**U-BIOPRED clinical adult asthma clusters linked to a subset of sputum *-omics***

Diane Lefaudeux1\* MSc, Bertrand De Meulder1\* PhD, Matthew J. Loza2 PhD, Nancy Peffer2 BS, Anthony Rowe3 PhD, Frédéric Baribaud2 PhD, Aruna T. Bansal4 PhD, Rene Lutter5 PhD, Ana R. Sousa6 PhD, Julie Corfield7 MSc, Ioannis Pandis8 PhD, Per S. Bakke9 MD, Massimo Caruso10 MD, Pascal Chanez11 MD, Sven-Erik Dahlén12MD , Louise J. Fleming13 MD, Stephen J. Fowler14 MD, Ildiko Horvath15 MD, Norbert Krug16 MD, Paolo Montuschi17 MD, Marek Sanak18 MD, Thomas Sandstrom19 MD, Dominic E. Shaw20 MD, Florian Singer 21 MD, Peter J. Sterk22 MD, PhD, Graham Roberts23 MD, Ian M. Adcock13 PhD, Ratko Djukanovic23 MD, Charles Auffray1 PhD, Kian. F. Chung13 MD and the U-BIOPRED Study Group #

* 1. \* Both contributed equally.
  2. #Other Consortium Study Group members in Acknowledgements
  3. 1European Institute for Systems Biology and Medicine, CIRI UMR5308, CNRS-ENS-UCBL-INSERM, Lyon, France
  4. 2Janssen Research and Development LLC, Spring House, USA
  5. 3Janssen Research and Development Ltd, High Wycombe, UK
  6. 4Acclarogen Ltd. St. John’s Innovation Centre, Cambridge, UK
  7. 5Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands
  8. 6Respiratory Therapeutic Unit, GSK, Stockley Park, UK
  9. 7AstraZeneca R&D Molndal, Sweden and Areteva R&D, Nottingham, UK
  10. 8Data Science Institute, Imperial College London, London, UK
  11. 9Department of Clinical Science, University of Bergen, Bergen, Norway
  12. 10Department of Clinical and Experimental Medicine, University of Catania, Italy
  13. 11Département des Maladies Respiratoires, Aix Marseille Université Marseille, France
  14. 12The Centre for Allergy Research, Karolinska Institutet, Stockholm, Sweden
  15. 13National Heart and Lung Institute, Imperial College & Biomedical Research Unit, Royal Brompton & Harefield NHS Trust, London, UK
  16. 14Centre for Respiratory Medicine and Allergy, The University of Manchester, UK
  17. 15Department of Pulmonology, Semmelweis University, Budapest, Hungary
  18. 16Fraunhofer Institute for Toxicology and Experimental Medicine, Hannover, Germany
  19. 17Faculty of Medicine, Catholic University of the Sacred Heart, Rome, Italy
  20. 18Department of Medicine, Jagiellonian University Medical School, Krakow, Poland
  21. 19Dept of Public Health and Clinical Medicine, Medicine, Umeå university, Umeå, Sweden 20Respiratory Research Unit, University of Nottingham, Nottingham, UK
  22. 21University Children’s Hospital Bern, Bern, Switzerland
  23. 22NIHR Respiratory Biomedical Research Unit, Clinical and Experimental Sciences, 23Faculty of Medicine, University of Southampton, Southampton, UK
  24. **Short Title:** U-BIOPRED clinical clusters of asthma

**Corresponding author:**

Professor K. F. Chung

National Heart and Lung Institute

Imperial College London

Dovehouse St,

London SW3 6LY,UK

Email : f.chung@imperial.ac.uk

This study is registered on ClinicalTrials.gov (identifier: NCT01982162).

**Declaration of all sources of funding:** 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).

**Word count:** 4124

## **ABSTRACT 246 w**

### **Background:** Asthma is a heterogeneous disease in which there is a differential response to asthma treatments. This heterogeneity needs to be evaluated so that a personalised management approach can be provided.

### **Objectives:** We stratified patients with moderate-to-severe asthma based on clinico-physiological parameters and performed an *-omics* analysis of sputum.

### **Methods:** Partition-around-medoid clustering was applied to a training set of 266 asthma participants from the European U-BIOPRED adult cohort using 8 pre-specified clinic-physiological variables. This was repeated in a separate validation set of 152 asthmatics. The clusters were compared based on sputum proteomic and transcriptomic data.

### **Results:** Four reproducible and stable clusters of asthmatics were identified. The training set cluster T1 consists of well-controlled moderate-to-severe asthmatics, while cluster T2 is a group of late-onset severe asthmatics with history of smoking and chronic airflow obstruction. Cluster T3 is similar to cluster T2 in terms of chronic airflow obstruction but is composed of non-smokers. Cluster T4 is predominantly composed of obese female uncontrolled severe asthmatics with increased exacerbations, but with normal lung function. The validation set exhibited similar clusters, demonstrating reproducibility of the classification. There were significant differences in sputum proteomics and transcriptomics between the clusters. The severe asthma clusters, T2, T3 and T4, had higher sputum eosinophilia than T1 with no differences in sputum neutrophil counts, exhaled nitric oxide and serum IgE levels.

### **Conclusion:** Clustering based on clinico-physiological parameters yielded 4 stable and reproducible clusters that associate with different pathobiological pathways.

1. **Clinical Implications:** The definition of four distinct clusters of asthma linked to different pathobiological pathways provides a better template for the phenotyping and personalised treatment of severe asthma, where high unmet needs remain.
2. **Capsule Summary:** Unsupervised clustering of asthma on clinical features alone has led to the definition of four phenotypes. Sputum ‘omics’ analysis has revealed different biological pathways pointing towards potential new treatments.
3. **Key words:** Severe asthma, clustering, sputum eosinophilia, partition-around-medoids algorithm
4. **Abbreviations:**

**ACQ:** Asthma Control Questionnaire

**ANOVA:** ANalysis Of VAriance

**BMI:** Body Mass Index

**CDF**: Cumulative Distribution Function

**COPD:** Chronic Obstructive Pulmonary Disease

**FDR:** False Discovery Rate

**FeNO:** Fractional exhaled Nitric Oxide

**FEV1:** Forced Expiratory Volume in 1 second

**FVC:** Force Vital Capacity

**ICS:** Inhaled CorticoSteroids

**IgE:** Immunoglobulin E

**OCS:** Oral CorticoSteroids

**PAM:** Partition-Around-Medoids

**SARP:** Severe Asthma Research Program

**Th2:** Lymphocyte T-helper 2

**U-BIOPRED:** Unbiased BIOmarkers for the PREDiction of respiratory diseases outcomes**INTRODUCTION**

Although clinicians have been focusing on the definition and classification of asthma severity and disease risk for the past decade, there is now a consensus that a deeper understanding of the basis of the heterogeneity of asthma is necessary in order to find targeted treatments for specific asthma phenotypes1. This is imperative for patients with severe asthma because this group of patients does not fully respond to currently-available asthma medications1 and is likely to constitute of a number of different asthma phenotypes2. There is therefore a need to improve the identification and definition of these phenotypes of asthma.

Cluster analysis using unsupervised statistical approaches has already led to the definition of clusters on the basis of similarities in clinical and inflammatory biomarkers3, 4. However, these studies have used relatively homogeneous populations and so may not reflect the real-life situation. One example is the exclusion of current or previous smokers with asthma, a group that may have an asthma-COPD overlap syndrome5. In addition, previously-derived clusters have not been linked to underlying biological profiles apart from the use of blood or sputum eosinophil counts. Other approaches have been to use unsupervised gene and protein omics data to cluster patients with asthma 6-8.

In this study, we used a robust clustering approach using clinical and physiological parameters that are available to the asthma physician in a broad range of participants with mild/moderate to severe asthma, including smokers and ex-smokers recruited in the U-BIOPRED project9. The second step was to explore the underlying pathobiological pathways of these clusters by examining the differential expression of the transcriptome of sputum cells and the proteome of sputum supernatants that exist between the clusters generated to determine whether they exhibit any differences in specific pathobiological pathways.

**METHODS**

### **U-BIOPRED cohorts**

We used a subset of the U-BIOPRED adult baseline data. The U-BIOPRED cohort comprises 509 asthmatics, both mild-moderate and severe and including non-smokers, ex-smokers and current smokers and 101 non-asthmatic control subjects. They had undergone detailed phenotypic characterisation using established standard operating procedures, as described previously9. The study participants were split randomly into training and validation datasets in a 2:1 ratio with the two groups being balanced in terms of asthma severity, age and gender. The validation group was used for internal replication. All participants gave signed informed consent to participate in the study which was approved by National Ethics Committees.

#### Clinical variables

The cluster analysis was focused on key variables that are readily accessible to the general practitioner representing important historical, clinical and physiological parameters underlying each participant with asthma. These variables were: age of onset of asthma symptoms, pack-years of cigarette smoking, body mass index (BMI), forced exhaled volume in 1 second (FEV1) as percentage of predicted value (FEV1% predicted), FEV1/FVC ratio (FVC: forced vital capacity), the average score of the 5 first questions of the Asthma Control Questionnaire (ACQ-5), self-reported numbers of exacerbations in the previous year and the daily dose of oral prednisolone or equivalent.

### **Data pre-processing and cluster analysis**

Box-Cox power transformation10 was used to approximate the data to a normal distribution using the powerTransform function from the R package *car* 11, which uses maximum likelihood to determine the best lambda. Data was then centred-scaled in order to ensure similar ranges for all the parameters and reduced using principal component analysis (PCA) ensuring that there was no correlation between the composite variables, thereby avoiding skewing of the analysis.

Clustering schemes are descriptive methods that group participants with similar characteristics. To determine similarity between participants, the Euclidean distance (which actually measures dissimilarity using the ordinary straight-line distance between 2 points) was used. Clustering was performed using the partition-around-medoid (PAM) algorithm, a more robust generalisation of the k-means method12. To assess the stability of clusters, bootstrapping (also known as consensus clustering) was performed by randomly removing 10% of the data and repeating the clustering for a total of 1,000 times13. The stability of the clusters was assessed by studying the cumulative distribution function (CDF) which, as represented in Figure 1A, describes the proportion of pairs of participants (on the y axis) that are clustered together in at most x percent of bootstrap iterations (on the x axis). Thus, if the curve is flat for example between x values of 0.2 and 0.8, this means that there are no pairs of participants that are clustered together between 20 and 80% of the iterations; *i.e.* they are almost never clustered together (<20%) or almost always clustered together (>80%). Clustering results were considered stable when the middle part of the CDF was flat.

To further define the stability of the clusters, we used an in-house objective called ‘deviation from ideal stability’ (Figure E1 in the Online Repository), which is at its best when it is close to zero. Additionally, to gain confidence in the existence of these clusters14, internal validity was checked using the Calinski and Harabasz index15, which measures the ratio of the between-clusters variance to the within-cluster variance such that the higher the value, the better defined are the clusters.

### **Sputum induction, transcriptomics and protein analytes**

### Sputum induction was performed following inhalation of hypertonic (0.9 to 4.5%) saline using a DeVilbiss 2000 Ultrasonic nebuliser (Somerset, PA, USA) according to a standardised protocol16. Sputum plugs were selected and liquefied using dithioerythritol. Differential cell counts were determined by assessment of a maximum of 500 to 1,000 inflammatory cells on Diff-Quick stained cytospins. Cytospin assessments were performed centrally with the outcome of the cytospin analysis determining the suitability of the sample for analysis by accepting only those samples with a cell viability of ≥ 50% and squamous cells of ≤ 40%.

Transcriptomic analysis was performed using the Affymetrix® HT HG-U133+ PM GeneChip on extracted RNA from sputum cells derived from cell pellets with a specific cut-off of ≤ 30% squamous cells. Technical and biological quality checks were performed following Affymetrix® recommendations with only RNA samples of high purity (RIN > 6.5) used for amplification; raw data were pre-processed using the robust microarray analysis (RMA) method from the *affy* R package17 to derive the expression matrix.

### From each one of the frozen aliquots of sputum supernatant, 1,129 analytes were quantified using the SomaScan™ v3 platform (SomaLogic®, Boulder, CO; *www.somalogic.com*) using SOMAmer® (Slow Off-rate Modified Aptamer) protein-binding reagents. These assays combine the best properties of antibodies and traditional aptamers, which are highly specific for the corresponding cognate proteins18. Analyte levels were reported as relative fluorescence units, cross-plate calibrated, and median normalised.

### **Statistical analysis**

All analyses were undertaken using the R software for statistical computing (version 3.1.2). Clinical variables were compared between clusters using analysis of variance (ANOVA) for multiple group comparison of normally-distributed variables. The Kruskal-Wallis test was used for multiple group comparison of ordered categorical or non-normally distributed variables and the χ² test was used for qualitative variables. To compare protein abundance or transcript expression, an ANOVA test was performed on the data (transformed with base 2 logarithm) adjusting for age and gender followed by a Tukey post-hoc pairwise comparison test. Protein analytes or probesets were defined to be consistently differentially abundant or expressed when their respective p-values were below 0.05 in both the training and validation sets analysed separately and when these sets were analysed together (preventing the inclusion of features that would have a different direction of change in the training and the validation sets). This allows for a reproducible and relatively stringent features selection process, and lowering the false positive rate despite not correcting the p-value for multiple testing. Both proteomics and transcriptomics datasets have been checked for any batch or site effect and corrected accordingly using ComBat method19.

Pathway enrichment analysis was performed using the results of the statistical analysis described above. The lists of contrast-specific features consistently found in both training and validation sets were submitted to the *g:Profiler* web-tool for enrichment analysis20. The p-values for the enrichment analysis were corrected for false discovery rate (FDR) with the Benjamini-Hochberg method21. Feature lists for each comparison were tested for enrichment against the KEGG22 and Reactome23 databases.

**RESULTS**

### **Participants**

### A total of 418 asthmatics out of 509 with a complete set of data for the 8 variables were available for analysis and were split into training (n=266) and validation (n=152) sets. The distribution of asthma severity, age and gender, and all 8 variables included in the clustering for the training and validation sets were not statistically different between the two sets although FEV1 (% predicted) and daily dose of oral corticosteroids (OCS) were incompletely balanced (p-values of 0.07 and 0.06 respectively) (Online Repository Table E1).

### **Training set clusters**

Consensus clustering on the training set was run in order to assess stability for a number of potential cluster numbers varying from two to ten. This resulted in the separation of two or four stable groups after resampling as defined by a flat middle part of the consensus CDF13 (Figure 1A), well-defined squares within the consensus matrix (Figure 1B) and by minimal values for the deviation from ideal stability index (Figure 1C top). These cluster numbers were also associated with the two highest Calinski and Harabasz indices (Figure 1C bottom), indicating that the clusters were more compact than the overall data. While separating into two and four clusters resulted in almost similar quality, four clusters were chosen for further analysis (denoted T1 to T4) to allow for a more precise sub-phenotype definition. Indeed, the two cluster allocation mainly regroups T1 with T4 and T2 with T3. Finally Figure 1D represents a heat-map of distances between the participants in the four clusters.

**Four-cluster analysis (T1 to T4)**

The four clusters are described in Table 1 and in Online Repository Table E2. Briefly, *Cluster T1* is composed of moderate-to-severe well-controlled asthmatics with normal FEV1, low sputum eosinophilia, almost no OCS use (6%) and a high proportion of atopic participants (84.1%). *Cluster T2* is mainly composed of overweight to obese (79% with BMI ≥ 25 kg/m² and 41% with BMI ≥ 30 kg/m²), late-onset severe asthmatics who smoked, with relatively poor control, severe airflow obstruction (mean FEV1: 58.9% predicted), and had the highest sputum and blood eosinophilia, with a lower proportion of atopic participants than in the other three clusters (55.6%). *Cluster T3* is similar to Cluster T2 except that the asthmatics were non-smokers, were less overweight, had poorer lung function and a higher proportion of atopic participants (70.6%). *Cluster T4* is mostly composed of obese female asthmatics (83% female, 88% with BMI ≥ 25 kg/m² and 56% with BMI ≥ 30 kg/m²), experiencing frequent exacerbations with poor asthma quality of life despite near normal lung function and 73.6% of positive atopy status. Fractional exhaled nitric oxide (FeNO) and serum immunoglobulin E (IgE) were not differentially distributed amongst the 4 clusters.

### **Validation set clusters**

### The same analysis was done on the validation set. It yielded five relatively stable clusters after resampling (denoted V1, V2, V3, V4a and V4b to align with the training set) as shown by a flat CDF and a low deviation from ideal stability (see Online Repository Figure E2 and Table E3). The Calinsky and Harasbaz index was slightly better for 4 clusters. The difference in the number of clusters compared to the training set might be due to the fact that the validation set was smaller. When comparing the training and validation clusters using the least statistical differences of clinical variables, cluster V1 was found to be similar to cluster T1, V2 to T2 and V3 to T3, while V4a combined with V4b was found to be similar to T4 (Online Repository Table E4). For ease of recall, clusters T1 and V1 will be referred to as Phenotype 1, T2 and V2 as Phenotype 2, T3 and V3 as Phenotype 3, T4 and V4a&4b as Phenotype 4.

### The distributions of the main clinical characteristics of the training and validation clusters were similar (Figure 2) with the exception of the V4a and V4b clusters covering 2 different parts of T4. V4a consists of less obese asthmatics associated with later onset of disease, lower OCS use and better asthma control when compared to V4b.

**Algorithm to predict clinical phenotype**

Support Vector Machine algorithm with a Gaussian radial basis kernel24 was used to predict phenotypes from the 8 clinical parameters. The model was trained on the training set only using a 10 fold cross-validation method to prevent overfitting with *caret25* and *kernlab24* R packages. The prediction model yielded an almost perfect accuracy of 97% on the training set. It predicted phenotype assignment on the validation set and achieved a very good accuracy rate of 86%. An xlsm file has been developed that can be used to predict the clinical phenotype (see Online Repository).

**Biological characterisation in a subset**

The results of proteomic and transcriptomic profiling in sputum samples were compared between the phenotypes to determine if they could represent a useful categorisation of asthma. Due to the fact that not all patients were able to produce any sputum or good-quality sputum for analysis and due to technical quality control, the number of participants used in the proteomic and transcriptomic analyses were 86 and 94, respectively. The clinical profiles of these participants who provided these samples was not different from those of the whole cohort as shown in Online Repository Table E5. Protein data was available in 86 participants (56 in the training set and 30 in the validation set). Ten proteins out of the 1,129 measured were identified as being consistently differentially abundant between phenotypes (Table II). This number of hits was too small to allow for any meaningful pathway enrichment analysis. Sputum transcript expression data was available for 94 participants (56 in the training set and 38 in the validation set). A total of 345 transcripts (291 annotated) were found to be consistently significantly differentially expressed in at least one of the pairwise comparisons between the phenotypes (Online Repository Table E6). Pathway enrichment results are shown in Table III.

#### Differential protein abundance in sputum supernatants

Both the comparison of Phenotype 2 (severe asthma (ex-)smokers) and Phenotype 3 (severe asthma non-smokers) to Phenotype 1 (well-controlled asthma) highlighted IL-16, a natural ligand of CD4 and CD9 that induces preferential migration of human T-regulatory cells26, as being elevated in the more severe phenotypes. Additionally, compared to Phenotype 2, there was greater abundance in Phenotype 1 of (i) CTAP-III (CXCL7), a potent chemoattractant and activator of neutrophils; (ii) GM-CSF, which controls the production, differentiation and function of granulocytes and macrophages; and (iii) Trypsin 2, which degrades the extracellular matrix. On the contrary, HAPLN1, involved in cell adhesion, was less abundant.

Phenotype 3 was associated with reduced levels of Cathepsin G involved in connective tissue remodelling at site of inflammation, as compared to Phenotypes 1 and 4. Moreover, Phenotype 3 also exhibited elevated levels of ARSB, an arylsulfatase involved in cell adhesion and migration regulation, and PSA2, a member of peptidase T1A family, when compared to Phenotype 1. Lastly, LYN kinase and FUT5 (fucosyl transferase 5) were found to be decreased in Phenotype 3 when compared to Phenotype 2.

**Differential transcript expression in sputum cells**

Comparing Phenotype 2 to Phenotype 1 yielded 8 differentially-expressed genes, 2 of which are linked to the haematopoietic cell lineage pathway (CSF1 and CD1B), both being more expressed in Phenotype 2. The comparison of Phenotype 3 to Phenotype 1 highlighted 147 genes, 5 of them encoding the proteins CTSB, PDIA3, CD4, CD74, CALR that are linked to antigen processing and presentation pathway. Pathway enrichment analysis of Phenotype 3 compared to Phenotype 2 revealed pathways related to the regulation of the actin cytoskeleton (ITGB1, ITGB8, FN1, DIAPH2, F2R, ACTN2), and to fibronectin matrix formation (ITGB1, FN1), potentially linked to the effect of smoking in severe asthma. 3 probesets, one of which annotated to a known gene, DAGLB, that encodes for the enzyme diacylglycerol lipase, were highlighted in the comparison of Phenotype 4 to Phenotype 2, therefore no pathway enrichment analysis was done. The comparison of Phenotype 4 to Phenotype 3 revealed 14 differentially-expressed genes including those encoding proteins related to the cell cycle and growth factor regulating pathways (MAPK1, E2F1 and SPRY2) and to the modulation of immune system responses, particularly the interferon signalling pathway (OASL, OAS3 and TRIM14).

## **DISCUSSION**

Using the partition-around-medoid (PAM) clustering algorithm and a bootstrapping method on the large U-BIOPRED cohort of participants with moderate to severe asthma, we have identified 4 clusters of asthma: one composed of well-controlled asthmatics with almost normal lung function but on low to high doses of ICS and the three others of severe asthmatics. Two of the clusters relate to chronic airflow obstruction with one cluster associated with smokers and ex-smokers with late onset asthma who had the highest blood and sputum eosinophil counts, while the third cluster is associated with non-smokers on oral corticosteroid therapy. Finally, the fourth cluster of severe asthma relates to obese female asthmatic patients with recurrent exacerbations with near normal lung function. Because of a bias in the patient recruitment process, there are significantly more patients from one site with a history of smoking and associated with cluster T2, so much so that the clinical variables for centre and for smoking status were confounded. We chose not to adjust the p-values for the centre effect as it would remove part or all of the variability associated with the smoking status.

Our phenotypes are quite distinguishable in terms of controlled versus uncontrolled asthma (Phenotype 1 versus Phenotypes 2, 3 and 4), airflow obstruction versus normal lung function (Phenotypes 2 and 3 versus Phenotypes 1 and 4), and infrequent exacerbations versus frequent exacerbations (Phenotypes 1, 2 and 3 versus Phenotype 4). Thus, our robust approach to clustering on the basis of clinico-physiological parameters has yielded phenotypes characterised on the basis of asthma control, airflow obstruction, recurrent exacerbations and oral corticosteroid dependence, which are all well-known features of severe asthma. The 3 clusters of predominantly severe asthma (clusters 2, 3 and 4) had the highest incidence of nasal polyps and exacerbations including admission to intensive care unit in the past year compared to Cluster 1. In addition, these 3 clusters also had the greatest use of rescue inhalers, of oral corticosteroid, and despite this, also had the higher ACQ5 scores. The differential expression of proteins and genes measured in sputum has also provided some insight into the potential pathophysiological pathways that may govern these phenotypes, particularly those related to the characteristics of severe asthma namely chronic airflow obstruction and frequent exacerbations.

In contrast to the training set, clustering of the validation set resulted in one additional cluster, even though the results were not as stable to resampling as in the training set, as shown by the cumulative distribution function and the deviation from ideal stability index. A potential reason for this may relate to the fact that we had fewer participants in the validation set compared to the training set thus increasing the difficulty of finding stable (to resampling) clusters. Furthermore, FEV1 (% predicted) and oral corticosteroid doses, two of the variables included in the clustering were slightly different between the training and validation sets. Nevertheless, the training and validation sets clusters shared similarities; cluster 4 being divided in the validation set into 2 clusters. This provides an internal replication of the clusters in our cohort..

Our clusters exhibit similarities with some of the clusters previously reported in two similar cohorts with mild/moderate and severe asthmatics, namely the SARP4 and the Leicester cohorts3, even though they used different clustering algorithms (Ward hierarchical clustering and k-means, respectively). In relation to the SARP cohorts, Phenotype 1 relates to SARP cluster 2, Phenotype 3 to SARP clusters 4 and 5, and Phenotype 4 to SARP cluster 3. One rather unique feature of our study is the inclusion of a smoking or ex-smoking cohort of severe asthma patients that were grouped principally in a late onset, severe airflow obstruction cluster with high blood and eosinophil counts and 55% of the group with evidence of atopy. These patients represent a group of asthma with features of COPD, namely chronic airflow obstruction, fulfilling the criteria of the asthma-COPD overlap syndrome5. A similar cluster has been previously reported27, 28, although in one cluster of smoking asthmatics, the degree of airflow obstruction was minimal but the cohort that was studied was not one of severe asthma29.

In the Leicester study3, the sputum eosinophil count was also used in the clustering, generating the late-onset, obese female severe asthmatics cluster with low sputum eosinophil counts, which is similar to Phenotype 4. In our clusters, blood and sputum eosinophil counts varied within each of these clusters with the highest to be found in the smoking and ex-smoking patients in Phenotype 2, supporting further the concept of the asthma-COPD overlap syndrome5. However, the clinical clusters did not segregate according to levels of serum IgE or FeNO. Furthermore, the levels of sputum periostin used as a biomarker of Th2-associated protein did not differ amongst the 4 groups.

These 4 distinct phenotypes of asthma that would be recognisable by the clinician experienced in seeing patients with severe asthma would allow the patients to be segregated into these clinical characteristics associated with severe asthma. These clusters were also characterised by different pathobiological pathways. Using very strict criteria for defining the differentially-abundant proteins by examining for consistency of their expression in both the training and validation sets analysed separately, we found that IL-16 (lymphocyte chemoattractant factor) was the only protein to be detected as differentially abundant when comparing both severe asthma clusters with airflow obstruction (present in Phenotypes 2 and 3) to well-controlled asthmatics (Phenotype 1). IL-16 has been previously associated with asthma, and shown to be expressed in abundance in epithelial cells after histamine challenge, in bronchoalveolar lavage fluid after an allergen challenge, in airway epithelium and CD4+ T-cells of airway biopsies30-32. Phenotype 2 showed higher levels than Phenotype 1 of CTAP III (CXCL7) and GM-CSF, which might be a reflection of the effect of smoking exposure, since CXCL7 has been used as a biomarker for the risk of lung cancer, and since GM-CSF mediates cigarette smoke-induced lung neutrophilia33, 34. In addition, higher levels of CXCL7 and GM-CSF has been shown in chronic obstructive pulmonary disease (COPD) secondary to cigarette smoking35, 36. On the other hand, Phenotype 3 showed decreased sputum levels of cathepsin G compared to Phenotype 1 and to Phenotype 4, in which raised systemic levels have been linked to neutrophilic asthma6.

LYN kinase was found to be reduced in Phenotype 3 (non-smoking severe obstructed asthmatics) when compared to Phenotype 2 (smoking or ex-smoking asthmatics). LYN kinase is a SRC kinase that controls GATA-3 and induces Th2 cell differentiation37 as well as the susceptibility of epithelial cells to their response to cigarette smoke extracts38. It has also been implicated in increasing asthma severity in mouse asthma models39.

Comparing the gene expression between Phenotypes 2 and 3 revealed pathways related to the regulation of actin cytoskeleton, and to fibronectin matrix formation. The comparison of Phenotype 4 (obese and exacerbation-prone asthmatics) with Phenotype 3 (airflow obstructed asthmatics), by contrast, yielded differential gene pathways related to immune cytokine signalling, particularly interferon signalling and regulation of fibroblast growth factor (FGF) and the signalling of FGF receptor (FGFR). These specific pathways may be involved in important pathophysiological aspects underlying the clinical phenotypes identified through this clustering approach based initially on clinico-physiological features.

Some of the limitations and biases within this analysis need to be highlighted. First, as in any clinical study, the cohort is biased by its inclusion and exclusion criteria (as discussed in detail in reference8) but we have been as inclusive as possible. Secondly, cluster analysis is a descriptive method and groups can be defined even when there is no underlying structure in the data; this limitation was addressed by assessing stability, separation and reproducibility of the clusters. Moreover, the choice of clinical variables may condition the type of clusters that is found but the choice of variables we used can be justified by their relevance to day-to-day clinical practice. The proof that the choice was reasonable is in the description of clinical cohorts that makes sense to the clinician. Finally, unsupervised clustering on the basis of the transcriptomic and proteomic data remains another powerful approach towards molecular phenotyping, work which is currently being performed in UBIOPRED.

**Conclusion**

The four phenotypes of asthma that we describe from the U-BIOPRED cohort have distinct clinical and molecular characteristics that should prove useful to the clinician in directing management of the particularly severe asthma phenotypes. One phenotype is associated with smoking, emphasising its influence on asthma. The differential molecular characteristics of the four phenotypes are not only potentially useful biomarkers of asthma severity, but they also represent a starting point for drug discovery efforts and the development of better treatments. This will pave the way towards a more personalised approach to asthma management.

**ACKNOWLEDGMENTS**

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 thank all the members of each recruiting Centre for the recruitment and assessment of the participants.

#**U-BIOPRED consortium study group members**

Nora Adriaens1, Hassan Ahmed2, Antonios Aliprantis3, Kjell Alving4, Philipp Badorek5, David Balgoma6, Clair Barber7, An Bautmans8, Annelie F. Behndig9, Elisabeth Bel1, Jorge Beleta10, Ann Berglind6,11, Alix Berton12, Jeanette Bigler13, Hans Bisgaard14, Grazyna Bochenek15, Michael J. Boedigheimer13, Klaus Bøonnelykke14, Joost Brandsma16, Armin Braun5, Paul Brinkman1, Dominic Burg17, Davide Campagna18, Leon Carayannopoulos19, João P. Carvalho da Purfição Rocha20, Amphun Chaiboonchoe2, Romanas Chaleckis6, Courtney Coleman21, Chris Compton22, Arnaldo D’Amico23, Barbro Dahlén6, 24, Jorge De Alba10, Pim de Boer25, Inge De Lepeleire8, Tamara Dekker1, Ingrid Delin6, Patrick Dennison6, 26, Annemiek Dijkhuis1, Aleksandra Draper27, Jessica Edwards21, Rosalia Emma18, Magnus Ericsson24, Veit Erpenbeck28, Damijan Erzen29, Cornelia Faulenbach5, Klaus Fichtner29, Neil Fitch27, Breda Flood21, Urs Frey30, Martina Gahlemann31, Gabriella Galffy32, Hector Gallart6, Trevor Garret27, Thomas Geiser33, Jilaiha Gent20, Maria Gerhardsson de Verdier12, David Gibeon34, Cristina Gomez6, Kerry Gove7, Neil Gozzard35, Yi-Ke Guo36, Simone Hashimoto1, John Haughney37, Gunilla Hedlin6, 11, Pieter-Paul Hekking1, Elisabeth Henriksson24, Lorraine Hewitt7, Tim Higgenbottam38, Uruj Hoda20, Jans Hohlfeld5, Cecile Holweg39, Peter Howarth7, Richard Hu13, Sile Hu34, Xugang Hu13, Val Hudson21, Anna J. James6, Juliette Kamphuis25, Erika J. Kennington21, Dyson Kerry40, Matthias Klüglich29, Hugo Knobel41, Richard Knowles42, Alan Knox43, Johan Kolmert6, Jon Konradsen6,11, Maxim Kots44, Linn Krueger30, Scott Kuo34, Maciej Kupczyk6, Bart Lambrecht45, Ann-Sofie Lantz6, 11, Lars Larsson12, Nikos Lazarinis24, Saeeda Lone-Satif1, Lisa Marouzet7, Jane Martin7, Sarah Masefield46, Caroline Mathon6, John G.Matthews39, Alexander Mazein2, Sally Meah34, Andrea Maiser34, Andrew Menzies-Gow20, Leanne Metcalf21, Roelinde Middelveld6, Maria Mikus47, Montse Miralpeix10, Philips Monk48, Nadia Mores49, Clare S. Murray50, 51, Jacek Musial15, David Myles22, Shama Naz6, Katja Nething29, Ben Nicholas52, Ulf Nihlen12, Peter Nilsson47, Björn Nordlund4, 6, Jörgen Östling12, Antonio Pacino53, Laurie Pahus54, Susanna Palkonnen55, Stelios Pavlidis34, Giorgio Pennazza23, Anne Petrén6, Sandy Pink7, Anthony Postle52, Pippa Powel46, Malayka Rahman-Amin21, Navin Rao56, Lara Ravanetti1, Emma Ray7, Stacey Reinke6, Leanne Reynolds21, Kathrin Riemann29, John Riley22, Martine Robberechts8, Amanda Roberts21, Christos Rossios34, Kirsty Russell34, Michael Rutgers25, Giuseppe Santini49, Marco Sentoninco26, Corinna Schoelch29, James P.R. Schofield17, Wolfgang Seibold29, Ralf Sigmund29, Marcus Sjödin6, Paul J.Skipp17, Barbara Smids1, Caroline Smith7, Jessica Smith21, Katherine M. Smith43, Päivi Söderman11, Adesimbo Sogbesan20, Doroteya Staykova57, Karin Strandberg24, Kai Sun34, David Supple21, Marton Szentkereszty32, Lilla Tamasi32, Kamran Tariq7, 26, John-Olof Thörngren24, Bob Thornton19, Jonathan Thorsen14, Salvatore Valente23, Wim van Aalderen1, Marianne van de Pol1, Kees van Drunen1, Marleen van Geest12, Jenny Versnel21, Jorgen Vestbo50, 51, Anton Vink41, Nadja Vissing14, Christophe von Garnier33, Arianne Wagerner1, Scott Wagers27, Frans Wald29, Samantha Walker21, Jonathan Ward58, Zsoka Weiszhart32, Kristiane Wetzel29, Craig E. Wheelock6, Coen Wiegman34, Siân Williams37, Susan J. Wilson58, Ashley Woodcock50, 51, Xian Yang34, Elizabeth Yeyashingham59, Wen Yu13, Wilhelm Zetterquist4, 6, Koos Zwinderman1

1: Academic Medical Centre, University of Amsterdam, The Netherlands; 2: European Institute for Systems Biology and Medicine, CIRI UMR5308, CNRS-ENS-UCBL-INSERM, Lyon, France; 3: Merck Research Laboratories, Boston, USA; 4: Department of Women’s & Children’s Health, Uppsala University, Sweden; 5: Fraunhover Institute for Toxicology and Experimental Medicine, Hannover, Germany; 6: Centre for Allergy Research, Karolinska Institutet, Stockhlom, Sweden; 7: NIHR Southampton Respiratory Biomedical Research Unit and Clinical and Experimental Sciences, Southampton, UK; 8: MSD, Brussels, Belgium; 9: Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden; 10: Almirall S.A., Barcelona, Spain; 11: Department of Women’s & Childern’s Health, Karolinska Institutet, Stockholm, Sweden; 12: AstraZeneca, Mölndal, Sweden; 13: Amgen Inc. Seattle, USA; 14: Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Genofte Hospital, University of Copenhagen, Copenhagen, Denmark; 15: Department of Internal Medicine, Jagiellonian University Medical College, Krakow, Poland; 16: Faculty of Medicine, Southampton University, Southampton, UK; 17: Centre for Proteomics Research, Institute of Life Sciences, University of Southampton, Southampton, UK; 18: Department of Clinical and Experimental Medicine, University of Catania, Catania, Italy; 19: MSD, Kenilworth, USA; 20: Royal Brompton and Harefield NHS Fondation Trust, London, UK; 21: Asthma UK, London, UK; 22: Respiratory Therapeutic Unit, GSK, London, UK; 23: University of Rome ‘Tor Vergata’, Rome, Italy; 24: University Hospital, Karolinska Institutet, Stockholm, Sweden; 25: Longfonds, Amersfoort, The Netherlands; 26: NIHR-Wellcome Trust Clinical Research Facility, Faculty of Medicine, University of Southampton, Southampton, UK; 27: BioSci Consulting, Maasmechelen, Belgium; 28: Translational Medicine, Respiratory Profiling, Novartis Institute for Biomedical Research, Basel, Switzerland; 29: Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany; 30: University Children’s Hospital, Basel, Switzerland; 31: Boehringer Ingelheim (Schweiz) GmbH, Basel, Switzerland; 32: Semmelweis Universty, Budapest, Hungary; 33: Department of Respiratory Medicine, University Hospital Bern, Bern, Switzerland; 34: National Hearth and Lund Institute, Imperial College, London, UK; 35: UCB, Slough, UK; 36: Data Science Institute, Imperial College, London, UK; 37: International Primary Care Respiratory Group, Aberdeen, Scotland; 38: Allergy Therapeutics, West Sussex, UK; 39: Respiratory and Allergy Diseases, Genentech, San Francisco, USA; 40: CromSource, Stirling, UK; 41: Philips Research Laboratories, Eindhoven, The Netherlands; 42: Arachos Pharma, Stevenage, UK; 43: Respiratory Research Unit, University of Nottingham, UK; 44: Chiesi Pharmaceuticals, SPA, Parma, Italy; 45: University of Gent, Gent, Belgium; 46: European Lung Foundation, Sheffield, UK; 47: Science for Life Laboratory & The Royal Institute of Technology, Stockholm, Sweden; 48: Synairgen Research Ltd, Southampton, UK; 49: Università Cattolica del Sacro Cuore, Roma, Italy; 50: Centre for Respiratory Medicine and Allergy, Institute of Inflammation and Repair, University of Manchester, Manchester, UK; 51: University Hospital of South Manchester, Manchester Academic Health Sciences Centre, Manchester, UK; 52: Faculty of Health Science, Southampton University, Southampton, UK; 53: Lega Italiana Anti Fumo, Catania, Italy; 54: Assistance Publique des Hôpitaux de Marseille, Clinique des bronches, allergies et sommeil, Espace EthiqueMéditerranéen, Aix-Marseille Université, Marseille, France; 55: European Federation of Allergy and Airways Diseases Patient’s Associations, Brussels, Belgium; 56: Janssen Research & Development, USA; 57: Centre for Biological Sciences, University of Southampton, Southampton, UK; 58: Histochemistry Research Unit, Faculty of Medicine, University of Southampton, Southampton, UK; 59: UK Clinical Operations, GSK, Stockley Park, UK

**Table I: Characteristics of the 4 asthma clusters from the training set□**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variables** | **Missing or uncertain**  **T1 / T2 / T3 / T4** | **Cluster T1 (n = 69)** (moderate to severe, well controlled) | **Cluster T2 (n = 56)** (severe late onset asthma with airway obstruction, high BMI, smoking and OCS use) | **Cluster T 3 (n = 68)** (severe asthma with airway obstruction, OCS use but no smoking history) | **Cluster T4 (n = 73)** (severe asthma with female predominance, high BMI, frequent exacerbations, OCS use but no history of smoking or airway obstruction) | **P-Value** |
| Age (years) |  | 42.9 ± 15.6 | 57.4 ± 10.1 | 52.5 ± 15 | 47.5 ± 13.6 | **< 0.001 1** |
| Female gender |  | 55.07% | 57.14% | 47.06% | **83.56%** | **< 0.001 3** |
| Total daily OCS dose (normalised to milligrams prednisolone)† |  | 0 (0 – 0) | 0 (0 – 10) | **4 (0 – 10)** | 0 (0 – 10) | **< 0.001 2** |
| Asthma onset (years)† |  | 17 (5 – 30) | **42.5 (30.8 – 52)** | 17.5 (5.75 – 37) | 20 (7 – 37) | **< 0.001 2** |
| FEV1% predicted† |  | **88.5 ± 16.9** | 58.9 ± 15.6 | 48.5 ± 13.8 | **79.2 ± 15.4** | **< 0.001 1** |
| FEV1/FVC† |  | 0.737 ± 0.0836 | **0.557 ± 0.0978** | **0.505 ± 0.0787** | 0.741 ± 0.0832 | **< 0.001 1** |
| ACQ\_5† |  | **0.8 (0.25 – 1.6)** | 2 (1.2 – 2.8) | 2.4 (1.6 – 3.6) | 2.6 (1.8 – 3.2) | **< 0.001 2** |
| Number of exacerbations in past year† |  | 0 (0 – 1) | 1 (0.75 – 3) | 2 (1 – 3.25) | **3 (2 – 4)** | **< 0.001 2** |
| BMI (kg/m²)† |  | 25.1 (21.8 – 28.4) | 29 (26 – 33.3) | 25.4 (23.4 – 28.5) | **31.6 (26.8 – 35.8)** | **< 0.001 2** |
| Pack years† |  | 0 (0 – 0) | **16 (4.94 – 25.9)** | 0 (0 – 0) | 0 (0 – 0) | **< 0.001 2** |
| High ICS dose | **2 / 5 / 2 / 5** | **37.3%** | 98.0% | 96.9% | 95.5% | **< 0.001 3** |
| Blood eosinophils (x103/µl) | 0 / 1 / 1 / 4 | 0.199 (0.1 – 0.3) | **0.301 (0.119 – 0.56)** | 0.22 (0.1 – 0.494) | 0.2 (0.0997 – 0.385) | **0.0273 2** |
| Blood neutrophils (x103/µl) | **0 / 1 / 1 / 4** | **3.98 (3.10 – 4.87)** | 5.09 (3.97 – 7.15) NA's: 1 | 5.13 (3.74 – 7.61) NA's: 1 | 4.73 (3.62 – 6.70) NA's: 4 | **< 0.001 2** |
| Sputum eosinophils (%) | **39 / 20 / 44 / 45** | **0.78 (0.252 – 5.88)** | 4.88 (1.42 – 20) NA's: 20 | 3.67 (1.01 – 29.6) NA's: 44 | 2.42 (0.288 – 7.06) NA's: 45 | **0.0147 2** |
| Sputum neutrophils (%) | 39 / 20 / 44 / 45 | 54.4 (27.4 – 64.1) | 59.7 (44.0 – 70.5) | 63.4 (40.2 – 86.3) | 49.4 (31.5 – 70.0) | 0.268 2 |
| FeNO (ppb) | 1 / 1 / 0 / 10 | 24 (14 – 42.9) | 29 (15.5 – 53) | 33.8 (21.2 – 56.5) | 26.5 (15.5 – 48) | 0.332 2 |
| Total IgE (iu/ml) | 0 / 1 / 2 / 3 | 131 (54 – 316) | 140 (47 – 310) | 113 (41.5 – 352) | 124 (41.2 – 319) | 0.995 2 |
| Atopy (positive) SPT or specific IgE | 0 / 2 / 0 / 1 | 84.1% | **55.6%** | 70.6% | 73.6% | **0.006 3** |

□ Values are presented as mean ± SD, median (1st – 3rd quartiles) or percentages.

1 ANOVA

2 Kruskal-Wallis test

3 χ² test

† Variables included in the clustering

**Table II:** **Differentially abundant proteins in sputum supernatants between clusters in a subset of patients**.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Target Name** | **GENE Symbol** | **Phenotype 2 vs 1**  **(n = 28 vs 22)** | | **Phenotype 3 vs 1**  **(n = 20 vs 22)** | | **Phenotype 3 vs 2**  **(n = 20 vs 28)** | | **Phenotype 4 vs 3**  **(n = 16 vs 20)** | |
| log2(FC)3 | P-value1, 2 | log2(FC)3 | P-value1, 2 | log2(FC)3 | P-value1, 2 | log2(FC)3 | P-value1, 2 |
| **IL-16** | **IL16** | 0.968 | **9.87e-6 (4.18e-3)** | 1.034 | **2.04e-5 (0.023)** | 0.066 |  | -0.153 |  |
| **CTAP-III** | **PPBP** | 1.096 | **4.61e-4 (9.13e-3)** | 1.141 |  | 0.045 |  | -0.432 |  |
| **Trypsin 2** | **PRSS2** | 0.930 | **1.85e-4 (6.72e-3)** | 0.833 |  | -0.097 |  | -0.283 |  |
| **GM-CSF** | **CSF2** | 0.898 | **6.69e-4 (0.0111)** | 0.242 |  | -0.656 |  | 0.0358 |  |
| **HPLN1** | **HAPLN1** | - 0.571 | **6.68e-5 (5.39e-3)** | -0.414 |  | 0.157 |  | -0.027 |  |
| **Cathepsin G** | **CTSG** | -0.682 |  | - 2.088 | **3.91e-5 (0.0221)** | -1.406 |  | 1.556 | 2.29e-3 (0.590) |
| **ARSB** | **ARSB** | 0.417 |  | 0.966 | 9.53e-4 (0.131) | 0.549 |  | -0.548 |  |
| **PSA2** | **PSMA2** | 0.140 |  | 0.955 | 7.82e-4 (0.126) | 0.815 |  | -0.456 |  |
| **LYN** | **LYN** | 1.038 |  | -1.338 |  | - 2.376 | **4.32e-5 (0.0488)** | 1.9002 |  |
| **FUT5** | **FUT5** | 0.283 |  | -1.821 |  | - 2.104 | 6.96e-4 (0.0982) | 1.537 |  |

1 Tukey p-value adjusted for age and gender in the training and validation set pooled together of the proteins that were consistently found both in the training and the validation for that specific contrast.

2 P-values are shown as Nominal (FDR corrected). Bold indicates that the difference is still significant even when correcting for false discovery rate using Benjamini-Hochberg method.

3 Fold change is given in base 2 logarithm, if it is positive in X vs Y it means that the analyte is more abundant in X than Y, negative values mean less abundant.

**Table III: Enriched pathways1 with differentially expressed genes between clusters in a subset of patients.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Term** | **Source** | **Size** | **Phenotype 2 vs 12,3**  **(n = 24 vs 31)** | **Phenotype 3 vs 12,3**  **(n = 20 vs 31)** | **Phenotype 3 vs 22,3**  **(n = 20 vs 24)** | **Phenotype 4 vs 32,3**  **(n = 19 vs 20)** |
| Hematopoietic cell lineage | KEGG | 84 | 0.05 (2) |  |  |  |
| Antigen processing and presentation | KEGG | 69 |  | 0.05 (5) |  |  |
| Regulation of actin cytoskeleton | KEGG | 216 |  |  | 0.05 (6) |  |
| Arrhythmogenic right ventricular cardiomyopathy | KEGG | 74 |  |  | 0.032 (4) |  |
| Melanoma | KEGG | 72 |  |  |  | 0.032 (2) |
| Glioma | KEGG | 66 |  |  |  | 0.027 (2) |
| Non-small cell lung cancer | KEGG | 57 |  |  |  | 0.020 (2) |
| MicroRNAs in cancer | KEGG | 258 |  |  |  | 0.024 (3) |
| Prostate cancer | KEGG | 90 |  |  |  | 0.05 (2) |
| Pancreatic cancer | KEGG | 67 |  |  |  | 0.028 (2) |
| Bladder cancer | KEGG | 38 |  |  |  | 0.009 (2) |
| Chronic myeloid leukemia | KEGG | 74 |  |  |  | 0.034 (2) |
| N-glycan trimming in the ER and Calnexin/Calreticulin cycle | REAC | 13 |  | 0.05 (3) |  |  |
| Fibronectin matrix formation | REAC | 74 |  |  | 0.05 (2) |  |
| Negative regulation of FGFR signaling | REAC | 16 |  |  |  | 0.002 (2) |
| Spry regulation of FGF signaling | REAC | 16 |  |  |  | 0.002 (2) |
| Oncogene Induced Senescence | REAC | 31 |  |  |  | 0.010 (2) |
| Pre-NOTCH Expression and Processing | REAC | 57 |  |  |  | 0.035 (2) |
| Cytokine Signaling in Immune system | REAC | 309 |  |  |  | 0.005 (4) |
| Interferon Signaling | REAC | 194 |  |  |  | 0.022(3) |
| Interferon gamma signaling | REAC | 91 |  |  |  | 0.002 (3) |
| Interferon alpha/beta signaling | REAC | 68 |  |  |  | 0.05 (2) |

1 Pathways from KEGG and REACTOME databases.

2 P-values are calculated using hypergeometric test with FDR correction for multiple testing, as described in Reimand *et.al.*20.

3 Number of genes from the list found in the pathway (hits) is indicated in brackets. Only pathways with 2 hits or more are shown.

**Legend to Figures:**

**Figure 1: Clustering using PAM algorithm on the training set.**

Panel A represents the consensus Cumulative Distribution Function (CDF) of the consensus matrix shown in panel B. Panel C represents 2 internal quality indexes: the deviation from ideal stability and the Calinski and Harabasz index. Panel D shows a heat map of the pairwise distances between participants.

**Figure 2: Boxplots of main variables**

Boxplots of some variables used for clustering (A-F) together with oral corticosteroid (OCS) use as binary variable (I), blood eosinophil counts (G) and sputum eosinophil percentage (H). Clusters found in the training and validation sets exhibit similar distributions (light and dark hue of the same colour). The V4a and V4b clusters relate to 2 subsets of T4.

**References:**

1. Chung KF, Wenzel SE, Brozek JL, Bush A, Castro M, Sterk PJ, et al. International ERS/ATS guidelines on definition, evaluation and treatment of severe asthma. The European respiratory journal 2014; 43:343-73.

2. Wenzel SE. Complex phenotypes in asthma: current definitions. Pulmonary pharmacology & therapeutics 2013; 26:710-5.

3. Haldar P, Pavord ID, Shaw DE, Berry MA, Thomas M, Brightling CE, et al. Cluster analysis and clinical asthma phenotypes. American journal of respiratory and critical care medicine 2008; 178:218-24.

4. Moore WC, Meyers DA, Wenzel SE, Teague WG, Li H, Li X, et al. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. American journal of respiratory and critical care medicine 2010; 181:315-23.

5. Bateman ED, Reddel HK, van Zyl-Smit RN, Agusti A. The asthma-COPD overlap syndrome: towards a revised taxonomy of chronic airways diseases? The Lancet. Respiratory medicine 2015; 3:719-28.

6. Baines KJ, Simpson JL, Wood LG, Scott RJ, Gibson PG. Transcriptional phenotypes of asthma defined by gene expression profiling of induced sputum samples. The Journal of allergy and clinical immunology 2011; 127:153-60, 60 e1-9.

7. Woodruff PG, Modrek B, Choy DF, Jia G, Abbas AR, Ellwanger A, et al. T-helper type 2-driven inflammation defines major subphenotypes of asthma. American journal of respiratory and critical care medicine 2009; 180:388-95.

8. Yan X, Chu JH, Gomez J, Koenigs M, Holm C, He X, et al. Noninvasive analysis of the sputum transcriptome discriminates clinical phenotypes of asthma. American journal of respiratory and critical care medicine 2015; 191:1116-25.

9. Shaw DE, Sousa AR, Fowler SJ, Fleming LJ, Roberts G, Corfield J, et al. Clinical and inflammatory characteristics of the European U-BIOPRED adult severe asthma cohort. Eur Respir J 2015.

10. Box GEP, Cox DR. An Analysis of Transformations (with discussion). Journal of the Royal Statistical Society Series B-Statistical Methodology 1964; 26:211-52.

11. Fox J, Weisberg S. An R companion to applied regression. 2nd ed. Thousand Oaks, Calif.: SAGE Publications; 2011.

12. Kaufman L, Rousseeuw P. Clustering by Means of Medoids. In: Faculty of Mathematics and Informatics DU, editor. Reports of the Faculty of Mathematics and Informatics. Delft; 1987.

13. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning 2003; 52:91-118.

14. Senbabaoglu Y, Michailidis G, Li JZ. Critical limitations of consensus clustering in class discovery. Sci Rep 2014; 4:6207.

15. Calinski T. A Dendrite Method for Cluster Analysis. Biometrics 1968; 24:207-&.

16. Pavord ID, Pizzichini MMM, Pizzichini E, Hargreave FE. The use of induced sputum to investigate airway inflammation. Thorax 1997; 52:498-501.

17. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003; 4:249-64.

18. Rohloff JC, Gelinas AD, Jarvis TC, Ochsner UA, Schneider DJ, Gold L, et al. Nucleic Acid Ligands With Protein-like Side Chains: Modified Aptamers and Their Use as Diagnostic and Therapeutic Agents. Molecular therapy. Nucleic acids 2014; 3:e201.

19. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007; 8:118-27.

20. Reimand J, Arak T, Vilo J. g:Profiler--a web server for functional interpretation of gene lists (2011 update). Nucleic Acids Res 2011; 39:W307-15.

21. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological 1995; 57:289-300.

22. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000; 28:27-30.

23. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic Acids Res 2014; 42:D472-7.

24. Karatzoglou A, Smola A, Hornik K, Zeileis A. kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software 2004; 11:20.

25. Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A. caret: Classification and Regression Training. 2012.

26. McFadden C, Morgan R, Rahangdale S, Green D, Yamasaki H, Center D, et al. Preferential migration of T regulatory cells induced by IL-16. Journal of immunology 2007; 179:6439-45.

27. Konno S, Taniguchi N, Makita H, Nakamaru Y, Shimizu K, Shijubo N, et al. Distinct Phenotypes of Cigarette Smokers Identified by Cluster Analysis of Patients with Severe Asthma. Annals of the American Thoracic Society 2015; 12:1771-80.

28. Weatherall M, Travers J, Shirtcliffe PM, Marsh SE, Williams MV, Nowitz MR, et al. Distinct clinical phenotypes of airways disease defined by cluster analysis. The European respiratory journal 2009; 34:812-8.

29. Kim TB, Jang AS, Kwon HS, Park JS, Chang YS, Cho SH, et al. Identification of asthma clusters in two independent Korean adult asthma cohorts. The European respiratory journal 2013; 41:1308-14.

30. Bellini A, Yoshimura H, Vittori E, Marini M, Mattoli S. Bronchial epithelial cells of patients with asthma release chemoattractant factors for T lymphocytes. The Journal of allergy and clinical immunology 1993; 92:412-24.

31. Cruikshank WW, Center DM, Nisar N, Wu M, Natke B, Theodore AC, et al. Molecular and functional analysis of a lymphocyte chemoattractant factor: association of biologic function with CD4 expression. Proceedings of the National Academy of Sciences of the United States of America 1994; 91:5109-13.

32. Laberge S, Ernst P, Ghaffar O, Cruikshank WW, Kornfeld H, Center DM, et al. Increased expression of interleukin-16 in bronchial mucosa of subjects with atopic asthma. American journal of respiratory cell and molecular biology 1997; 17:193-202.

33. Vlahos R, Bozinovski S, Chan SP, Ivanov S, Linden A, Hamilton JA, et al. Neutralizing granulocyte/macrophage colony-stimulating factor inhibits cigarette smoke-induced lung inflammation. American journal of respiratory and critical care medicine 2010; 182:34-40.

34. Yee J, Sadar MD, Sin DD, Kuzyk M, Xing L, Kondra J, et al. Connective tissue-activating peptide III: a novel blood biomarker for early lung cancer detection. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2009; 27:2787-92.

35. Di Stefano A, Caramori G, Gnemmi I, Contoli M, Bristot L, Capelli A, et al. Association of increased CCL5 and CXCL7 chemokine expression with neutrophil activation in severe stable COPD. Thorax 2009; 64:968-75.

36. Rovina N, Koutsoukou A, Koulouris NG. Inflammation and immune response in COPD: where do we stand? Mediators of inflammation 2013; 2013:413735.

37. Charles N, Watford WT, Ramos HL, Hellman L, Oettgen HC, Gomez G, et al. Lyn kinase controls basophil GATA-3 transcription factor expression and induction of Th2 cell differentiation. Immunity 2009; 30:533-43.

38. Wang W, Ye Y, Li J, Li X, Zhou X, Tan D, et al. Lyn regulates cytotoxicity in respiratory epithelial cells challenged by cigarette smoke extracts. Current molecular medicine 2014; 14:663-72.

39. Beavitt SJ, Harder KW, Kemp JM, Jones J, Quilici C, Casagranda F, et al. Lyn-deficient mice develop severe, persistent asthma: Lyn is a critical negative regulator of Th2 immunity. Journal of immunology 2005; 175:1867-75.