.

searching and quantification (e.g., due to homologues or protein to protein ambiguity) (Supplementary Table S3).

The variability in quantitation was nonuniform throughout the data set; 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 S2). The database was used to identify highly cited respiratory proteins associated with disease, and then, in a posthoc 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 subset of these proteins are presented in Figure 5. Low variability in the pooled samples, but high

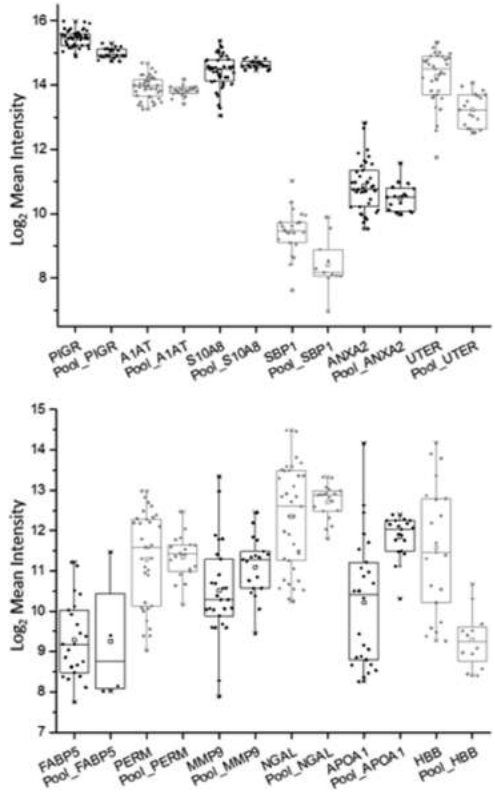


Figure 5. Distribution and variation in measurement of potential inflammatory biomarkers in the healthy population compared to pooled controls. Proteins showed varying levels of spread indicating that the sample size required for statistical power will vary significantly per analyte, with the top graph showing those proteins with lower variability and the bottom, those with higher variability. Note that the pooled samples were taken from asthmatics and nonasthmatics, 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 subclinical inflammatory effects in healthy participants.

variability in the population, indicated that the majority of variability arose from interindividual 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.

DOI: 10.1021/acs.jproteome.8b00018
J. Proteome Res. 2018, 17, 2072−2091

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 Figure 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

		1.5-fold change		2-fold change	
protein	SD ($\log_2(x)$)	P0 95%	P0 97.5%	P0 95%	P0 97.5%
PIGR	0.264	9	10	5	6
BPIB1	0.303	11	12	6	6
A1AT	0.349	14	15	7	7
LYSC	0.352	14	15	7	7
ACTB	0.379	16	17	7	8
TRFL	0.478	23	26	10	11
S10A8	0.525	27	30	11	13
S10A9	0.609	36	40	14	16
CO3	0.631	38	43	15	17
SBP1	0.711	48	53	18	20
VTDB	0.714	48	53	18	20
CFAH	0.743	52	58	20	22
ANXA2	0.776	56	63	21	24
PRDX1	0.804	60	67	23	25
UTER	0.825	63	70	24	26
PEDF	0.879	72	80	26	29
B2MG	0.882	72	80	27	30
FABP5	0.956	84	94	31	34
PERM	1.166	124	137	44	49
DEF1	1.168	124	138	44	49
MMP9	1.201	131	146	47	52
NGAL	1.307	155	172	55	61
APOA1	1.588	227	252	80	88
HBB	1.609	233	259	82	91

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.,[24] Nicholas et al.,[2] and Titz et al.[21] (Figure 6A,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.,[24] 81% of the 171 the proteins identified by Nicholas et al.,[2] but only 31% of the 2178

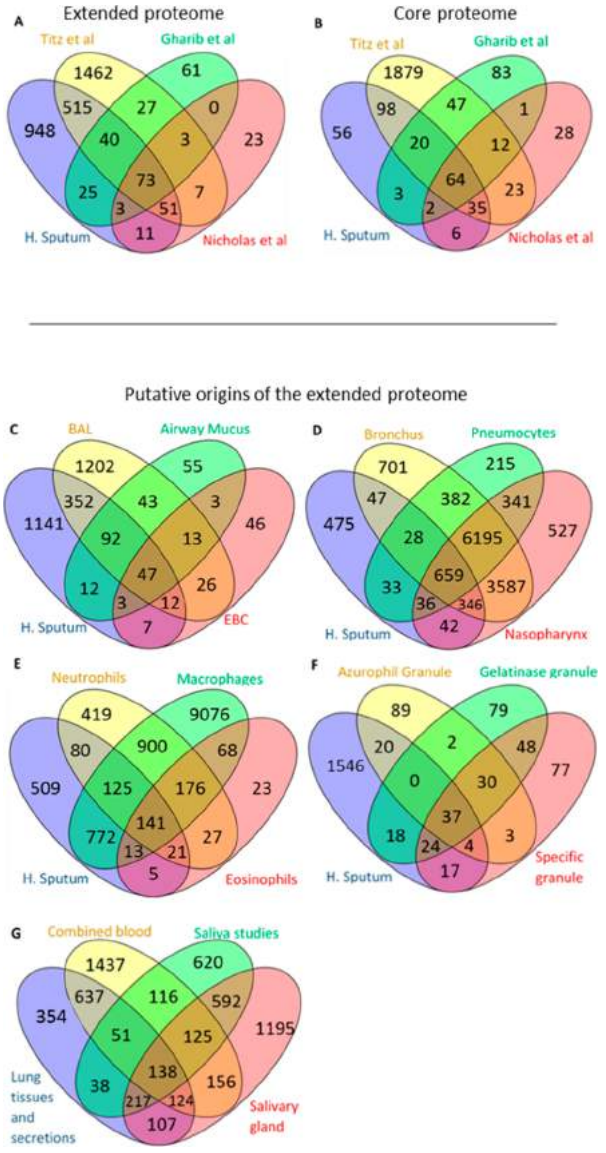


Figure 6. Comparison of coverage of U-BIOPRED sputum with other studies. Including other sputum studies (A,B), other proteomic studies of airway fluids and secretions (C), and respiratory tissues found in the protein atlas normal expression database (D). U-BIOPRED (Healthy) sputum proteome was compared 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).

proteins found in the study by Titz et al.,[21] 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[E].

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

DOI: 10.1021/acs.jproteome.8b00018
J. Proteome Res. 2018, 17, 2072−2091

proteome coverage were also observed when comparing airway mucus,[55] exhaled breath condensate (EBC)[53,54] and studies analyzing bronchial alveolar lavage (BAL)[36,37,49−52] (Figure 6C). 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[45] (Figure 6D). There was also overlap in the identifications of measured proteins to those of proteomes measured for eosinophils, macrophages and granulocyte (Figure 6E) and neutrophil granules (Figure 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 Figure S6). 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 that are harder to detect.

To further understand the protein composition of sputum, we also investigated possible systemic or nonrespiratory origins of our identified proteins through comparison to other relevant biofluids. We identified ∼50% overlap with proteomes of sputum, the upper airway, saliva and blood, similar to that reported previously[2,61] (Figure 6G).

Functional Analysis

Subcellular localization of proteins was analyzed using Ingenuity Pathways Analysis (IPA) (Figure 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.

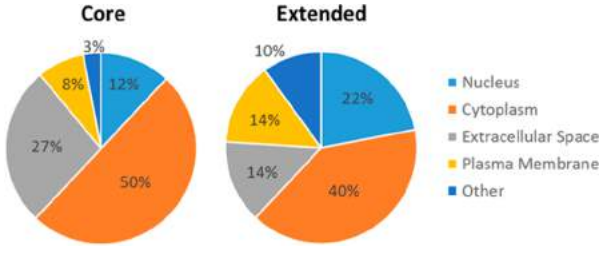


Figure 7. Subcellular localization prediction results from IPA analyses of the proteins from the core and extended sputum proteome against a background gene set restricted to Ingenuity Knowledge Base genes from expected sources of sputum proteins.

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) signaling refers to the ILK-mediated control of cytoskeleton remodelling. eNOS Signaling 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 utilized REVIGO to collapse and summarize like terms as treemaps. The results of these GO enrichment analyses for the core sputum proteome are shown in Supplementary Figure 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 Figure S7D,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 color, providing a landscape overview of induced sputum. Enrichment was observed for vesicle related components; GORilla $q = 1.92 \times 10^{-7}$, with 111 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 neutrophils.[131] 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 organization (integrin linked kinase signaling, actin cytoskeleton signaling, leukocyte extravasation signaling), 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 homologue 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 data set (Supplementary Tables S4 and S5).

Investigating protein interaction networks (http://string-db.org) in our core sputum proteome data set (Supplementary Figure 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.

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

DOI: 10.1021/acs.jproteome.8b00018
J. Proteome Res. 2018, 17, 2072−2091

Table 3. Top 20 Enriched Canonical Pathways in the Core Sputum Proteome Set

ingenuity canonical pathways	enrichment *p* value	total number matched proteins
LXR/RXR Activation	3.16×10^{-12}	17
Glycolysis I	2.00×10^{-11}	9
Acute Phase Response Signaling	6.31×10^{-11}	17
FXR/RXR Activation	5.01×10^{-10}	15
Gluconeogenesis I	7.76×10^{-10}	8
Actin Cytoskeleton Signaling	2.69×10^{-07}	15
ILK Signaling	2.00×10^{-06}	13
Primary Immunodeficiency Signaling	3.89×10^{-06}	7
Pyruvate Fermentation to Lactate	1.45×10^{-05}	3
Complement System	1.70×10^{-05}	6
RhoGDI Signaling	2.95×10^{-05}	11
Hematopoiesis from Pluripotent Stem Cells	3.16×10^{-05}	6
Leukocyte Extravasation Signaling	3.89×10^{-05}	12
RhoA Signaling	3.89×10^{-05}	9
Clathrin-mediated Endocytosis Signaling	5.50×10^{-05}	11
Glucocorticoid Receptor Signaling	7.08×10^{-05}	14
Cell Cycle: G2/M DNA Damage Checkpoint Regulation	8.13×10^{-05}	6
Epithelial Adherens Junction Signaling	1.69×10^{-04}	9
Coagulation System	1.90×10^{-04}	5
eNOS Signaling	1.99×10^{-04}	9

airway lining fluid proteins and, thus, significantly advances on the sputum proteome we have previously reported.[2] The study has identified 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[2] identified 191 proteins in the sputum of a single individual using 2-dimensional gel electrophoresis and mass spectrometry, which is biased toward 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. Gharib et al., applied shotgun mass spectrometry to assess 5 healthy and 10 asthmatic participants,[24,25] and identified 254 proteins in all participants' sputum. Using a nonparametric test developed by the authors, called the spectral index,[26] 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.,[21] conducted an impressive study applying LC−MS/MS of tryptic peptides labeled with Tandem Mass Tags (TMT) to sputum samples of 216 participants equally composed of healthy nonsmokers, healthy never smokers, COPD patients and current smokers. 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 analyzed 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 nonsmokers.

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,[132−134] were not identified in our study, likely due to postdigestion 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 data set: technical reproducibility in measurement of each protein, interindividual 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 data set. 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 data sets in the current study, and as such, influenced our approach to developing our "Top-90 intensity normalization" 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 normalization strategies based upon total MS intensity. We have found that using the region with the greatest coverage for adjusting samples, the top90 method, is an effective method of normalization for protein load and intensity.

Comparison of variability across injection replicates and pooled QC samples within the healthy sputum proteome data set 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^{E}. 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 recognize 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

DOI: 10.1021/acs.jproteome.8b00018
J. Proteome Res. 2018, 17, 2072−2091

Table 4. Top 20 Enriched Functions and Diseases Found by IPA of the Core Sputum Proteome

disease and function category hierarchy	specific annotation	enrichment p value	number of matched proteins
Infectious Diseases, Inflammatory Disease, Respiratory Disease	Severe acute respiratory syndrome	1.92×10^{-09}	14
Infectious Diseases	Viral infection	4.17×10^{-09}	18
Inflammatory Response	Inflammatory response	4.40×10^{-08}	18
Connective Tissue Disorders, Inflammatory Disease, Skeletal and Muscular Disorders	Rheumatic Disease	9.30×10^{-08}	19
Connective Tissue Disorders, Inflammatory Disease, Skeletal and Muscular Disorders	Arthritis	1.30×10^{-07}	18
Cellular Movement, Immune Cell Trafficking	Leukocyte migration	2.22×10^{-07}	20
Cellular Movement, Hematological System Development and Function, Immune Cell Trafficking, Inflammatory Response	Cell movement of phagocytes	2.87×10^{-07}	16
Immunological Disease	Systemic autoimmune syndrome	3.05×10^{-07}	16
Inflammatory Disease	Chronic inflammatory disorder	5.92×10^{-07}	16
Cellular Movement, Hematological System Development and Function, Immune Cell Trafficking	Cell movement of leukocytes	6.79×10^{-07}	19
Connective Tissue Disorders, Immunological Disease, Inflammatory Disease, Skeletal and Muscular Disorders	Rheumatoid arthritis	1.02×10^{-06}	15
Cellular Movement, Hematological System Development and Function, Immune Cell Trafficking	Cell movement of myeloid cells	1.73×10^{-06}	15
Cell-to-Cell Signaling and Interaction, Hematological System Development and Function, Immune Cell Trafficking	Adhesion of immune cells	3.58×10^{-06}	13
Cellular Movement, Hematological System Development and Function, Immune Cell Trafficking, Inflammatory Response	Chemotaxis of phagocytes	5.26×10^{-06}	12
Cellular Movement, Hematological System Development and Function, Immune Cell Trafficking, Inflammatory Response	Chemotaxis of leukocytes	6.22×10^{-06}	14
Cellular Movement, Hematological System Development and Function, Immune Cell Trafficking, Inflammatory Response	Chemotaxis of myeloid cells	8.82×10^{-06}	12
Cellular Movement, Hematological System Development and Function, Immune Cell Trafficking	Cell movement of granulocytes	1.06×10^{-05}	11
Cellular Assembly and Organization	Formation of rosettes	1.45×10^{-05}	3
Organismal Survival	Organismal death	1.86×10^{-05}	7
Cellular Movement, Hematological System Development and Function, Immune Cell Trafficking, Inflammatory Response	Cell movement of neutrophils	2.08×10^{-05}	9

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 postprocessing errors. Many of these have potential homologues 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 interindividual variability was significant (Supplementary Table S1), in stark contrast to the low interindividual variability in serum from the same individuals.

Upon the basis of 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 interindividual 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 hemoglobin 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,[129,130] HBB may be a consequence of oral-derived blood contaminating the sputum. The other highly variable proteins, neutrophil gelatinase associated lipocalin, MMP9, and neutrophil defensin are all known to be involved in inflammation, and several of these seen in the current study correlated with sputum neutrophil numbers.

Protein biomarkers have been reported in previous respiratory studies, however, high variability in protein concentrations, in combination with small sample size[30] can result in false positive biomarker identifications. Additionally, as we have shown, the protein to protein differences in variability means that predicting sample size to sufficiently power biomarker investigations is difficult to achieve without extensive prior investigation.

DOI: 10.1021/acs.jproteome.8b00018
J. Proteome Res. 2018, 17, 2072−2091

These problems are likely strong contributors to the poor record of translating many biomarkers to the clinic.[27,28]

Confounding Factors in Sputum Analysis

Previous studies have shown intermediate levels of inflammation and even tissue remodelling in healthy atopic individuals relative to those with a clear diagnosis of asthma (Djukanovic ERJ). However, in the current study, there were no significant differences between atopic and nonatopic participants, 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,[124,125] but is often confounded by smoking status and steroid treatment, influencing neutrophil half-life, activity and migration.[126−128] A number of neutrophil proteins correlated with neutrophil cell counts (Figure 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 while 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 localization of the proteins from the core and extended sputum proteomes (Figure 7 and 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 defense, inflammatory responses via complement and acute phase proteins, phagocytic cells (macrophages and neutrophils) and, to a lesser extent, eosinophils.

A number of signaling 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.[133,135,136] Such regulatory pathways have complex roles in inflammatory biology: FXR inhibits inflammation,[137] while LXR agonists have been shown to increase airway reactivity and smooth muscle growth in an asthma model.[138] LXR has also been implicated in counter-regulation of toll like receptor induced inflammatory responses,[139,140−142] which are involved in inflammatory respiratory diseases.[146,147] We also observed an enrichment of the RhoA and RhoGDI signaling pathways, known to be involved in hyper-responsiveness in asthma,[143,144] and superoxide generation in macrophages,[145] respectively.

An interesting observation was the significant enrichment of vesicles and vesicular components in the sputum (Figure S7). These small membraned 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.[146] Secretory vesicles have recently become an area of interest for their potential pro and anti-inflammatory functions.[146] In response to nonspecific complement mediated inflammation neutrophils produce ectosomes that are coated in and loaded with proteins often associated with granules.[148] 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,[149] and eosinophils produce cytokine containing vesicles, e.g., degranulation, that may also be important for asthma (reviewed in Spencer et al.[150]).

Secretory vesicles contain high levels of cytoplasmic proteins.[77] For example, exosomes isolated in the BAL of asthmatics are enriched for inflammatory leukotriene production[149] 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 signaling,[151] and have been shown to be different between BAL samples of healthy and asthmatic individuals.[152]

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 defense 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 data set for proteomes from patients suffering from inflammatory lung disease.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.8b00018.

DOI: 10.1021/acs.jproteome.8b00018
J. Proteome Res. 2018, 17, 2072−2091

Figures S1−S8, Tables S1−S5 (PDF)
Proteins of the core and extended sputum proteomes (XLSX)
Database of potential respiratory biomarkers (XLSX)

AUTHOR INFORMATION

Corresponding Author

*E-mail: j.p.r.schofield@soton.ac.uk. Phone: +44 (0) 23 80594204.

ORCID

James P. R. Schofield: 0000-0003-4542-6614
Kian Fan Chung: 0000-0001-7101-1426

Author Contributions

†D. Burg and J.P.R. Schofield are equally contributing joint first authors.

Author Contributions

‡R. Djukanović and P.J. Skipp are joint senior authors.

Author Contributions

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

Notes

The authors declare no competing financial interest.
The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the data set identifier PXD005949 and DOI: 10.6019/PXD005949.

ACKNOWLEDGMENTS

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 center 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. Badorrek, Fraunhofer ITEM; S. Ballereau, European Institute for Systems Biology and Medicine, University of Lyon, France; F. Baribaud, Janssen R&D, USA; M.K. Batuwitage, Imperial College, London, UK; A. Bedding, Roche Diagnostics GmbH, 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, 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. Chaiboonchoe, 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 Instituet, 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, Ph.D., Project manager, Verona Italy; M. Gahlemann, Boehringer Ingelheim (Schweiz) GmbH; G. Galffy, 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, Stevenge, UK; J.R. Konradsen, Karolinska University Hospital and Karolinska Institutet; K. Kretsos, 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, Mölndal, 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

DOI: 10.1021/acs.jproteome.8b00018
J. Proteome Res. 2018, 17, 2072−2091

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. Rossios, Imperial College; G. Santini, Università Cattolica del Sacro Cuore; M. Saqi, European Institute for Systems Biology and Medicine, University of Lyon, France; S. Scott, North West Severe Asthma Network; Countess of Chester NHS Trust; N. Sehgal, North West Severe Asthma Network; Pennine Acute Hospital NHS Trust; A. Selby, NIHR Southampton Respiratory Biomedical Research Unit, Clinical and Experimental Sciences and Human Development and Health, Southampton, UK; P. Söderman, Astrid Lindgren Children's Hospital, Karolinska University Hospital, Stockholm, Sweden; Department of Women's and Children's Health, Stockholm, Sweden; A. Sogbesan, Royal Brompton and Harefield NHS Foundation Trust; F. Spycher, University Hospital, Department of Pulmonary Medicine, Bern, Switzerland; S. Stephan, Centre for Respiratory Medicine and Allergy, Institute of Inflammation and Repair, University Hospital of South Manchester, NHS Foundation Trust, Manchester, UK; J. Stokholm, University of Copenhagen and Danish Pediatric Asthma Center, Gentofte Hospital, University of Copenhagen, Denmark; M. Sunther, Centre for Respiratory Medicine and Allergy, Institute of Inflammation and Repair, University Hospital of South Manchester, NHS Foundation Trust, Manchester, UK; M. Szentkereszty, Semmelweis University, Budapest, Hungary; L. Tamasi, Semmelweis University, Budapest, Hungary; K. Tariq, NIHR Southampton Respiratory Biomedical Research Unit and University of Southampton; S. Valente, Università Cattolica del Sacro Cuore; W.M. van Aalderen, Academic Medical Centre, University of Amsterdam; C.M. van Drunen, Academic Medical Centre, University of Amsterdam; J. Van Eyll, UCB, Slough, UK; A. Vyas, North West Severe Asthma Network; Lancashire Teaching Hospitals NHS Trust; W. Yu, Amgen Inc.; W. Zetterquist, Department of Woman and Child Health, Karolinska Institutet, Department of Woman and Child Health, Karolinska Institutet, Stockholm, Sweden; Z. Zolkipli, NIHR Southampton Respiratory Biomedical Research Unit, University Hospital Southampton NHS Foundation Trust, Southampton, UK; 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, Department 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 Bautmans, 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; Romanas Chaleckis, Centre of Allergy Research, Karolinska Institutet, Stockholm, Sweden; Arnaldo D'Amico, University of Rome "Tor Vergata", Rome Italy; Jorge De Alba, Almirall S.A., Barcelona, Spain; Inge De Lepeleire, MSD, Brussels, Belgium; Tamara Dekker, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands; Annemiek Dijkhuis, 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; Kerry Gove, NIHR Southampton Respiratory Biomedical Research Unit and Clinical and Experimental Sciences, Southampton, UK; Neil Gozzard, UCB, Slough, UK; John Haughney, International 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 Kupczyk, Centre for Allergy Research, Karolinska Institutet, Stockholm, Sweden; Bart Lambrecht, 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,

DOI: 10.1021/acs.jproteome.8b00018
J. Proteome Res. 2018, 17, 2072−2091

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 Petrén, 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; Malayka Rahman-Amin, previously at Asthma UK, London, UK; Navin Rao, Janssen R&D, USA; Lara Ravanetti, 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 organizations: 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, Krakow, 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, biomedical 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 (cochair), 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; Juliëtte Kamphuis, The Netherlands; Lehanne Sergison, UK; Luigi

DOI: 10.1021/acs.jproteome.8b00018
J. Proteome Res. 2018, 17, 2072−2091

Visintin, Italy; Pim de Boer (cochair), 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 pathobiology (chair); Martina Gahlemann, patient safety advice and clinical care (cochair); Ralf Sigmund, bioinformatician. 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.

■ 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

■ REFERENCES

(1) Nicholas, B.; Djukanovic, R. Induced sputum: a window to lung pathology. *Biochem. Soc. Trans.* **2009**, *37*, 868−72.
(2) Nicholas, B.; Skipp, P.; Mould, R.; Rennard, S.; Davies, D. E.; O'Connor, C. D.; et al. Shotgun proteomic analysis of human-induced sputum. *Proteomics* **2006**, *6*, 4390−401.
(3) Nicholas, B. L.; O'Connor, C. D.; Djukanovic, R. From proteomics to prescription-the search for COPD biomarkers. *Copd* **2009**, *6*, 298−303.
(4) Hastie, A. T.; Moore, W. C.; Li, H.; Rector, B. M.; Ortega, V. E.; Pascual, R. M.; et al. Biomarker surrogates do not accurately predict sputum eosinophil and neutrophil percentages in asthmatic subjects. *J. Allergy Clin. Immunol.* **2013**, *132*, 72−80.
(5) Westerhof, G. A.; Korevaar, D. A.; Amelink, M.; de Nijs, S. B.; de Groot, J. C.; Wang, J.; et al. Biomarkers to identify sputum eosinophilia in different adult asthma phenotypes. *Eur. Respir. J.* **2015**, *46*, 688−96.
(6) Tak, T.; Hilvering, B.; Tesselaar, K.; Koenderman, L. Similar activation state of neutrophils in sputum of asthma patients irrespective of sputum eosinophilia. *Clin. Exp. Immunol.* **2015**, *182*, 204−12.
(7) Bergquist, M.; Jonasson, S.; Hjoberg, J.; Hedenstierna, G.; Hanrieder, J. Comprehensive multiplexed protein quantitation delineates eosinophilic and neutrophilic experimental asthma. *BMC Pulm. Med.* **2014**, *14*, 110.
(8) Chua, J. C.; Douglass, J. A.; Gillman, A.; O'Hehir, R. E.; Meeusen, E. N. Galectin-10, a potential biomarker of eosinophilic airway inflammation. *PLoS One* **2012**, *7*, e42549.
(9) Emmanouil, P.; Loukides, S.; Kostikas, K.; Papatheodorou, G.; Papaporfyriou, A.; Hillas, G.; et al. Sputum and BAL Clara cell secretory protein and surfactant protein D levels in asthma. *Allergy* **2015**, *70*, 711−4.
(10) Yilmaz, I.; Bayraktar, N.; Ceyhan, K.; Secil, D.; Yuksel, S.; Misirligil, Z.; et al. Evaluation of vascular endothelial growth factor-A and Endostatin levels in induced sputum and relationship to bronchial hyperreactivity in patients with persistent allergic rhinitis monosensitized to house dust. *Rev. Port. Pneumol.* **2015**, *21*, 321.
(11) Desai, D.; Gupta, S.; Siddiqui, S.; Singapuri, A.; Monteiro, W.; Entwisle, J.; et al. Sputum mediator profiling and relationship to airway wall geometry imaging in severe asthma. *Respir Res.* **2013**, *14*, 17.
(12) Mastalerz, L.; Celejewska-Wojcik, N.; Wojcik, K.; Gielicz, A.; Cmiel, A.; Ignacak, M.; et al. Induced sputum supernatant bioactive lipid mediators can identify subtypes of asthma. *Clin. Exp. Allergy* **2015**, *45*, 1779−89.
(13) Yan, X.; Chu, J. H.; Gomez, J.; Koenigs, M.; Holm, C.; He, X.; et al. Noninvasive analysis of the sputum transcriptome discriminates clinical phenotypes of asthma. *Am. J. Respir. Crit. Care Med.* **2015**, *191*, 1116−25.
(14) Gao, J.; Ohlmeier, S.; Nieminen, P.; Toljamo, T.; Tiitinen, S.; Kanerva, T.; et al. Elevated sputum BPIFB1 levels in smokers with chronic obstructive pulmonary disease: a longitudinal study. *Am. J. Physiol Lung Cell Mol. Physiol* **2015**, *309*, L17−26.
(15) Tangedal, S.; Aanerud, M.; Persson, L. J.; Brokstad, K. A.; Bakke, P. S.; Eagan, T. M. Comparison of inflammatory markers in induced and spontaneous sputum in a cohort of COPD patients. *Respir. Res.* **2014**, *15*, 138.
(16) Grabcanovic-Musija, F.; Obermayer, A.; Stoiber, W.; Krautgartner, W. D.; Steinbacher, P.; Winterberg, N.; et al. Neutrophil extracellular trap (NET) formation characterises stable and exacerbated COPD and correlates with airflow limitation. *Respir. Res.* **2015**, *16*, 59.
(17) Zuiker, R. G.; Kamerling, I. M.; Morelli, N.; Calderon, C.; Boot, J. D.; de Kam, M.; et al. Reproducibility of biomarkers in induced sputum and in serum from chronic smokers. *Pulm. Pharmacol. Ther.* **2015**, *33*, 81−6.
(18) Casado, B.; Iadarola, P.; Pannell, L. K.; Luisetti, M.; Corsico, A.; Ansaldo, E.; et al. Protein expression in sputum of smokers and chronic obstructive pulmonary disease patients: a pilot study by CapLC-ESI-Q-TOF. *J. Proteome Res.* **2007**, *6*, 4615−23.
(19) Nicholas, B. L.; Skipp, P.; Barton, S.; Singh, D.; Bagmane, D.; Mould, R.; et al. Identification of lipocalin and apolipoprotein A1 as biomarkers of chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* **2010**, *181*, 1049−60.
(20) Baraniuk, J. N.; Casado, B.; Pannell, L. K.; McGarvey, P. B.; Boschetto, P.; Luisetti, M.; et al. Protein networks in induced sputum from smokers and COPD patients. *Int. J. Chronic Obstruct. Pulm. Dis.* **2015**, *10*, 1957−75.
(21) Titz, B.; Sewer, A.; Schneider, T.; Elamin, A.; Martin, F.; Dijon, S.; et al. Alterations in the sputum proteome and transcriptome in smokers and early-stage COPD subjects. *J. Proteomics* **2015**, *128*, 306−20.
(22) Suojalehto, H.; Kinaret, P.; Kilpelainen, M.; Toskala, E.; Ahonen, N.; Wolff, H.; et al. Level of Fatty Acid Binding Protein 5 (FABP5) Is Increased in Sputum of Allergic Asthmatics and Links to Airway Remodeling and Inflammation. *PLoS One* **2015**, *10*, e0127003.
(23) Terracciano, R.; Preiano, M.; Palladino, G. P.; Carpagnano, G. E.; Barbaro, M. P.; Pelaia, G.; et al. Peptidome profiling of induced sputum by mesoporous silica beads and MALDI-TOF MS for non-invasive biomarker discovery of chronic inflammatory lung diseases. *Proteomics* **2011**, *11*, 3402−14.
(24) Gharib, S. A.; Nguyen, E. V.; Lai, Y.; Plampin, J. D.; Goodlett, D. R.; Hallstrand, T. S. Induced sputum proteome in healthy subjects and asthmatic patients. *J. Allergy Clin. Immunol.* **2011**, *128*, 1176−84.
(25) Diamandis, E. P. The failure of protein cancer biomarkers to reach the clinic: why, and what can be done to address the problem? *BMC Med.* **2012**, *10*, 87.
(26) Fu, X.; Gharib, S. A.; Green, P. S.; Aitken, M. L.; Frazer, D. A.; Park, D. R.; et al. Spectral index for assessment of differential protein expression in shotgun proteomics. *J. Proteome Res.* **2008**, *7*, 845−54.
(27) Drucker, E.; Krapfenbauer, K. Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine. *EPMA J.* **2013**, *4*, 7.
(28) Frantzi, M.; Bhat, A.; Latosinska, A. Clinical proteomic biomarkers: relevant issues on study design & technical considerations in biomarker development. *Clin Transl Med.* **2014**, *3*, 7.
(29) Kern, S. E. Why your new cancer biomarker may never work: recurrent patterns and remarkable diversity in biomarker failures. *Cancer Res.* **2012**, *72*, 6097−101.
(30) Mischak, H.; Allmaier, G.; Apweiler, R.; Attwood, T.; Baumann, M.; Benigni, A.; et al. Recommendations for biomarker identification and qualification in clinical proteomics. *Sci. Transl. Med.* **2010**, *2*, 46ps2.
(31) Skates, S. J.; Gillette, M. A.; LaBaer, J.; Carr, S. A.; Anderson, L.; Liebler, D. C.; et al. Statistical Design for Biospecimen Cohort Size in Proteomics-based Biomarker Discovery and Verification Studies. *J. Proteome Res.* **2013**, *12*, 5383−94.
(32) Silva, J. C.; Denny, R.; Dorschel, C. A.; Gorenstein, M.; Kass, I. J.; Li, G. Z.; et al. Quantitative proteomic analysis by accurate mass retention time pairs. *Anal. Chem.* **2005**, *77*, 2187−200.

DOI: 10.1021/acs.jproteome.8b00018
J. Proteome Res. 2018, 17, 2072−2091

(33) Silva, J. C.; Gorenstein, M. V.; Li, G. Z.; Vissers, J. P.; Geromanos, S. J. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell. Proteomics* **2006**, *5*, 144−56.
(34) Neilson, K. A.; Ali, N. A.; Muralidharan, S.; Mirzaei, M.; Mariani, M.; Assadourian, G.; et al. Less label, more free: approaches in label-free quantitative mass spectrometry. *Proteomics* **2011**, *11*, 535−53.
(35) Ow, S. Y.; Salim, M.; Noirel, J.; Evans, C.; Rehman, I.; Wright, P. C. iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly". *J. Proteome Res.* **2009**, *8*, 5347−55.
(36) Schulze, W. X.; Usadel, B. Quantitation in mass-spectrometry-based proteomics. *Annu. Rev. Plant Biol.* **2010**, *61*, 491−516.
(37) Foster, M. W.; Thompson, J. W.; Que, L. G.; Yang, I. V.; Schwartz, D. A.; Moseley, M. A.; et al. Proteomic analysis of human bronchoalveolar lavage fluid after subsgemental exposure. *J. Proteome Res.* **2013**, *12*, 2194−205.
(38) Shaw, D. E.; Sousa, A. R.; Fowler, S. J.; Fleming, L. J.; Roberts, G.; Corfield, J.; et al. Clinical and inflammatory characteristics of the European U-BIOPRED adult severe asthma cohort. *Eur. Respir. J.* **2015**, *46*, 1308−21.
(39) Djukanovic, R.; Sterk, P. J.; Fahy, J. V.; Hargreave, F. E. Standardised methodology of sputum induction and processing. *Eur. Respir. J.* **2002**, *37*, 1s−2s.
(40) Muntel, J.; Fromion, V.; Goelzer, A.; Maa*β*, S.; Mäder, U.; Büttner, K.; Hecker, M.; Becher, D. Comprehensive absolute quantification of the cytosolic proteome of Bacillus subtilis by data independent, parallel fragmentation in liquid chromatography/mass spectrometry (LC/MSE). *Mol. Cell. Proteomics* **2014**, *13* (4), 1008−1019.
(41) Silva, J. C.; Denny, R.; Dorschel, C.; Gorenstein, M. V.; Li, G. Z.; Richardson, K.; et al. Simultaneous qualitative and quantitative analysis of the Escherichia coli proteome: a sweet tale. *Mol. Cell. Proteomics* **2006**, *5*, 589−607.
(42) Li, G. Z.; Vissers, J. P.; Silva, J. C.; Golick, D.; Gorenstein, M. V.; Geromanos, S. J. Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures. *Proteomics* **2009**, *9*, 1696−719.
(43) Johnson, W. E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **2012**, *8* (1), 118−27.
(44) Polpitiya, A. D.; Qian, W. J.; Jaitly, N.; Petyuk, V. A.; Adkins, J. N.; Camp, D. G., 2nd; et al. DAnTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics* **2008**, *24*, 1556−8.
(45) Uhlen, M.; Fagerberg, L.; Hallstrom, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; et al. Proteomics. Tissue-based map of the human proteome. *Science* **2015**, *347*, 1260419.
(46) Farrah, T.; Deutsch, E. W.; Omenn, G. S.; Campbell, D. S.; Sun, Z.; Bletz, J. A.; et al. A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol. Cell. Proteomics* **2011**, *10*, M110.006353.
(47) Omenn, G. S.; States, D. J.; Adamski, M.; Blackwell, T. W.; Menon, R.; Hermjakob, H.; et al. Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* **2005**, *5*, 3226−45.
(48) Schenk, S.; Schoenhals, G. J.; de Souza, G.; Mann, M. A high confidence, manually validated human blood plasma protein reference set. *BMC Med. Genomics* **2008**, *1*, 41.
(49) Chen, J.; Ryu, S.; Gharib, S. A.; Goodlett, D. R.; Schnapp, L. M. Exploration of the normal human bronchoalveolar lavage fluid proteome. *Proteomics: Clin. Appl.* **2008**, *2*, 585−95.
(50) Nguyen, E. V.; Gharib, S. A.; Palazzo, S. J.; Chow, Y. H.; Goodlett, D. R.; Schnapp, L. M. Proteomic profiling of bronchoalveolar lavage fluid in critically ill patients with ventilator-associated pneumonia. *PLoS One* **2013**, *8*, e58782.
(51) Nguyen, E. V.; Gharib, S. A.; Schnapp, L. M.; Goodlett, D. R. Shotgun MS proteomic analysis of bronchoalveolar lavage fluid in normal subjects. *Proteomics: Clin. Appl.* **2014**, *8*, 737−47.
(52) Wu, J.; Kobayashi, M.; Sousa, E. A.; Liu, W.; Cai, J.; Goldman, S. J.; et al. Differential proteomic analysis of bronchoalveolar lavage fluid in asthmatics following segmental antigen challenge. *Mol. Cell. Proteomics* **2005**, *4*, 1251−64.
(53) Fumagalli, M.; Ferrari, F.; Luisetti, M.; Stolk, J.; Hiemstra, P. S.; Capuano, D.; et al. Profiling the proteome of exhaled breath condensate in healthy smokers and COPD patients by LC-MS/MS. *Int. J. Mol. Sci.* **2012**, *13*, 13894−910.
(54) Muccilli, V.; Saletti, R.; Cunsolo, V.; Ho, J.; Gili, E.; Conte, E.; et al. Protein profile of exhaled breath condensate determined by high resolution mass spectrometry. *J. Pharm. Biomed. Anal.* **2015**, *105*, 134−49.
(55) Joo, N. S.; Evans, I. A.; Cho, H. J.; Park, I. H.; Engelhardt, J. F.; Wine, J. J. Proteomic analysis of pure human airway gland mucus reveals a large component of protective proteins. *PLoS One* **2015**, *10*, e0116756.
(56) Fang, X.; Yang, L.; Wang, W.; Song, T.; Lee, C. S.; DeVoe, D. L.; et al. Comparison of electrokinetics-based multidimensional separations coupled with electrospray ionization-tandem mass spectrometry for characterization of human salivary proteins. *Anal. Chem.* **2007**, *79*, 5785−92.
(57) Guo, T.; Rudnick, P. A.; Wang, W.; Lee, C. S.; Devoe, D. L.; Balgley, B. M. Characterization of the human salivary proteome by capillary isoelectric focusing/nanoreversed-phase liquid chromatography coupled with ESI-tandem MS. *J. Proteome Res.* **2006**, *5*, 1469−78.
(58) Hu, S.; Xie, Y.; Ramachandran, P.; Ogorzalek Loo, R. R.; Li, Y.; Loo, J. A.; et al. Large-scale identification of proteins in human salivary proteome by liquid chromatography/mass spectrometry and two-dimensional gel electrophoresis-mass spectrometry. *Proteomics* **2005**, *5*, 1714−28.
(59) Ramachandran, P.; Boontheung, P.; Xie, Y.; Sondej, M.; Wong, D. T.; Loo, J. A. Identification of N-linked glycoproteins in human saliva by glycoprotein capture and mass spectrometry. *J. Proteome Res.* **2006**, *5*, 1493−503.
(60) Xie, H.; Rhodus, N. L.; Griffin, R. J.; Carlis, J. V.; Griffin, T. J. A catalogue of human saliva proteins identified by free flow electrophoresis-based peptide separation and tandem mass spectrometry. *Mol. Cell. Proteomics* **2005**, *4*, 1826−30.
(61) Yan, W.; Apweiler, R.; Balgley, B. M.; Boontheung, P.; Bundy, J. L.; Cargile, B. J.; et al. Systematic comparison of the human saliva and plasma proteomes. *Proteomics: Clin. Appl.* **2009**, *3*, 116−34.
(62) Cervero, P.; Himmel, M.; Kruger, M.; Linder, S. Proteomic analysis of podosome fractions from macrophages reveals similarities to spreading initiation centres. *Eur. J. Cell Biol.* **2012**, *91*, 908−22.
(63) Eligini, S.; Brioschi, M.; Fiorelli, S.; Tremoli, E.; Colli, S.; Banfi, C. Data for proteomic analysis of Human monocyte-derived macrophages. *Data Brief* **2015**, *4*, 177−9.
(64) Acharya, K. R.; Ackerman, S. J. Eosinophil granule proteins: form and function. *J. Biol. Chem.* **2014**, *289*, 17406−15.
(65) Kahn, J. E.; Dutoit-Lefevre, V.; Duban-Deweer, S.; Chafey, P.; Pottiez, G.; Lefranc, D.; et al. Comparative proteomic analysis of blood eosinophils reveals redox signaling modifications in patients with FIP1L1-PDGFRA-associated chronic eosinophilic leukemia. *J. Proteome Res.* **2011**, *10*, 1468−80.
(66) Straub, C.; Burnham, J. P.; White, A. C., Jr.; Pazdrak, K.; Sanchez, C.; Watanabe, L. C.; et al. Altered eosinophil proteome in a patient with hypereosinophilia from acute fascioliasis. *Clin. Vaccine Immunol.* **2011**, *18*, 1999−2002.
(67) Straub, C.; Pazdrak, K.; Young, T. W.; Stafford, S. J.; Wu, Z.; Wiktorowicz, J. E.; et al. Toward the Proteome of the Human Peripheral Blood Eosinophil. *Proteomics: Clin. Appl.* **2009**, *3*, 1151−73.
(68) Yoon, S. W.; Kim, T. Y.; Sung, M. H.; Kim, C. J.; Poo, H. Comparative proteomic analysis of peripheral blood eosinophils from healthy donors and atopic dermatitis patients with eosinophilia. *Proteomics* **2005**, *5*, 1987−95.
(69) Tomazella, G. G.; da Silva, I.; Laure, H. J.; Rosa, J. C.; Chammas, R.; Wiker, H. G.; et al. Proteomic analysis of total cellular proteins of human neutrophils. *Proteome Sci.* **2009**, *7*, 32.
(70) Tomazella, G. G.; daSilva, I.; Thome, C. H.; Greene, L. J.; Koehler, C. J.; Thiede, B.; et al. Analysis of detergent-insoluble and whole cell

DOI: 10.1021/acs.jproteome.8b00018
J. Proteome Res. 2018, 17, 2072−2091

lysate fractions of resting neutrophils using high-resolution mass spectrometry. *J. Proteome Res.* **2010**, *9*, 2030−6.
(71) Trusch, M.; Tillack, K.; Kwiatkowski, M.; Bertsch, A.; Ahrends, R.; Kohlbacher, O.; et al. Displacement chromatography as first separating step in online two-dimensional liquid chromatography coupled to mass spectrometry analysis of a complex protein sample−the proteome of neutrophils. *J. Chromatogr A* **2012**, *1232*, 288−94.
(72) Zhu, J.; Zhang, H.; Guo, T.; Li, W.; Li, H.; Zhu, Y.; et al. Quantitative proteomics reveals differential biological processes in healthy neonatal cord neutrophils and adult neutrophils. *Proteomics* **2014**, *14*, 1688−97.
(73) Brinkmann, V.; Reichard, U.; Goosmann, C.; Fauler, B.; Uhlemann, Y.; Weiss, D. S.; et al. Neutrophil extracellular traps kill bacteria. *Science* **2004**, *303*, 1532−5.
(74) Guimaraes-Costa, A. B.; Nascimento, M. T.; Wardini, A. B.; Pinto-da-Silva, L. H.; Saraiva, E. M. ETosis: A Microbicidal Mechanism beyond Cell Death. *J. Parasitol. Res.* **2012**, *2012*, 929743.
(75) Dalli, J.; Montero-Melendez, T.; Norling, L. V.; Yin, X.; Hinds, C.; Haskard, D.; et al. Heterogeneity in neutrophil microparticles reveals distinct proteome and functional properties. *Mol. Cell. Proteomics* **2013**, *12*, 2205−19.
(76) Lominadze, G.; Powell, D. W.; Luerman, G. C.; Link, A. J.; Ward, R. A.; McLeish, K. R. Proteomic analysis of human neutrophil granules. *Mol. Cell. Proteomics* **2005**, *4*, 1503−21.
(77) Rorvig, S.; Ostergaard, O.; Heegaard, N. H.; Borregaard, N. Proteome profiling of human neutrophil granule subsets, secretory vesicles, and cell membrane: correlation with transcriptome profiling of neutrophil precursors. *J. Leukocyte Biol.* **2013**, *94*, 711−21.
(78) Huang da, W.; Sherman, B. T.; Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **2009**, *37*, 1−13.
(79) Huang da, W.; Sherman, B. T.; Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **2009**, *4*, 44−57.
(80) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2012.
(81) *Venny: An interactive tool for comparing lists with Venn's diagrams*; 2007−2015; http://bioinfogp.cnb.csic.es/tools/venny/index.html.
(82) Bederson, B. B.; Shneiderman, B.; Wattenberg, M. Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies, In *The Craft of Information Visualization*; Bederson, B. B., Shneiderman, B., Eds.; Morgan Kaufmann: San Francisco, 2003; pp 257−78.
(83) Pathan, M.; Keerthikumar, S.; Ang, C. S.; Gangoda, L.; Quek, C. Y.; Williamson, N. A.; et al. FunRich: An open access standalone functional enrichment and interaction network analysis tool. *Proteomics* **2015**, *15*, 2597−601.
(84) Chen, Y.; Zhang, Y.; Yin, Y.; Gao, G.; Li, S.; Jiang, Y.; et al. SPD−a web-based secreted protein database. *Nucleic Acids Res.* **2005**, *33*, D169−73.
(85) Szklarczyk, D.; Franceschini, A.; Kuhn, M.; Simonovic, M.; Roth, A.; Minguez, P.; et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **2011**, *39*, D561−8.
(86) Bateman, A.; Martin, M. J.; O'Donovan, C.; Magrane, M.; Apweiler, R.; Alpi, E.; et al. UniProt: a hub for protein information. *Nucleic Acids Res.* **2015**, *43*, D204−12.
(87) Eden, E.; Navon, R.; Steinfeld, I.; Lipson, D.; Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinf.* **2009**, *10*, 48.
(88) Supek, F.; Bosnjak, M.; Skunca, N.; Smuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **2011**, *6*, e21800.
(89) Hinks, T. S. C.; Zhou, X.; Staples, K. J.; Dimitrov, B. D.; Manta, A.; Petrossian, T.; Lum, P. Y.; Smith, C. G.; Ward, J. A.; Howarth, P. H.; et al. Innate and adaptive T cells in asthmatic patients: relationship to severity and disease mechanisms. *J. Allergy Clin. Immunol.* **2015**, *136*, 323−333.
(90) Li, L.; Cheng, W. Y.; Glicksberg, B. S.; Gottesman, O.; Tamler, R.; Chen, R.; Bottinger, E. P.; Dudley, J. T. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci. Transl. Med.* **2015**, *7*, 311ra174.
(91) Nielson, J. L.; Paquette, J.; Liu, A. W.; Guandique, C. F.; Tovar, C. A.; Inoue, T.; Irvine, K. A.; Gensel, J. C.; Kloke, J.; Petrossian, T. C.; et al. Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nat. Commun.* **2015**, *6*, 8581.
(92) Lum, P. Y.; Singh, G.; Lehman, A.; Ishkanov, T.; Vejdemo-Johansson, M.; Alagappan, M.; Carlsson, J.; Carlsson, G. Extracting insights from the shape of complex data using topology. *Sci. Rep.* **2013**, *3*, 1236.
(93) Carlsson, G. Topology and data. *Bull. Am. Math Soc.* **2009**, *46*, 255−308.
(94) Hoffman, G. E.; Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinf.* **2016**, *17* (1), 483.
(95) Ryckman, C.; Vandal, K.; Rouleau, P.; Talbot, M.; Tessier, P. A. Proinflammatory activities of S100: proteins S100A8, S100A9, and S100A8/A9 induce neutrophil chemotaxis and adhesion. *J. Immunol.* **2003**, *170*, 3233−42.
(96) Stockley, R. A. The multiple facets of alpha-1-antitrypsin. *Ann. Transl. Med.* **2015**, *3*, 130.
(97) Gaber, F.; Acevedo, F.; Delin, I.; Sundblad, B. M.; Palmberg, L.; Larsson, K.; et al. Saliva is one likely source of leukotriene B4 in exhaled breath condensate. *Eur. Respir. J.* **2006**, *28*, 1229−35.
(98) Lee, T. H.; Jang, A. S.; Park, J. S.; Kim, T. H.; Choi, Y. S.; Shin, H. R.; et al. Elevation of S100 calcium binding protein A9 in sputum of neutrophilic inflammation in severe uncontrolled asthma. *Ann. Allergy, Asthma, Immunol.* **2013**, *111*, 268.
(99) Gray, R. D.; MacGregor, G.; Noble, D.; Imrie, M.; Dewar, M.; Boyd, A. C.; et al. Sputum proteomics in inflammatory and suppurative respiratory diseases. *Am. J. Respir. Crit. Care Med.* **2008**, *178*, 444−52.
(100) Louten, J.; Mattson, J. D.; Malinao, M. C.; Li, Y.; Emson, C.; Vega, F.; et al. Biomarkers of disease and treatment in murine and cynomolgus models of chronic asthma. *Biomarker Insights* **2012**, *7*, 87−104.
(101) Zhao, J.; Zhu, H.; Wong, C. H.; Leung, K. Y.; Wong, W. S. Increased lungkine and Chitinase levels in allergic airway inflammation: a proteomics approach. *Proteomics* **2005**, *5*, 2799−807.
(102) Bloemen, K.; Hooyberghs, J.; Desager, K.; Witters, E.; Schoeters, G. Non-invasive biomarker sampling and analysis of the exhaled breath proteome. *Proteomics: Clin. Appl.* **2009**, *3*, 498−504.
(103) Goswami, S.; Angkasekwinai, P.; Shan, M.; Greenlee, K. J.; Barranco, W. T.; Polikepahad, S.; et al. Divergent functions for airway epithelial matrix metalloproteinase 7 and retinoic acid in experimental asthma. *Nat. Immunol.* **2009**, *10*, 496−503.
(104) North, M. L.; Khanna, N.; Marsden, P. A.; Grasemann, H.; Scott, J. A. Functionally important role for arginase 1 in the airway hyperresponsiveness of asthma. *Am. J. Physiol Lung Cell Mol. Physiol* **2009**, *296*, L911−20.
(105) Calvo, F. Q.; Fillet, M.; de Seny, D.; Meuwis, M. A.; Maree, R.; Crahay, C.; et al. Biomarker discovery in asthma-related inflammation and remodeling. *Proteomics* **2009**, *9*, 2163−70.
(106) Jeong, H.; Rhim, T.; Ahn, M. H.; Yoon, P. O.; Kim, S. H.; Chung, I. Y.; et al. Proteomic analysis of differently expressed proteins in a mouse model for allergic asthma. *J. Korean Med. Sci.* **2005**, *20*, 579−85.
(107) Chu, H. W.; Thaikoottathil, J.; Rino, J. G.; Zhang, G.; Wu, Q.; Moss, T.; et al. Function and regulation of SPLUNC1 protein in Mycoplasma infection and allergic inflammation. *J. Immunol.* **2007**, *179*, 3995−4002.
(108) Kawakami, M.; Narumoto, O.; Matsuo, Y.; Horiguchi, K.; Horiguchi, S.; Yamashita, N.; et al. The role of CCR7 in allergic airway inflammation induced by house dust mite exposure. *Cell. Immunol.* **2012**, *275*, 24−32.
(109) Brebner, J. A.; Stockley, R. A. Recent advances in alpha-1-antitrypsin deficiency-related lung disease. *Expert Rev. Respir. Med.* **2013**, *7*, 213−29.

DOI: 10.1021/acs.jproteome.8b00018
J. Proteome Res. 2018, 17, 2072−2091

(110) Ishikawa, N.; Hattori, N.; Kohno, N.; Kobayashi, A.; Hayamizu, T.; Johnson, M. Airway inflammation in Japanese COPD patients compared with smoking and nonsmoking controls. *Int. J. Chronic Obstruct. Pulm. Dis.* **2015**, *10*, 185−92.
(111) Ohlmeier, S.; Mazur, W.; Linja-Aho, A.; Louhelainen, N.; Ronty, M.; Toljamo, T.; et al. Sputum proteomics identifies elevated PIGR levels in smokers and mild-to-moderate COPD. *J. Proteome Res.* **2012**, *11*, 599−608.
(112) Merkel, D.; Rist, W.; Seither, P.; Weith, A.; Lenter, M. C. Proteomic study of human bronchoalveolar lavage fluids from smokers with chronic obstructive pulmonary disease by combining surface-enhanced laser desorption/ionization-mass spectrometry profiling with mass spectrometric protein identification. *Proteomics* **2005**, *5*, 2972−80.
(113) Shiratsuchi, N.; Asai, K.; Kanazawa, H.; Kyoh, S.; Tochino, Y.; Kodama, T.; et al. Measurement of soluble perforin, a marker of CD8+ T lymphocyte activation in epithelial lining fluid. *Respir Med.* **2011**, *105*, 1885−90.
(114) Gohy, S. T.; Detry, B. R.; Lecocq, M.; Bouzin, C.; Weynand, B. A.; Amatngalim, G. D.; et al. Polymeric immunoglobulin receptor down-regulation in chronic obstructive pulmonary disease. Persistence in the cultured epithelium and role of transforming growth factor-beta. *Am. J. Respir. Crit. Care Med.* **2014**, *190*, 509−21.
(115) Stewart, C. E.; Sayers, I. Urokinase receptor orchestrates the plasminogen system in airway epithelial cell function. *Lung* **2013**, *191*, 215−25.
(116) Calero, C.; Arellano, E.; Lopez-Villalobos, J. L.; Sanchez-Lopez, V.; Moreno-Mata, N.; Lopez-Campos, J. L. Differential expression of C-reactive protein and serum amyloid A in different cell types in the lung tissue of chronic obstructive pulmonary disease patients. *BMC Pulm. Med.* **2014**, *14*, 95.
(117) Ohlmeier, S.; Vuolanto, M.; Toljamo, T.; Vuopala, K.; Salmenkivi, K.; Myllarniemi, M.; et al. Proteomics of human lung tissue identifies surfactant protein A as a marker of chronic obstructive pulmonary disease. *J. Proteome Res.* **2008**, *7*, 5125−32.
(118) Hara, A.; Sakamoto, N.; Ishimatsu, Y.; Kakugawa, T.; Nakashima, S.; Hara, S.; et al. S100A9 in BALF is a candidate biomarker of idiopathic pulmonary fibrosis. *Respir Med.* **2012**, *106*, 571−80.
(119) Landi, C.; Bargagli, E.; Carleo, A.; Bianchi, L.; Gagliardi, A.; Prasse, A.; et al. A system biology study of BALF from patients affected by idiopathic pulmonary fibrosis (IPF) and healthy controls. *Proteomics: Clin. Appl.* **2014**, *8*, 932−50.
(120) Mukae, H.; Ishimoto, H.; Yanagi, S.; Ishii, H.; Nakayama, S.; Ashitani, J.; et al. Elevated BALF concentrations of alpha- and beta-defensins in patients with pulmonary alveolar proteinosis. *Respir Med.* **2007**, *101*, 715−21.
(121) Jaffar, J.; Unger, S.; Corte, T. J.; Keller, M.; Wolters, P. J.; Richeldi, L.; et al. Fibulin-1 predicts disease progression in patients with idiopathic pulmonary fibrosis. *Chest* **2014**, *146*, 1055−63.
(122) Bhargava, M.; Becker, T. L.; Viken, K. J.; Jagtap, P. D.; Dey, S.; Steinbach, M. S.; et al. Proteomic profiles in acute respiratory distress syndrome differentiates survivors from non-survivors. *PLoS One* **2014**, *9*, e109713.
(123) Sloane, A. J.; Lindner, R. A.; Prasad, S. S.; Sebastian, L. T.; Pedersen, S. K.; Robinson, M.; et al. Proteomic analysis of sputum from adults and children with cystic fibrosis and from control subjects. *Am. J. Respir. Crit. Care Med.* **2005**, *172*, 1416−26.
(124) Bullens, D. M.; Truyen, E.; Coteur, L.; Dilissen, E.; Hellings, P. W.; Dupont, L. J.; et al. IL-17 mRNA in sputum of asthmatic patients: linking T cell driven inflammation and granulocytic influx? *Respir. Res.* **2006**, *7*, 135.
(125) Fahy, J. V.; Kim, K. W.; Liu, J.; Boushey, H. A. Prominent neutrophilic inflammation in sputum from subjects with asthma exacerbation. *J. Allergy Clin. Immunol.* **1995**, *95*, 843−52.
(126) Cox, G. Glucocorticoid treatment inhibits apoptosis in human neutrophils. Separation of survival and activation outcomes. *J. Immunol.* **1995**, *154*, 4719−25.
(127) Dale, D. C.; Fauci, A. S.; Wolff, S. M. Alternate-day prednisone. Leukocyte kinetics and susceptibility to infections. *N. Engl. J. Med.* **1974**, *291*, 1154−8.
(128) Pedersen, B.; Dahl, R.; Karlstrom, R.; Peterson, C. G.; Venge, P. Eosinophil and neutrophil activity in asthma in a one-year trial with inhaled budesonide. The impact of smoking. *Am. J. Respir. Crit. Care Med.* **1996**, *153*, 1519−29.
(129) Durdiakova, J.; Fabryova, H.; Koborova, I.; Ostatnikova, D.; Celec, P. The effects of saliva collection, handling and storage on salivary testosterone measurement. *Steroids* **2013**, *78*, 1325−31.
(130) Kamodyova, N.; Banasova, L.; Jansakova, K.; Koborova, I.; Tothova, L.; Stanko, P.; et al. Blood Contamination in Saliva: Impact on the Measurement of Salivary Oxidative Stress Markers. *Dis. Markers* **2015**, *2015*, 479251.
(131) Lőrincz, Á.M.; Schütte, M.; Timár, C. I.; Veres, D. S.; Kittel, Á.; McLeish, K. R.; Merchant, M. L.; Ligeti, E. Functionally and morphologically distinct populations of extracellular vesicles produced by human neutrophilic granulocytes. *J. Leukocyte Biol.* **2015**, *98* (4), 583−589.
(132) Kobayashi, K.; Ogata, H.; Morikawa, M.; Iijima, S.; Harada, N.; Yoshida, T.; et al. Distribution and partial characterisation of IgG Fc binding protein in various mucin producing cells and body fluids. *Gut* **2002**, *51*, 169−76.
(133) Thornton, D. J.; Sheehan, J. K. From mucins to mucus: toward a more coherent understanding of this essential barrier. *Proc. Am. Thorac. Soc.* **2004**, *1*, 54−61.
(134) Andersson, C.; Zaman, M. M.; Jones, A. B.; Freedman, S. D. Alterations in immune response and PPAR/LXR regulation in cystic fibrosis macrophages. *J. Cystic Fibrosis* **2008**, *7*, 68−78.
(135) Hong, C.; Walczak, R.; Dhamko, H.; Bradley, M. N.; Marathe, C.; Boyadjian, R.; et al. Constitutive activation of LXR in macrophages regulates metabolic and inflammatory gene expression: identification of ARL7 as a direct target. *J. Lipid Res.* **2011**, *52*, 531−9.
(136) Renga, B.; Migliorati, M.; Mencarelli, A.; Fiorucci, S. Reciprocal regulation of the bile acid-activated receptor FXR and the interferon-γ-STAT-1 pathway in macrophages. *Biochim. Biophys. Acta, Mol. Basis Dis.* **2009**, *1792*, 564−73.
(137) Stojancevic, M.; Stankov, K.; Mikov, M. The impact of farnesoid X receptor activation on intestinal permeability in inflammatory bowel disease. *Can. J. Gastroenterol* **2012**, *26*, 631−7.
(138) Birrell, M. A.; De Alba, J.; Catley, M. C.; Hardaker, E.; Wong, S.; Collins, M.; et al. Liver X receptor agonists increase airway reactivity in a model of asthma via increasing airway smooth muscle growth. *J. Immunol.* **2008**, *181*, 4265−71.
(139) Castrillo, A.; Joseph, S. B.; Vaidya, S. A.; Haberland, M.; Fogelman, A. M.; Cheng, G.; et al. Crosstalk between LXR and Toll-like Receptor Signaling Mediates Bacterial and Viral Antagonism of Cholesterol Metabolism. *Mol. Cell* **2003**, *12*, 805−16.
(140) Tall, A. R.; Yvan-Charvet, L. Cholesterol, inflammation and innate immunity. *Nat. Rev. Immunol.* **2015**, *15*, 104−16.
(141) Bezemer, G. F.; Sagar, S.; van Bergenhenegouwen, J.; Georgiou, N. A.; Garssen, J.; Kraneveld, A. D.; et al. Dual role of Toll-like receptors in asthma and chronic obstructive pulmonary disease. *Pharmacol. Rev.* **2012**, *64*, 337−58.
(142) Phipps, S.; Lam, C. E.; Foster, P. S.; Matthaei, K. I. The contribution of toll-like receptors to the pathogenesis of asthma. *Immunol. Cell Biol.* **2007**, *85*, 463−70.
(143) Chiba, Y.; Misawa, M. MicroRNAs and their therapeutic potential for human diseases: MiR-133a and bronchial smooth muscle hyperresponsiveness in asthma. *J. Pharmacol. Sci.* **2010**, *114*, 264−8.
(144) Yoshii, A.; Iizuka, K.; Dobashi, K.; Horie, T.; Harada, T.; Nakazawa, T.; et al. Relaxation of contracted rabbit tracheal and human bronchial smooth muscle by Y-27632 through inhibition of Ca2+ sensitization. *Am. J. Respir. Cell Mol. Biol.* **1999**, *20*, 1190−200.
(145) Bastarache, J. A.; Fremont, R. D.; Kropski, J. A.; Bossert, F. R.; Ware, L. B. Procoagulant alveolar microparticles in the lungs of patients with acute respiratory distress syndrome. *AJP: Lung Cellular and Molecular Physiology* **2009**, *297* (6), L1035−L1041.
(146) Sadallah, S.; Eken, C.; Schifferli, J. A. Ectosomes as modulators of inflammation and immunity. *Clin. Exp. Immunol.* **2011**, *163* (1), 26−32.
(147) Pick, E.; Gorzalczany, Y.; Engel, S. Role of the rac1 p21-GDP-dissociation inhibitor for rho heterodimer in the activation of the

DOI: 10.1021/acs.jproteome.8b00018
J. Proteome Res. 2018, 17, 2072−2091

superoxide-forming NADPH oxidase of macrophages. *Eur. J. Biochem.* **1993**, *217*, 441−55.
(148) Robbins, P. D.; Morelli, A. E. Regulation of immune responses by extracellular vesicles. *Nat. Rev. Immunol.* **2014**, *14*, 195−208.
(149) Hess, C.; Sadallah, S.; Hefti, A.; Landmann, R.; Schifferli, J. A. Ectosomes released by human neutrophils are specialized functional units. *Mol. Immunol.* **1998**, *35163*, 354.
(150) Admyre, C.; Telemo, E.; Almqvist, N.; Lotvall, J.; Lahesmaa, R.; Scheynius, A.; et al. Exosomes - nanovesicles with possible roles in allergic inflammation. *Allergy* **2008**, *63*, 404−8.
(151) Spencer, L. A.; Bonjour, K.; Melo, R. C.; Weller, P. F. Eosinophil secretion of granule-derived cytokines. *Front. Immunol.* **2014**, *5*, 496.
(152) Torregrosa Paredes, P.; Esser, J.; Admyre, C.; Nord, M.; Rahman, Q. K.; Lukic, A.; et al. Bronchoalveolar lavage fluid exosomes contribute to cytokine and leukotriene production in allergic asthma. *Allergy* **2012**, *67*, 911−9.

DOI: 10.1021/acs.jproteome.8b00018
J. Proteome Res. 2018, 17, 2072−2091

Supporting Information

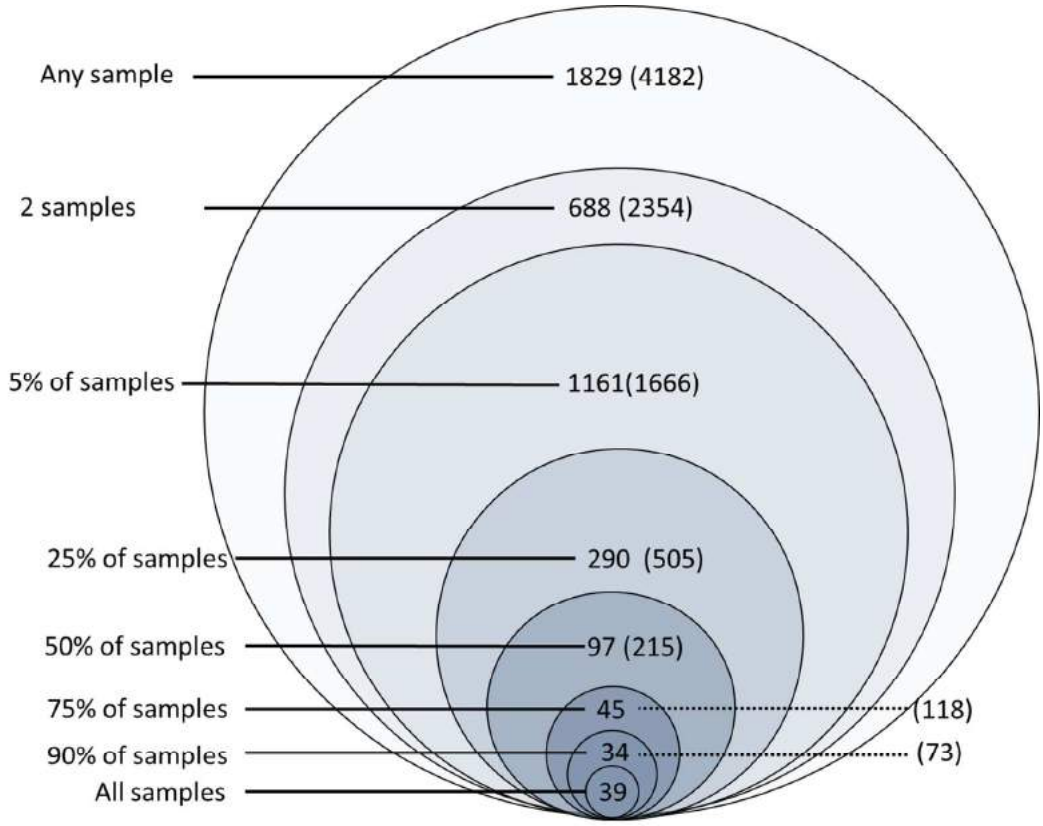
Large-scale label-free quantitative mapping of the sputum proteome

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[2], Rene Lutter[7, 8], Louise Fleming[9], Dominick E Shaw[10], Per S Bakke[11], 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[4], Charles Auffray[22], Ana R Sousa[23], Ian M Adcock[24], Kian Fan Chung[9], Peter J Sterk[7], Ratko Djukanović[2, #], Paul J Skipp[1, #] and the U-BIOPRED Study Group[25]

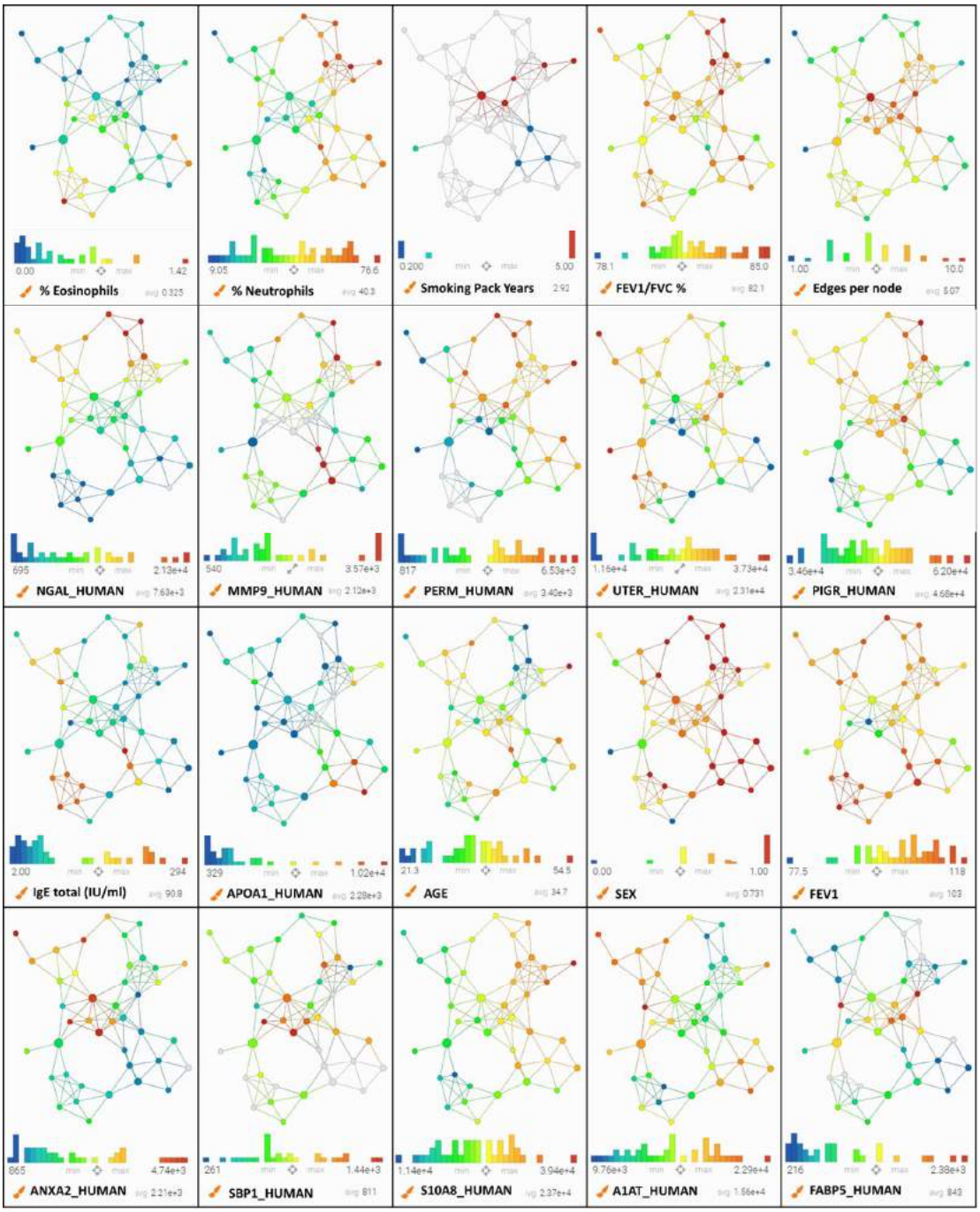
[1]Centre for Proteomic Research, University of Southampton, UK
[2]NIHR Southampton Respiratory Biomedical research unit, University Hospital Southampton, UK
[3]Acclarogen Ltd, Cambridge, UK
[4]Data Science Institute, Imperial College, London, UK
[5]Janssen Research & Development, Buckinghamshire, UK
[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.
[18]Faculty of Medicine, Catholic University of the Sacred Heart, Rome, Italy.
[19]Laboratory of Molecular Biology and Clinical Genetics, Medical College, Jagiellonian University, Krakow, Poland
[20]Dept. of Medicine, Dept of Public Health and Clinical Medicine Respiratory Medicine Unit, Umeå University, Umeå, Sweden.
[21]University Children's Hospital Zurich, Zurich, Switzerland.
[22]European Institute for Systems Biology and Medicine, CNRS-ENS-UCBL-INSERM, Université de Lyon, France
[23]Respiratory Therapeutic Unit, GSK, Stockley Park, UK.
[24]Cell and Molecular Biology Group, Airways Disease Section, National Heart and Lung Institute, Imperial College London, Dovehouse Street, London, UK
[25]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

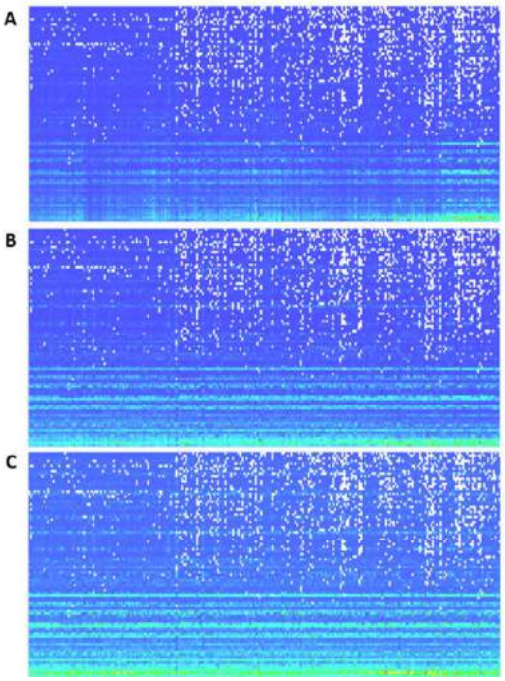
Table of Contents



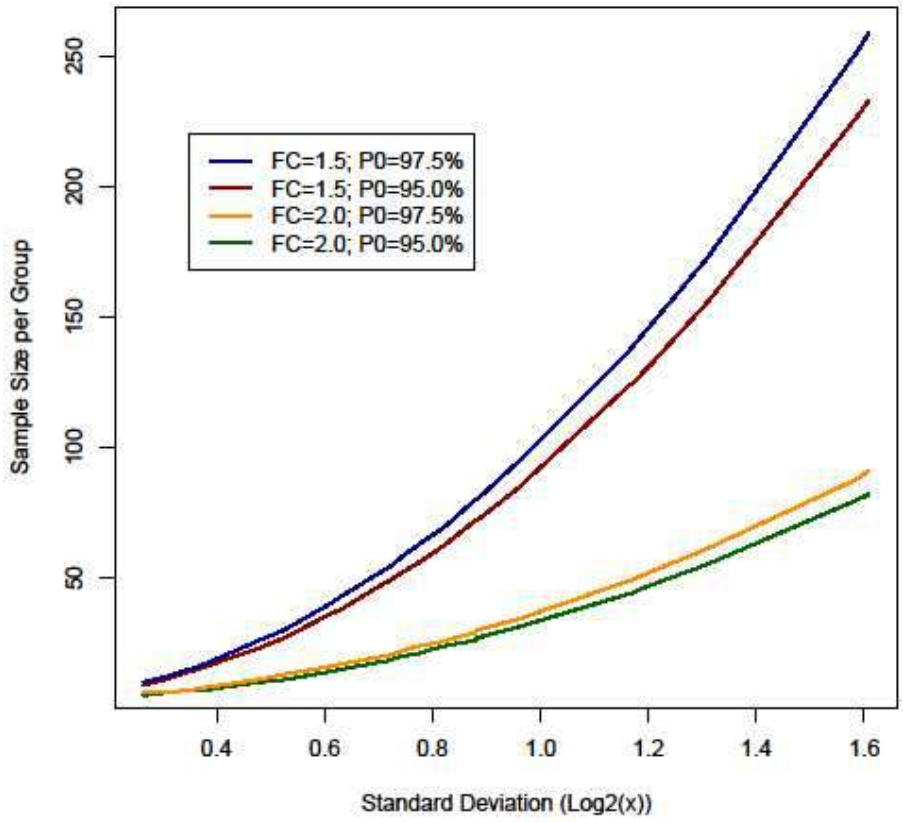
Supplementary Figure S1. Distribution of sputum proteins across samples. Each circle represents a specific subset of protein coverage across samples and the numbers within each circle represent the number of identifications in that coverage region and cumulative identifications in brackets. Each sample had a core of ~60 proteins that were repeatedly identified across samples (found in at least 90% of samples). Above this level, sparsity in the dataset increased dramatically with approximately 95% of the total protein pool identified in fewer than 25% of analysed samples



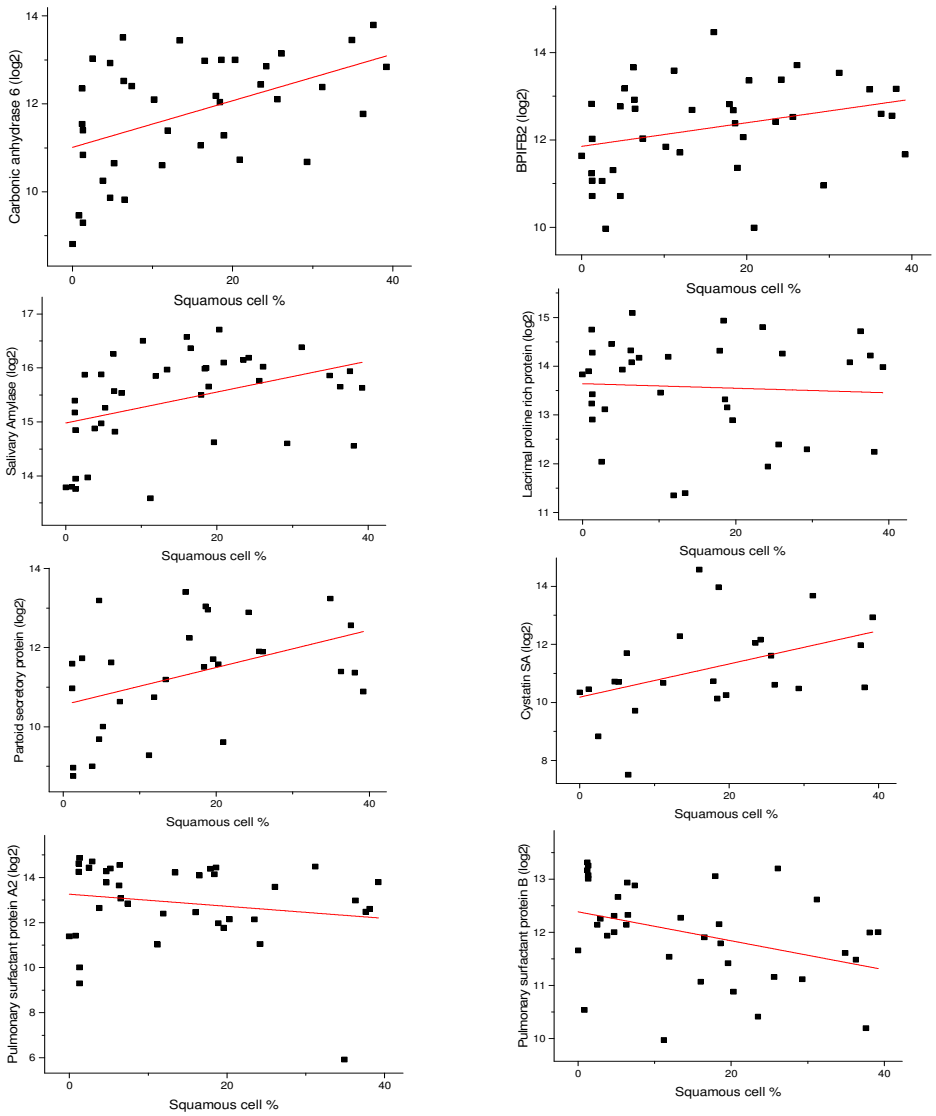
Supplementary Figure S2. Topological data analysis (TDA) network showing no clustering according to abundant proteins or clinical variables. TDA was used to assess distribution of variables across participants' samples. Participant samples could not be separated according to protein expression or clinical variables, these variables did not cluster within the population studied.



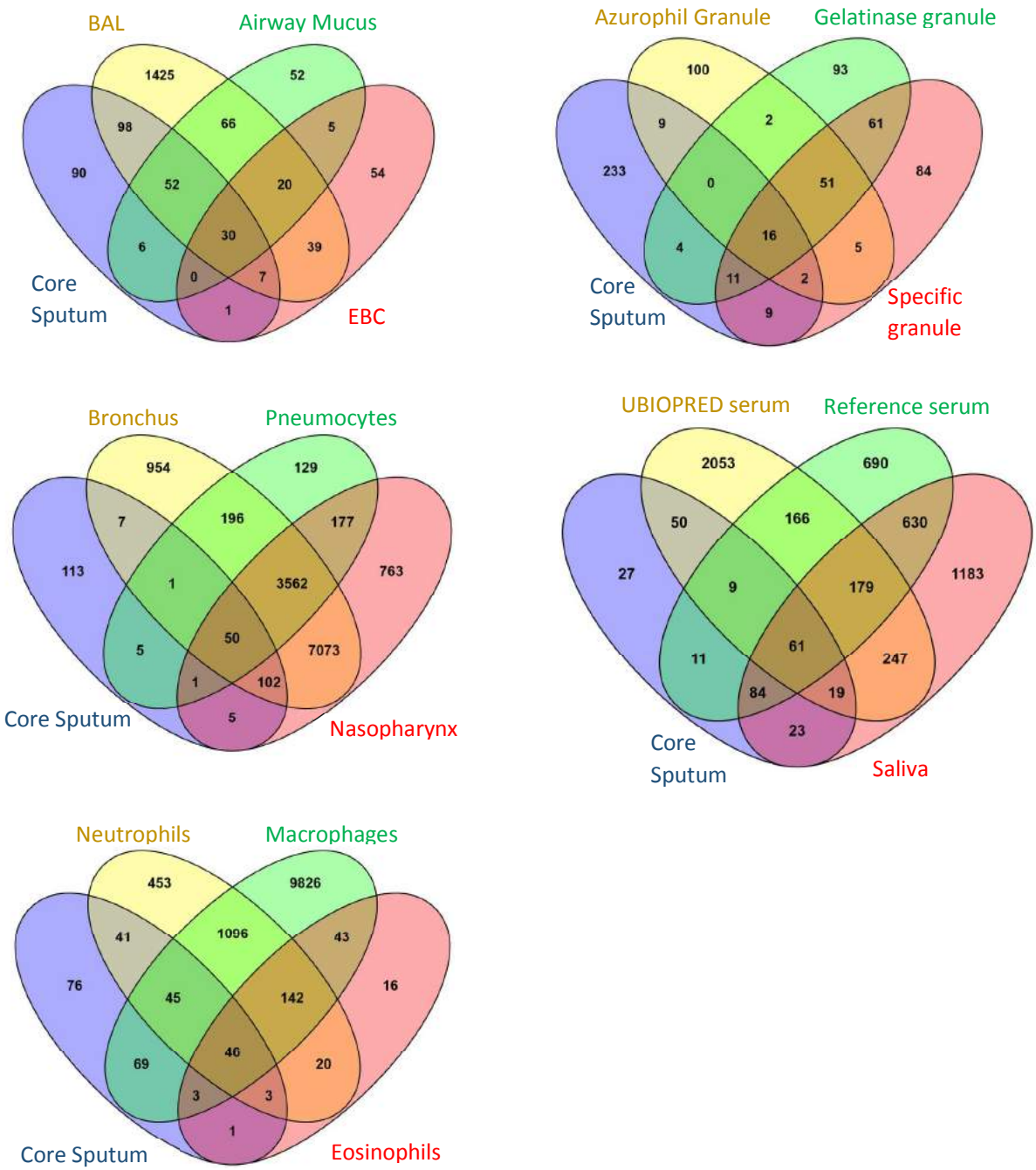
Supplementary Figure S3. Heatmaps of the U-BIOPRED sputum proteomics data. X axes are ranked samples in the order in which they were run, and the Y axes are proteins with ≤20% missing values. A) raw top3 intensity data. B) 'top90' normalised data. C) data after batch effect correction using Combat. Systematic variability can be seen in the raw data which is largely corrected for by 'top90' correction.



Supplementary Figure S4. The sample size required per group relative to variability of the samples given by the standard deviation of the measurements.



Supplementary figure S5. Correlations between proteins and squamous cell counts. Moderate correlations are seen between proteins reported to be salivary proteins and squamous cell counts, and there is a moderately negative correlation between lung secretory proteins and squamous cells counts.

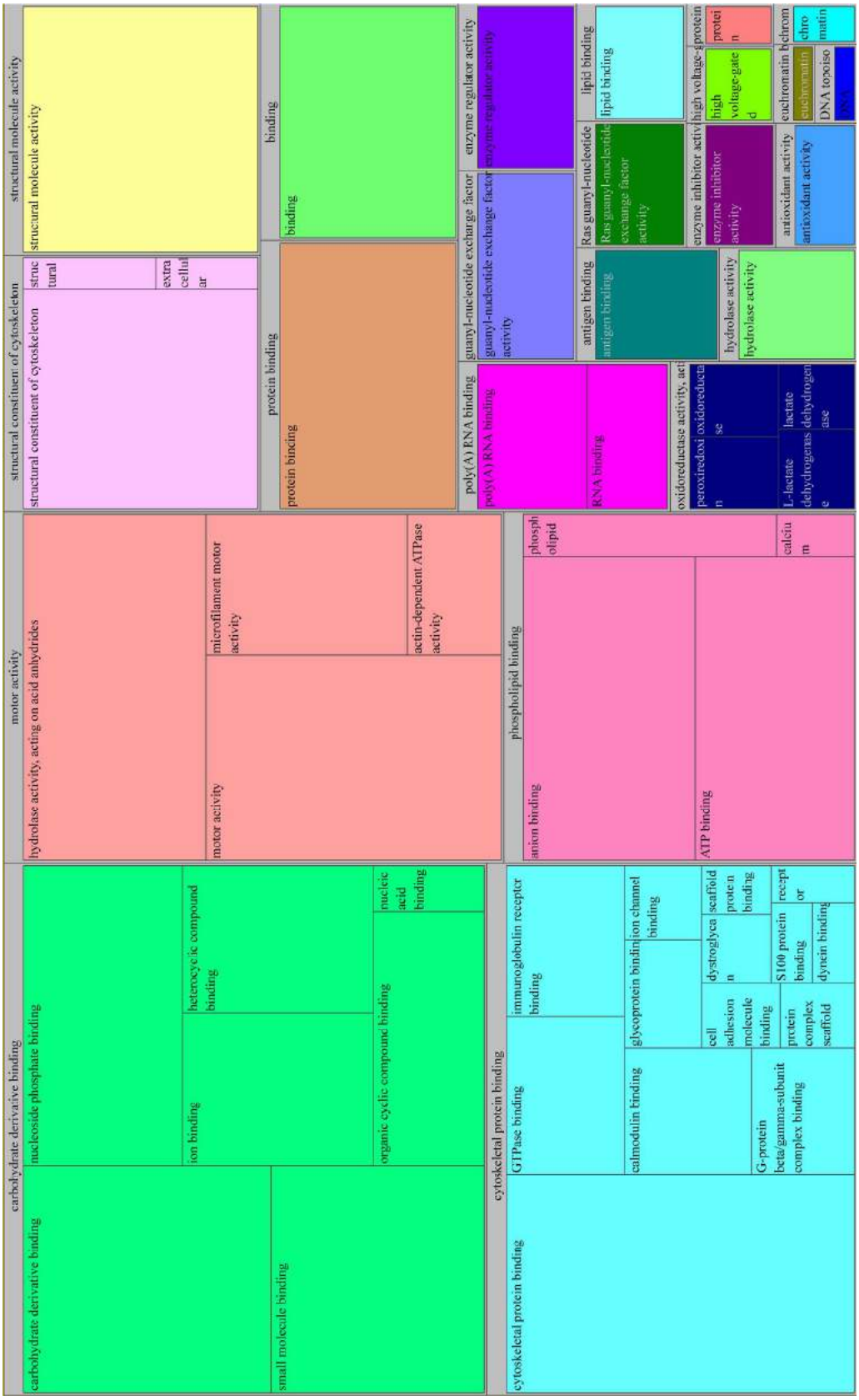


Supplementary figure S6. Overlap of proteins in the U-BIOPRED core sputum proteome in comparison to other studies.

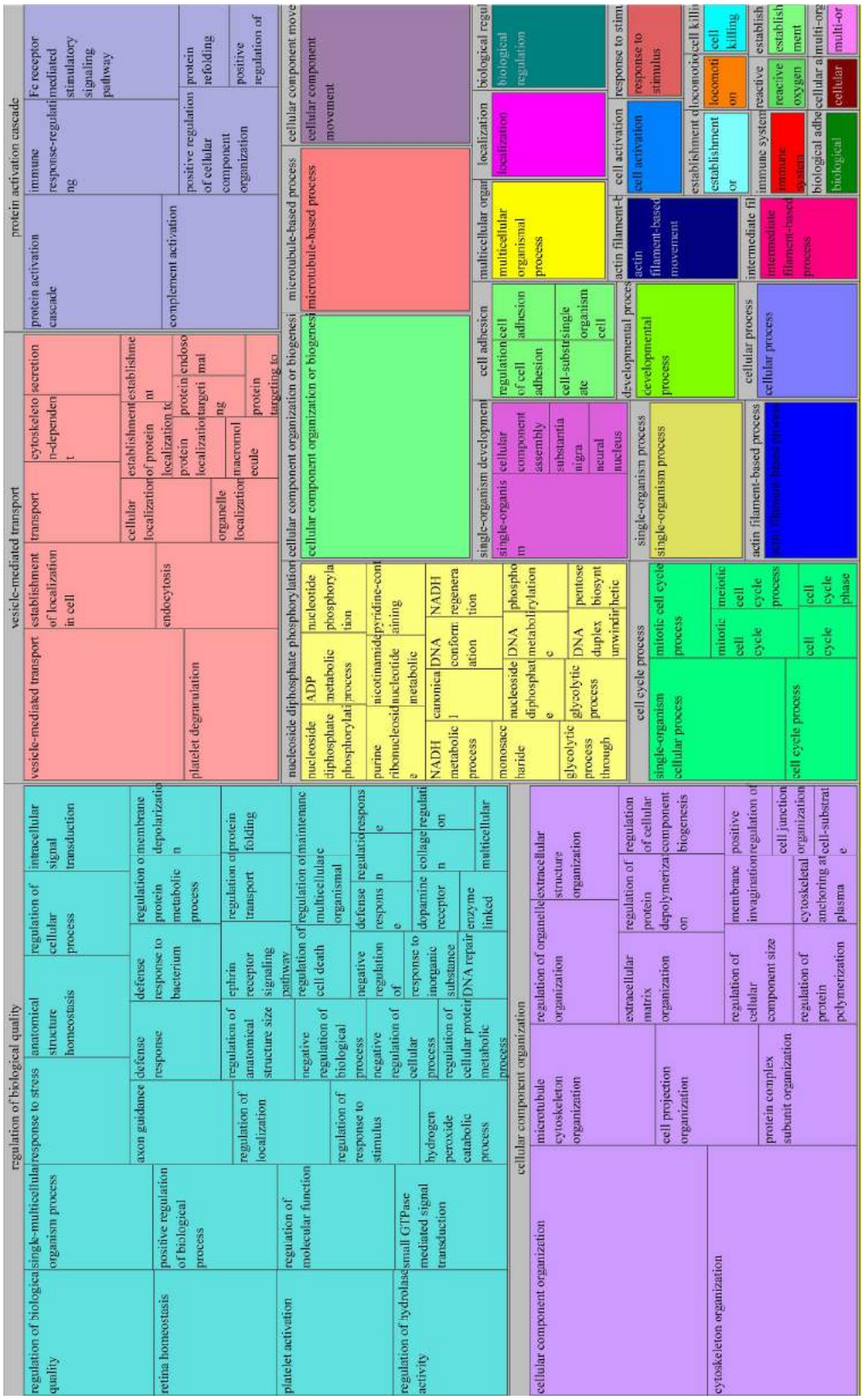
A

extracellular vesicular exosome
vesicle
extracellular vesicular exosome
blood microparticle
cytoskeletal part
cytoskeleton
membrane-bounded organelle
contractile fiber part
chromosome
contractile
smooth
nuclear
cytosol
cytoplasm
cytosol
intracellular part
cell part
cell cortex part
cell cortex
cytoplasmic part
perinuclear region of
macromolecular complex
macromolecular complex
organelle
organelle
immunoglobulin complex
protein complex
immunoglobulin
heterotrimeric
DNA packaging
voltage-gated
plasma membran
collagen trimer
WASH complex
haptoglobin-hem
lateral plasma
adherens junction
adherens junction
cell junction
cell junction
cell projection
cell projection
cell projection part
neuron part
neuron part
extracellular m
extracellular matrix part
extracellular
extracellular matrix
filopodium tip
myelin sheath
myelin sheath
cell periphery
cell periphery
site of polarized growth
site of pol
midbody
midbody
extrinsic c
extrinsic component
sarcolemma
sarcolemma
cell leadin
cell leading edge
cell body
cell body
contract
contrac
membra
membrane r
membr

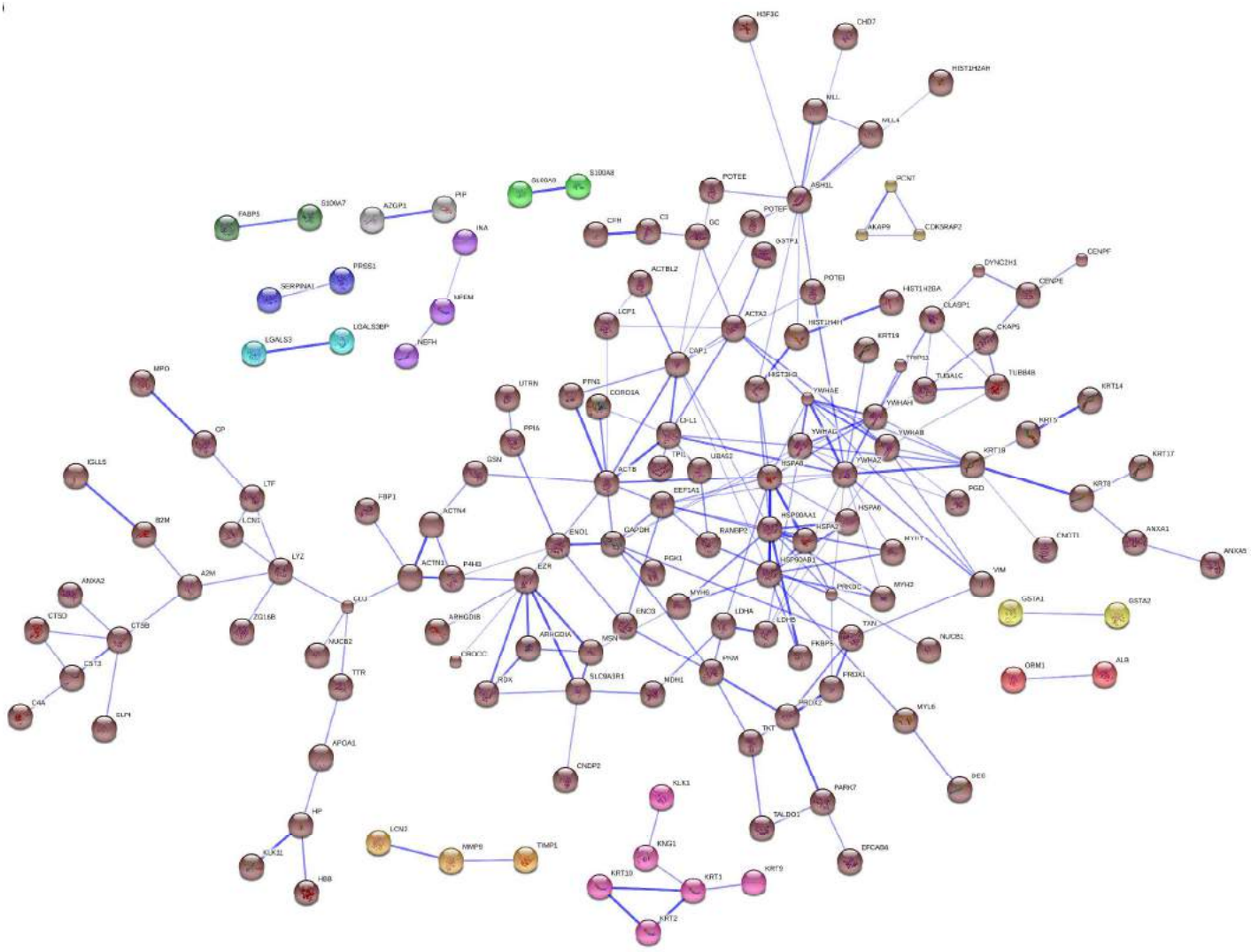
B


carbohydrate derivative binding
nucleoside phosphate binding
ion binding
heterocyclic compound binding
small molecule binding
organic cyclic compound binding
nucleic acid binding
motor activity
hydrolase activity, acting on acid anhydrides
microfilament motor activity
motor activity
actin-dependent ATPase activity
structural constituent of cytoskeleton
structural constituent of cytoskeleton
struc tural
extra cellul ar
protein binding
protein binding
binding
binding
poly(A) RNA binding
poly(A) RNA binding
guanyl-nucleotide exchange factor
guanyl-nucleotide exchange factor activity
enzyme regulator activity
enzyme regulator activity
cytoskeletal protein binding
cytoskeletal protein binding
GTPase binding
immunoglobulin receptor binding
calmodulin binding
glycoprotein bindin
ion channel binding
cell adhesion molecule binding
dystroglyca n
scaffold protein binding
G-protein beta/gamma-subunit complex binding
prctein complex scaffold
S100 protein binding
rec-ept or
dynein binding
phospholipid binding
anion binding
phosph olipid
ATP binding
calciu m
antigen binding
antigen binding
RNA binding
oxidoreductase activity, aci
peroxiredoxi n
oxidoreducta se
L-lactate dehydrogena e
lactate dehydrogen ase
Ras guanyl-nucleotide
Ras guanyl-nucleotide exchange factor activity
lipid binding
lipid binding
enzyme inhibitor activ
enzyme inhibitor activity
hydrolase activity
hydrolase activity
high voltage-g
high voltage-gate d
protein
protei n
antioxidant activity
antioxidant activity
euchromatin b
euchromatin
chro matin
DNA topoiso
DNA

C.



Supplementary figure S7 A). Treemap of healthy sputum proteome: cellular component. B) Treemap of healthy sputum proteome: Molecular function. C) Treemap of healthy sputum proteome: Biological process



Supplementary figure S8. Interaction network of the Healthy core sputum proteome. The network map was prepared using STRING (http://string.embl-heidelberg.de/).

Supplementary Table S1. Proteins with high population variance compared to measurement variance

UNIPROT ID	Protein name	pool CV %	Healthy Sputome CV %
S10A8_HUMAN	Protein S100A8 (Calgranulin-A)	7.978	365.521
HPT_HUMAN	Haptoglobin	9.576	651.593
AACT_HUMAN	Alpha antichymotrypsin	10.276	331.670
AMY1_HUMAN	Salivary Amylase	10.787	525.235
FETUA_HUMAN	Alpha-2-HS-glycoprotein	11.355	384.569
S10A9_HUMAN	Protein S100A9 (Calgranulin-B)	12.240	256.063
ILEU_HUMAN	Leukaocyte elastase inhibitor (SERPINB1)	14.431	338.942
CYTS_HUMAN	Cystatin-S	16.125	536.631
IGHM_HUMAN	Ig mu chain C region	16.305	404.459
PLSL_HUMAN	Plastin-2	16.830	206.644

These proteins are likely biologically relevant. Many of these are listed in the biomarker summary database. Supplementary excel file 1

Supplementary Table S2. Proteins with poor repeatability due to sample processing

Protein	Rank	Sample Coverage %	pool CV%	% poorly replicated pool	% poorly replicated sputome	Homologues?	Comments
A1ATR_HUMAN	241	43.6	425.12	**0.00**	12.84	No proteins with homolgous regions in search database	
SKT_HUMAN	234	44.8	266.94	**0.00**	11.61	No proteins with homolgous regions in search database	
TRRAP_HUMAN	246	42.8	241.75	**0.00**	13.08	No proteins with homolgous regions in search database	Nuclear
K1671_HUMAN	255	41.2	222.14	**0.00**	14.56	No proteins with homolgous regions in search database	
CNTLN_HUMAN	249	41.6	213.01	**0.00**	11.54	No proteins with homolgous regions in search database	Centrosome, centriole
MYH11_HUMAN	272	40	200.40	**0.00**	**0.00**	Multiple Myosin family members	
BASP1_HUMAN	124	67.2	196.75	5.26	1.79	No proteins with homolgous regions in search database	membrane protein
1433G_HUMAN	237	44	190.22	**0.00**	10.91	Similar to other 14-3-3 proteins but homology with 14-3-3F	
LDHB_HUMAN	87	80.8	180.97	9.52	4.46	Homologous regions to LDHA see below	
K1C18_HUMAN	252	41.2	169.52	17.65	10.68	Some proteins with high identity but minimal-no tryptic homology	
HPTR_HUMAN	260	40.8	163.28	**0.00**	10.78	Homologous regions with haptoglobin	
TRNK1_HUMAN	266	40.4	158.58	**0.00**	10.89	No proteins with homolgous regions in search database	
1433F_HUMAN	254	41.2	154.77	**0.00**	0.97	Similar to other 14-3-3 proteins but homology with 14-3-3E	
K1C10_HUMAN	89	79.2	145.96	4.55	3.54	Similar to other keratins. Some tryptic regions of homolgy	
LDH6A_HUMAN	171	56	126.66	**0.00**	12.14	Homologous regions to LDHA and B proteins see above	

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. Many of these have potential homologues which could interfere with their quantitation. The majority of these are at the lower end of the coverage spectrum.

Supplementary Table S3. Poor quantifying proteins in the mass spectrometer

Protein	Rank	Sample Coverage %	pool CV%	% poorly replicated pool	% poorly replicated sputome	Homologues?	Comments
POTEE_HUMAN	40	95.6	184.42	50.00	41.00	Multiple POTE family members	
GOGB1_HUMAN	132	64	140.09	18.75	33.75	No proteins with homolgous regions in search database	Membrane protein
PCNT_HUMAN	145	60.8	73.69	17.65	32.89	No proteins with homolgous regions in search database	Centrosome, centrioles, cilia
MYH6_HUMAN	189	51.6	146.93	42.86	32.56	Multiple Myosin family members	
HS90B_HUMAN	97	76.8	129.94	50.00	32.29	Multiple proteins with homologous regions	
MYH7_HUMAN	169	56.4	571.91	18.75	31.21	Multiple Myosin family members	
MYH7B_HUMAN	209	48.8	130.44	21.43	29.51	Multiple Myosin family members	
GOGA4_HUMAN	120	68	49.90	33.33	29.41	No proteins with homolgous regions in search database	Membrane associated
MYH4_HUMAN	232	45.2	135.92	15.38	28.32	Multiple Myosin family members	
TMPSD_HUMAN	170	56	149.17	20.00	27.86	No proteins with homolgous regions in search database	Membrane protein
CE290_HUMAN	146	60.8	445.27	26.67	27.63	No proteins with homolgous regions in search database	Centrosome, cilia
MYH3_HUMAN	222	47.2	128.56	20.00	27.12	Multiple Myosin family members	
ACTN3_HUMAN	263	40.4	402.26	53.33	26.73	Multiple actinin family members	
CK5P2_HUMAN	155	59.2	270.85	17.65	26.35	No proteins with homolgous regions in search database	Centrosome, centrioles
H90B2_HUMAN	175	55.2	979.89	20.00	25.36	Multiple proteins with homologous regions	

Proteins that display high CV of measurement across pools and high levels of poor replication between injections are likely to be poor quantifiers in the MS instrument. Many of these are from hydrophobic compartments or have potential homologues which could interfere with their quantitation. The majority of these are at the lower end of the coverage spectrum.

Supplementary Table S4. Top Canonical pathways enriched in the extended Healthy sputum proteome.

Ingenuity Canonical Pathways	Enrichment p-value)
Actin Cytoskeleton Signaling	2.51189E-22
Epithelial Adherens Junction Signaling	3.16228E-19
Remodeling of Epithelial Adherens Junctions	1.99526E-14
Signaling by Rho Family GTPases	7.94328E-14
RhoGDI Signaling	3.16228E-13
Cellular Effects of Sildenafil (Viagra)	1.25893E-12
RhoA Signaling	1.38038E-10
ILK Signaling	2.23872E-10
LXR/RXR Activation	2.69153E-10
Germ Cell-Sertoli Cell Junction Signaling	5.49541E-10
Ephrin B Signaling	2.29087E-09
Glycolysis I	8.91251E-09
14-3-3-mediated Signaling	4.16869E-08
Regulation of Actin-based Motility by Rho	4.7863E-08
Sertoli Cell-Sertoli Cell Junction Signaling	9.54993E-08
Gap Junction Signaling	1.20226E-07
FXR/RXR Activation	1.23027E-07
Breast Cancer Regulation by Stathmin1	2.75423E-07
Thrombin Signaling	3.16228E-07
Virus Entry via Endocytic Pathways	4.36516E-07

Supplementary Table S5. Top function and disease representatives in the extended healthy sputum proteome.

Diseases and Functions Category heirarchy	Diseases or Functions Annotation	Enrichment p-Value
Infectious Diseases, Inflammatory Disease, Respiratory Disease	Severe acute respiratory syndrome	1.02E-06
Cellular Movement, Immune Cell Trafficking	Leukocyte migration	2.43E-05
Infectious Diseases	Viral Infection	3.76E-05
Cellular Movement, Hematological System Development and Function, Immune Cell Trafficking	Cell movement of leukocytes	4.84E-05
Cell-To-Cell Signaling and Interaction	Response of granulocytes	1.30E-04
Inflammatory Response	Inflammatory response	1.94E-04
Cell-To-Cell Signaling and Interaction, Cellular Compromise, Cellular Function and Maintenance, Inflammatory Response	Respiratory burst of granulocytes	5.22E-04
Free Radical Scavenging	Generation of reactive oxygen species	7.85E-04

Cell-To-Cell Signaling and Interaction, Hematological System Development and Function, Inflammatory Response	Immune response of neutrophils	9.60E-04
Cell-To-Cell Signaling and Interaction, Cellular Compromise, Cellular Function and Maintenance, Hematological System Development and Function, Inflammatory Response	Respiratory burst of neutrophils	1.04E-03
Cell-To-Cell Signaling and Interaction, Hematological System Development and Function, Immune Cell Trafficking	Binding of antigen presenting cells	1.14E-03
Cell-To-Cell Signaling and Interaction, Hematological System Development and Function, Immune Cell Trafficking	Adhesion of granulocytes	1.16E-03
Organismal Survival	Organismal death	1.16E-03
Cell-To-Cell Signaling and Interaction, Hematological System Development and Function, Immune Cell Trafficking, Inflammatory Response	Binding of macrophages	1.21E-03
Cellular Movement, Hematological System Development and Function, Immune Cell Trafficking, Inflammatory Response	Chemotaxis of leukocytes	1.81E-03
Cellular Movement	Chemotaxis of cells	1.92E-03
Cell-To-Cell Signaling and Interaction, Cellular Growth and Proliferation, Hematological System Development and Function	Stimulation of leukocyte cell lines	2.13E-03
Cellular Movement, Hematological System Development and Function, Immune Cell Trafficking, Inflammatory Response	Chemotaxis of natural killer cells	2.35E-03
Cell-To-Cell Signaling and Interaction, Hematological System Development and Function, Immune Cell Trafficking	Adhesion of immune cells	2.45E-03
Cellular Assembly and Organization	Formation of rosettes	3.05E-03