Novel views on endotyping asthma, its remission, and COPD

Carpaij, Orestes

DOI:
10.33612/diss.136744640

Chapter 11

Bronchial gene expression clustering in COPD identifies a subgroup with higher level of bronchial T-cells and accelerated lung function decline

Orestes A. Carpaij [#], Alen Faiz [#], Jeunard Boekhoudt, Beth Becker, Judith M. Vonk, Wim Timens, Huib A.M. Kerstjens, Gaik W. Tew, Michelle Grimbaldeston, Maggie Neighbors, Lisette I.Z. Kunz, Pieter S. Hiemstra, Katrina A. Steiling, Avi Spira, Victor Guryev, Corry-Anke Brandsma [*], Maarten van den Berge [*]

#: shared first author
*: shared last author

Submitted

Abstract

Chronic Obstructive Pulmonary Disease (COPD) is usually diagnosed, staged, and monitored by clinical features, while it is known that these parameters poorly reflect underlying pathology. More insight to distinguish pathophysiological mechanisms, so-called endotyping, is needed to predict disease progression in COPD and improve treatment. Linking genome-wide gene expression profiling to disease pathology has the potential to contribute to this endotyping.

The aim of the study was to relate gene expression-based clusters of COPD patients to physiological and histopathological parameters in COPD to identify new endotypes.

An existing COPD-associated gene signature was applied on our bronchial biopsy RNA-sequencing dataset derived from mild-moderate COPD patients, to perform an unsupervised clustering analysis using ConsensusClusterPlus. The gene expression-based clusters of COPD patients were related to cross-sectional clinical and histopathological features, as well as longitudinal lung function decline over 7 years.

We identified two clusters of COPD patients: COPD-associated Airway Gene Expressed 1 (CAGE1)-cluster (n = 39) and CAGE2-cluster (n = 17). CAGE2 was characterized by higher baseline sputum lymphocyte percentage, higher biopsy CD4+ and CD8+ T-cell counts, more rapid lung function decline over 7 years and inhaled corticosteroid unresponsiveness compared to CAGE1. The CAGE2 gene signature was associated with more severe COPD in the validation cohort.

Unsupervised clustering analysis based on an existing COPD gene expression signature enabled the identification of new COPD endotypes with significant differences in bronchial T-cell count and acceleration of lung function decline.

Introduction

Chronic Obstructive Pulmonary Disease (COPD) is primarily diagnosed, staged, and monitored by pulmonary function tests and symptom severity [1], even though these parameters poorly reflect the underlying pathology [2]. Several COPD studies have focussed on identification of new biomarkers, such as blood neutrophils, $CD8^+$ T-cells, IL-6 and genetics [3,4]. However, with the exception of α-1-antitrypsin deficiency and the newly identified rare dominant mutation in protein tyrosine phosphatase non-receptor type 6 (PTPN6) [5], most biomarkers have fallen short of predicting long-term outcome in COPD. More insight in the pathologic changes and identification of so-called endotypes of COPD, i.e. subgroups with distinct pathologic features, is highly needed.

Combining genome-wide gene expression profiling with clinical outcomes and pathological features represents a promising contribution for endotyping of obstructive pulmonary diseases [6]. Several unsupervised gene expression cluster analyses have been performed on sputum samples [7,8] and bronchial biopsies [9,10]. In asthmatics, Baines *et al.* distinguished three different clusters based on gene expression differences in asthma linked to distinct clinical features [8]. Additionally, their sputum-derived gene signature could predict the clinical response to inhaled corticosteroids (ICS). Unsupervised gene expression clustering in COPD patients has been studied less; Chang *et al.* performed unsupervised clustering in 141 COPD patients and 88 smoker controls using blood microarray gene expression data and identified four distinct endotypes [11]. One of the endotypes was associated with severe lung function impairment, respiratory symptoms and CT-scan characterised emphysema, whereas another endotype was characterized by a more preserved lung function and less emphysema. These studies suggest that combining genome-wide gene expression profiling with clinical outcomes and pathological features enables us to identify key COPD endotypes that may help in determining prognosis, treatment response and in the end novel treatments.

Previously we have shown that a 98-gene signature from airway epithelial brushes can distinguish COPD patients from healthy controls [12]. This gene expression-signature was found to be enriched for genes involved in a variety of pathogenic categories, e.g. glycoproteins, proteins involved in acute inflammation (both up-regulated) and

epidermal growth factor-like domains (down-regulated), suggesting an association with inflammatory response and altered epithelial restoration [12].

In the current study, we performed unsupervised clustering using this COPD-associated gene-signature in RNA-seq data obtained from bronchial biopsies of well-characterized COPD patients from the Groningen and Leiden Universities Study of Corticosteroids in Obstructive Lung Disease (GLUCOLD) study. We related these clusters to physiological and histopathological parameters to identify clinically relevant COPD endotypes.

Methods

Patient inclusion

We included subjects who participated in the GLUCOLD study [13]. Briefly, patients were 45 – 75 years old, had an FEV_1/FVC ratio <70%, ≥10 packyears of smoking and no history of asthma. COPD patients were characterised with the following tests, i.e. spirometry, peripheral blood samples (for cell differential), sputum induction (for inflammatory cell percentages) and bronchoscopy with bronchial biopsies, which were immunostained for inflammatory cells (i.e. $CD3^+$, $CD4^+$, CD8+, $CD68^+$, neutrophil elastase, tryptase, $EG2^+$). GLUCOLD was a randomized, double-blind, placebo-controlled study that consisted of four treatment regimens for 30 months, i.e. fluticasone, fluticasone/salmeterol, placebo, or fluticasone the first 6 months followed by placebo for the remaining 24 months of the study. After the end of this 30-month double-blind treatment period, spirometry was repeated once yearly for an additional 5 years (7.5 years in total) [14]. After the 30 month trial, ICS treatment was considered ongoing when a subject used ICS for >50% of the total follow-up period [15]. The local ethics committee approved the study protocol and all subjects gave written informed consent.

Cluster algorithm and statistical analysis

Methods for RNA isolation from bronchial biopsies and RNA-seq were described previously [16]. Processing of RNA libraries and RNA sequencing methods were outlined previously as well [17]. The 98 gene-signature associated with COPD was selected from a previous study [12]. Of the 98 genes in this signature, 93 genes were detected in our baseline RNA-seq dataset. Genes with an expression value of less than one fragment per million reads in all samples were excluded.

The R package ConsensusClusterPlus was used to identify clusters based on the 93 gene-signature list and its algorithm is described by Wilkerson *et al.* [18]. In brief, using this algorithm, a consensus value was calculated per number of clusters, resulting in Cumulative Distribution Functions (CDF). The number of clusters with the lowest CDF value (i.e. the least gene expressional overlap between the clusters) was selected for further analysis. Subsequently, baseline and longitudinal clinical and histopathological features were compared between the clusters, using SPSS 23.0.03 software (SPSS Inc., Chicago, IL, USA). When data was non-normally distributed, a log2 transformation was performed. If these variables remained non-normally distributed, we used the original values. Differences in baseline and longitudinal clinical features between clusters were compared using independent sample T-tests in case of normal distribution, Mann-Whitney U test for non-normally distributed data and chi-square tests for categorical variables. Subsequently, a logistic regression was used to correct for smoking status, age, and sex.

A linear mixed effects model was used to pool treatment arms and analyse annual FEV_1 decline (in millilitres per year). In order to use all subjects for a linear mixed effects model and expel treatment changes during the study as much as possible, the six months' time point was used as baseline, as described previously [15]. This way, a linear mixed effects model was applied between six months and 7.5 years, correcting for smoking status, ICS treatment (i.e. >50% ICS use between 30 months – 7.5 years) and inflammatory cell count as possible confounders.

Differential expression and pathway analysis

To identify genes differentially expressed between the identified clusters whole genome differential gene expression analysis was performed using a linear model correcting for smoking status, gender, and age (R-package Limma version 3.40.6), using a False Discovery Rate (FDR) of <0.05. G-Profiler and Gene Set Enrichment Analysis (GSEA) were used for pathway analysis of genes associated with the different clusters, as described previously [19]. For GSEA, genes were ranked based on their T-value statistic, comparing the number of clusters.

Validation of clusters associated with severity of COPD

To determine if the identified clusters could be replicated, we examined the signature of the differentially expressed genes in a dataset previously published by Steiling *et al.*, consisting of bronchial brushings from ever-smokers with and without COPD enrolled in the British Columbia Lung Health Study (BCLHS), using gene set variation analysis (GSVA) between the clusters in a subset of this dataset [12]. We used gene symbols to match the genes identified by sequencing and affymatrix micro-array. Separate GSVA analyses were performed for the genes that were up- or down-regulated in the comparison between CAGE1 and CAGE2 (t < or > 0), resulting in separate GSVA scores. These two GSVA scores were then projected into the validation dataset and linear models were used to test whether the GSVA scores were associated with baseline and longitudinal lung function tests, correcting for sex, age, packyears, and smoking status.

Results

Of the original 114 randomised COPD patients in the GLUCOLD study at baseline, 89 frozen biopsies were available, and 56 had RNA of sufficient quality for RNA sequencing. All subjects had longitudinal pulmonary function measurements until 30 months of follow-up and 31 individuals had measurements until 7.5 years. The mean duration of the follow-up was 5.7 ±2.6 years.

Based on the existing COPD gene signature, CDF clustering analysis showed that a model with two clusters resulted in the lowest inter-consensus value and therefore these two clusters were selected for further analysis (figure 1). One cluster, termed the COPD-associated Airway Gene Expressed 1 (CAGE1), consisted of 39 COPD patients. The second cluster, termed the COPD-associated Airway Gene Expressed 2 (CAGE2) cluster, was comprised of 17 COPD patients.

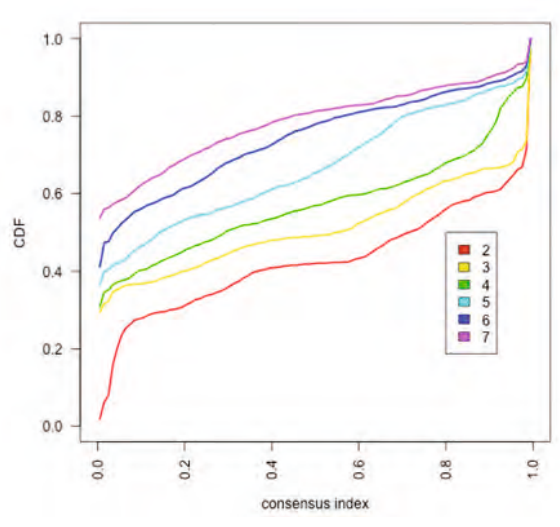


Figure 1: cluster quantity and least inter-cluster consensus. The red line represents the cluster quantity (red is two clusters, yellow is three, green is four, etc.) with the least inter-cluster consensus, which is based on Cumulative Distribution Functions (CDF on y-axis).

CAGE2 is associated with higher lymphocyte levels

Baseline demographics, clinical and histopathological characteristics for CAGE1 and CAGE2 are presented in table 1. There was no significant difference in gender, age, smoking status, pack years and baseline pulmonary function tests between the CAGE1

and CAGE2 clusters. Sputum lymphocyte percentage was higher in CAGE2 compared to CAGE1 (2.2% (IQR 1.8 – 3.5) vs. 1.8% (IQR 1.2 – 2.2), P=0.045, figure 2A), as well as $CD3^+$, $CD4^+$ and $CD8^+$ T-cell counts in biopsies ($CD3^+$: 153.8/0.1 mm^2 (IQR 107.8 – 211.3) vs. 103.4/0.1 mm^2 (IQR 64.0 – 169.5), P=0.01, $CD4^+$: 68/0.1 mm^2 (IQR 57 – 97) vs. 44/0.1 mm^2 (IQR 25 – 76), P=0.02, $CD8^+$: 30/0.1 mm^2 (IQR 20 – 37) vs. 17 /0.1 mm^2 (IQR 8 – 28), P=0.006, figure 2B-D). After correcting for either smoking status or age, both bronchial $CD3^+$ and $CD8^+$ T-cell count remained significantly associated with CAGE2 (P=0.035 and P=0.044, respectively), but significance was lost after correcting for both. Bronchial $CD4^+$ T-cell count was not significantly different between CAGE1 and CAGE2 when correcting for smoking or age.

Table 1: baseline characteristics of COPD patients in CAGE1 and CAGE2

	CAGE1 n = 39	CAGE2 n = 17	p	p^S	p^A	p^{SSA}
Sex, male (n, %)	35 (89.7%)	15 (88.2%)	1.0 #	-	-	-
Mean age in years	61.5 (±7.1)	60.3 (±9.1)	.59 ¶	-	-	-
Mean BMI in kg/m^2	25.2 (±3.1)	25.5 (±4.6)	.83 ¶	-	-	-
Current smoking (n, %)	29 (74.4%)	9 (52.9%)	.13 #	-	-	-
Median pack years	42.0 (31.9 – 55.6)	36.5 (31.5 – 51.9)	.45 §	-	-	-
Reversibility (≥12% and >200ml improvement)	8 (20.5%)	4 (23.5%)	1.0 #	-	-	-
GOLD II to III Classification ratio (n, %)	36 (92.3%)	15 (88.2%)	.63 #	-	-	-
Mean FEV_1 % predicted Post-BD	61.5 (±8.7)	64.9 (±9.0)	.46 ¶	-	-	-
Mean FEV_1/IVC ratio (%)	45.3 (±9.0)	47.9 (±8.9)	.32 ¶	-	-	-
Mean RV/TLC ratio (%)	48.8 (±8.3)	46.9 (±6.6)	.42 ¶	-	-	-
Mean TLCO % predicted	63.9 (±22.6)	63.9 (±14.5)	1.0 ¶	-	-	-
PC_{20} methacholine mg/ml $	0.7 (0.2 – 2.8)	0.5 (0.1 – 1.6)	.46 ¶	-	-	-
PC_{20} methacholine ≤8 mg/ml (n, %)	36 (94.7%)	16 (94.1%)	1.0 #	-	-	-
Median CCQ, total score	1.2 (0.8 – 1.8)	1.4 (0.9 – 1.8)	.75 §	-	-	-
Mean SGRQ, symptom score	31 (±14)	29 (±15)	.68 ¶	-	-	-
Treatment arm: placebo (n, %)	8 (20.5%)	6 (35.3%)				
Treatment arm: fluticasone/salmeterol (n, %)	10 (25.6%)	6 (35.3%)				
Treatment arm: fluticasone ≥0 months (n, %)	9 (23.1%)	4 (23.5%)	.21#	-	-	-
Treatment arm: fluticasone ≥6 months (n, %)	12 (30.8%)	1 (5.9%)				
Treatment arm: placebo (n, %)	8 (20.5%)	6 (35.3%)				

Table 1: (continued)

	CAGE1 n = 39	CAGE2 n = 17	p	p^{S}	p^{A}	p^{SSA}
ICS use until 30 months (n, %)	31 (79.5%)	11 (64.7%)	.32 #	-	-	-
>50% ICS use of the period 2.5-7.5 years (n, %)	9 (29.0%)	2 (15.4%)	.46 #	-	-	-
Mean % eosinophils $	1.9 (1.3 – 3.2)	2.4 (1.2 – 4.4)	.24 ¶	-	-	-
Mean % basophils $	0.5 (0.3 – 0.7)	0.4 (0.3 – 1.1)	.49 ¶	-	-	-
Mean % neutrophils	60.2 (±9.3)	57.7 (±10.1)	.38 ¶	-	-	-
Median % monocytes	8.8 (7.5 – 9.5)	8.3 (7.4 – 10.6)	.99 §	-	-	-
Mean % lymphocytes	28.3 (±7.6)	29.8 (±10.6)	.55 ¶	-	-	-
Mean % eosinophils $	1.1 (0.3 – 2.0)	0.9 (0.3 – 2.5)	.96 ¶	-	-	-
Median % neutrophils	70.0 (59.7 – 81.5)	72.0 (64.0 – 75.0)	.98 §	-	-	-
Median % macrophages	25.2 (15.0 – 34.5)	21.7 (18.5 – 25.7)	.68 §	-	-	-
Median % lymphocytes	1.8 (1.2 – 2.2)	2.2 (1.8 – 3.5)	**.045** §	.176	**.027**	.086
CD3^{+} T-cells, Count / 0.1mm^{2} $	103.4 (64.0 – 169.5)	153.8 (107.8 – 211.3)	**.013** ¶	**.037**	**.022**	.051
CD4^{+} T-cells, Count / 0.1mm^{2} $	44.2 (25.0 – 75.5)	67.4 (56.5 – 96.5)	**.022** ¶	.094	.063	.144
CD8^{+} T-cells, Count / 0.1mm^{2} $	17.0 (8.0 – 28.0)	29.7 (20.3 – 36.5)	**.006** ¶	**.042**	**.040**	.071
NE^{+} neutrophils, Count / 0.1mm^{2} $	3.9 (2.5 – 8.0)	4.6 (3.5 – 6.5)	.55 ¶	-	-	
Tryptase^{+} mast cells, Count / 0.1mm^{2} $	27.2 (20.5 – 37.0)	26.1 (23.0 – 32.8)	.78 ¶	-	-	
CD68^{+} macrophages, Count / 0.1mm^{2} $	8.4 (5.0 – 13.5)	9.8 (9.0 – 13.5)	.49 ¶	-	-	
EG2^{+} eosinophils, Count / 0.1mm^{2}	0.4 (0.3 – 0.7)	0.2 (0.5 – 0.8)	.94 §	-	-	

BMI: body mass index, ICS: inhaled corticosteroids, BD: bronchodilator (salbutamol 400mcg), PC_{20}: substance dosage needed for a 20% FEV_1 drop, CCQ: COPD Control Questionnaire, SGRQ: St. George Respiratory Questionnaire, NE: neutrophil elastase, # : Fisher's Exact Test, ¶ : Independent T-test, § : Mann-Whitney U Test. $: log2 transformed and presented in geometric mean (IQR). S: logistic regression corrected for smoking status, A: logistic regression corrected for age, SSA: logistic regression corrected for sex, smoking status, and age.

CAGE2 is associated with faster lung function decline and ICS unresponsiveness

To investigate the association between CAGE status and the longitudinal decline of lung function, we analysed the change in FEV_1 over 7 years in both groups. Between 6 months and 7.5 year follow-up, patients in the CAGE2 cluster showed a faster decline of FEV_1 compared to patients in CAGE1 (-69.9 ml/year (95% CI -55.7 to -84.0 ml) versus -44.0 ml/year (95% CI -30.3 to -57.7 ml/year), P=0.002), after correcting for smoking status, ICS treatment and biopsy CD4^{+} and CD8^{+} T-cell counts (figure 2E). In addition, CAGE2-patients responded less after 30 months of ICS treatment compared to CAGE1-patients as reflected by the change in $_{11}$ (-29.1ml/year (95% CI -89.2 – 30.9ml) versus 24.4ml/year (95% CI -30.9 – 79.8ml) (P=0.048, corrected for smoking status and, biopsy CD8^{+} T-cell count; P=0.146 when additionally corrected for CD4^{+} T-cell count, figure 2F). In addition, we investigated whether patients with high or low CD4^{+} or CD8^{+} T-cell counts at baseline numbers (i.e. high >50th percentile versus low <50th percentile) showed differences in FEV_1 decline, but found no significant difference (P=0.94 and P=0.92 respectively), corrected for smoking status, supplementary figure 1A). No correlations were found between baseline FEV_1 % predicted with either biopsy CD4^{+} or CD8^{+} T-cell count (Pearson's R=-0.036, P=0.79 and R=0.13, P=0.33 respectively). Finally, since the CD4^{+} and CD8^{+} T-cells were different between clusters, which could be a possible confounder for findings related to lung function decline, we also analysed their direct association, which was not significant (CD4^{+} Spearman's R=0.139, P=0.34 and CD8^{+} R=0.062, P=0.67).

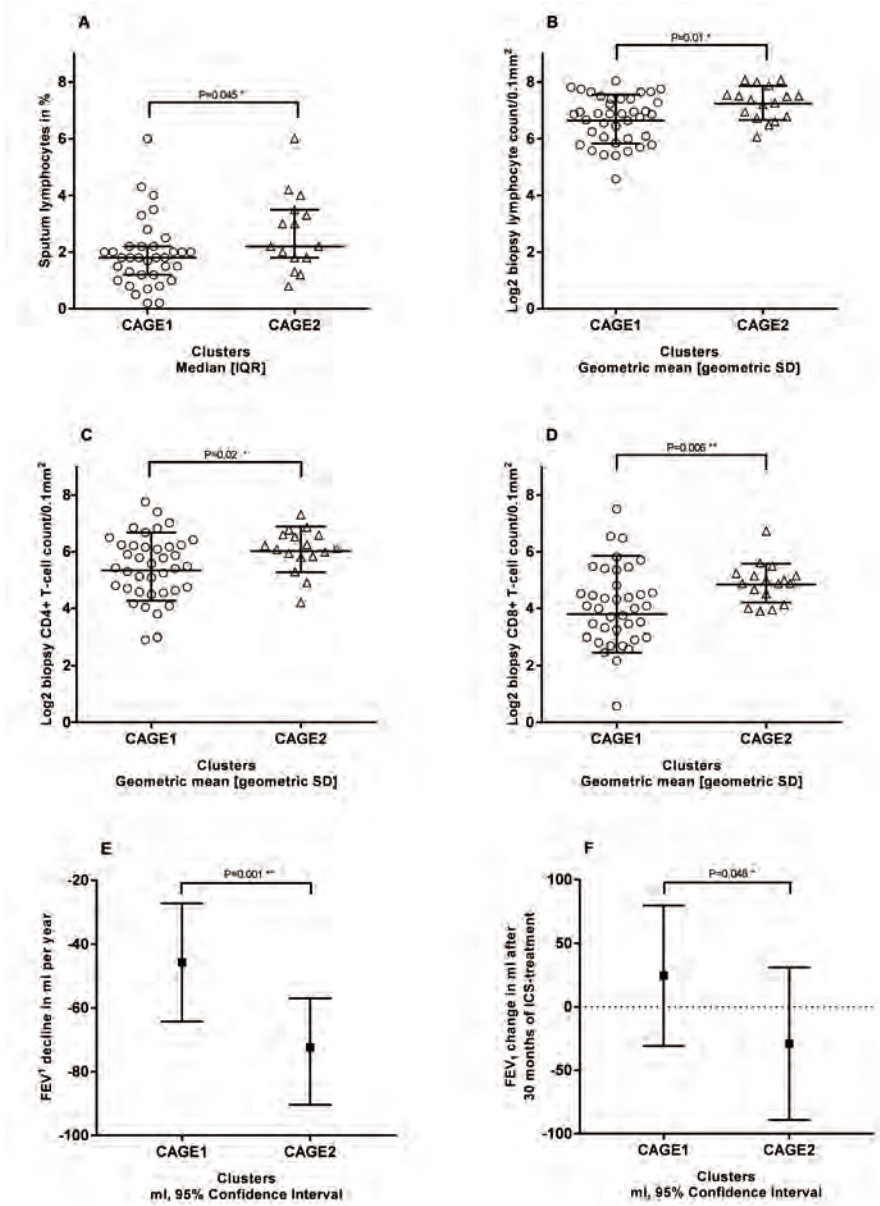


Figure 2: inflammatory level parameters in CAGE1 and CAGE2. **2A:** sputum lymphocyte percentage of total amount of cells per cluster, presented in median with IQR, **2B:** log2-transformed biopsy $CD3^+$ T-cell count in 0.1mm² per cluster, presented in geometric mean and SD, **2C:** log2-transformed baseline $CD4^+$ T-cell count count in 0.1mm² per cluster, presented in geometric mean and SD, **2D:** log2-transformed baseline biopsy $CD8^+$ T-cell count in 0.1mm² per cluster, presented in geometric mean and SD, **2E:** annualized FEV_1 change (ml per year, 95% CI) between 6 months and 7.5 years of follow-up per cluster, corrected for smoking status, ICS treatment and log2-transformed $CD4^+$ and $CD8^+$ T-cell biopsy count variance from geometric mean, **2F:** FEV_1 change in ml (and 95% CI) between 0 and 30 months of ICS treatment per cluster, corrected for smoking status and log2-transformed $CD8^+$ T-cell biopsy count variance from geometric mean.

Gene expression differences between CAGE1 & CAGE2 and pathway analyses

Given that the CAGE1 and CAGE2 patients clusters differed in COPD-related phenotypic variables such as T-cell counts, and lung function decline, we next examined differences in genome-wide gene expression profiles between the 17 patients in CAGE1 and 39 patients in CAGE2. We identified 200 genes that were differentially expressed between the two groups with fold change FC>|2| and FDR<0.05, with 186 higher expressed in CAGE2 as compared to CAGE1, and 14 genes whose expression levels were lower in CAGE2 as compared to CAGE1 (table 2). Pathway analysis using G-profiler on the 186 higher expressed genes showed an enrichment in pro-inflammatory pathways (e.g. antigen binding, humoral immune response mediated by circulating immunoglobin and B-cell mediated immunity, table 3). No pathways were associated with the 14 lower expressed genes.

Table 2: top 20 genes differentially expressed in CAGE2 compared to CAGE1

	Increased	FDR	Decreased	FDR
1.	NXPE2	2.05E-24	SCEL	2.57E-05
2.	LYZ	7.58E-21	CLCA4	8.80E-05
3.	LTF	8.37E-18	CAPN14	2.51E-04
4.	AZGP1	8.81E-22	TGM3	3.24E-04
5.	DMBT1	5.26E-16	NMRAL2P	1.07E-04
6.	ZG16B	5.07E-19	SERPINB2	4.12E-04
7.	MGAM2	8.00E-17	A2ML1	4.24E-04
8.	PPP1R1B	1.19E-18	MUC21	3.88E-04
9.	CRISP3	2.91E-15	HS6ST2	1.46E-06
10.	CA2	5.75E-18	CYP3A5	4.59E-04
11.	PRR4	1.09E-17	SLC7A11	3.27E-04
12.	PIP	7.90E-15	FGFBP1	1.02E-03
13.	CCDC129	2.68E-15	KLK13	1.46E-03
14.	GP2	3.20E-14	LYPD3	1.44E-03
15.	TTYH1	1.27E-14	-	-
16.	LINC02009	1.09E-13	-	-
17.	C6orf58	3.26E-13	-	-
18.	MYCN	8.00E-17	-	-
19.	LPO	9.81E-13	-	-
20.	NPY1R	4.08E-15	-	-

Table 3: clusters gene-expression pathways (G-profiler)

Rank	Term name	Term ID	Term genes (N)	Query genes (N)	Common genes (N)	Corrected P
1.	Antigen binding	Go:0003823	189	164	19	6.07×10^{-12}
2.	Complement activation, classical pathway	Go:0006958	133	164	15	1.08×10^{-9}
3.	Immune response-activating signal transduction	Go:0002757	576	164	27	1.10×10^{-9}
4.	Activation of immune response	Go:0002253	638	164	28	$1,91 \times 10^{-9}$
5.	Immune response-regulating signalling pathway	Go:0002764	605	164	27	3.48×10^{-9}
6.	Humoral immune response mediated by circulating immunoglobin	Go:0002455	144	164	15	3.50×10^{-9}
7.	Immunoglobin mediated immune response	Go:0016064	205	164	17	4.51×10^{-9}
8.	B cell mediated immunity	Go:0019724	207	164	17	5.28×10^{-9}
9.	Positive regulation of immune response	Go:0050778	775	164	30	6.41×10^{-9}
10.	Adaptive immune response	Go:0002250	488	164	24	9.10×10^{-9}

To provide complementary functional assessment of the gene expression signature, GSEA was used to compare the cluster's pathways. Figure 3A illustrates a pathway analysis on the higher expressed genes, including antigen processing and presentation (e.g. antigen processing and presentation, antigen presentation folding assembly and peptide loading of class I MHC) and T-cell pathways (e.g. immunoregulatory interactions between lymphoid and non-lymphoid cells, TCR signalling, phosphorylation of $CD3^{+}$ and TCR ζ chains, interferon γ signalling). Figure 3B presents a pathway analysis of the lower expressed genes, e.g. cell senescence (meiosis, packaging of telomere ends and telomere maintenance) and mitochondrial pathways (citric acid and respiratory electron transport, mitochondrial transcription).

Figure 3: GSEA pathway analysis of increased and decreased genes. **3A:** GSEA pathway analysis of increased and decreased genes, **3B:** GSEA pathway analysis of increased and decreased genes.

Replication of CAGE2 using an independent cohort

Analysis of the BCLHS data was used for replication of CAGE2 in an independent cohort [12]. Baseline demographics of this study are presented in supplementary table 1. In brief, the validation cohort consisted of 87 COPD patients. Thirty subjects were current-, and 57 were former-smokers. The mean age was 65 ±6 years and the mean amount of packyears was 51 ±25. The mean baseline FEV_1 % predicted was 60.3% ±13.8 and follow-up period was four years.

Of the 186 higher and 14 lower expressed genes between the clusters, 143 were expressed in the validation cohort. Of these 143 genes, 140 were higher expressed and 13 were lower expressed in CAGE2 compared to CAGE1. In the COPD patients, we found that a higher expression of CAGE2 associated genes were associated with more severe COPD, i.e. a trend for a lower baseline FEV_1 % predicted (T=-1.70, P=0.09) and lower FEV_1/FVC ratio (T=-2.74, P=7.44 x 10^{-3}) (table 4). GSVA for genes decreased with CAGE2 were not associated with severity of COPD. In addition, GSVA for genes increased and decreased with CAGE2 were not associated with lung function decline in COPD.

Table 4: validation of CAGE2 signature in British Columbia Lung Health Study

	T GSVA of up-regulated genes in CAGE2 vs. CAGE1	P-value	T GSVA of down-regulated genes in CAGE2 vs. CAGE1	P-value
COPD only				
ΔFEV_1 COPD only (ml per year)	-0.35	0.73	0.15	0.88
Baseline FEV_1 % predicted (%)	-1.70	0.09	-1.42	0.16
FEV_1/FVC ratio (%)	**-2.74**	***7.44 x 10^{-3}***	-0.04	0.97

Corrected for sex, age, pack years, and smoking status.

Discussion

In the current study, we identified two COPD subgroups based on bronchial biopsy gene expression data, i.e. CAGE1 and CAGE2. The CAGE2 endotype is characterized by a higher sputum lymphocyte percentage, and higher biopsy $CD4^+$ T-cell and $CD8^+$ T-cell counts, but most importantly, associated with more rapid lung function decline, independent of smoking, ICS treatment and baseline CD8+ T-cell counts when compared to CAGE1. Pathway analysis confirmed that genes who's expression is associated within CAGE2 are involved in T-cell immune responses supporting the known relationship between chronic inflammation and lung function decline in COPD patients; these signatures now allow the discrimination of slow (CAGE1) and rapid (CAGE2) decline in lung function.

We found CAGE2 to be associated with higher $CD4^+$ and $CD8^+$ T-cell numbers in the airway wall compared to CAGE1. Several studies found that bronchial $CD8^+$ T-cells are associated with lung function impairment in COPD [20–22]. Using surgically resected lung tissue, Hogg et al. showed that the percentage of airways containing $CD8^+$ T-cells was higher in the higher COPD stages [22]. $CD8^+$ T-cells play a dominant role in airway inflammation [23–25] and have been associated with increased epithelial apoptosis as is often observed in COPD [26]. Whereas CAGE2 when compared to CAGE1 shows differences in T-cell profile, it is important to note that the association between CAGE2 COPD and more rapid loss of lung function remained significant after adjusting for baseline bronchial $CD8^+$ and $CD4^+$ T-cell counts. Therefore, other pathways defined by the CAGE2 signature are likely to be involved in decline of FEV_1. The latter is supported by the finding that bronchial $CD8^+$ and $CD4^+$ T-cells themselves were not associated with either FEV_1 impairment or decline between 6 months and 7.5 years follow-up.

The CAGE2 endotype may be useful as a biomarker for the following reasons. First, it is associated with greater lung function decline, and could potentially be used to identify COPD patients who are at risk for rapid lung function deterioration. Second, we found that CAGE2 patients show less improvement in FEV_1 after ICS treatment than patients with CAGE1 COPD. Thus, these clusters potentially distinguish between patients who are responsive and unresponsive to ICS therapy, and perhaps other (future) treatments. So far, a limited number of studies have been performed that applied unsupervised

clustering on COPD patients based on gene expression signatures. Chang et al. analysed genome-wide blood gene expression from 229 former smokers in the ECLIPSE study [11], and identified four distinct clinical subtypes of COPD, which were successfully reproduced in an independent sample. The four groups were well differentiated by baseline FEV_1/FVC and FEV_1, and had significant differences in emphysema and symptom severity. Blood sampling would be easier in terms of clinical application and be more patient friendly compared to taking bronchial biopsies. Nevertheless, our current analyses of parameters tested indicate that the baseline identified CAGE2 gene expression cluster is linked to prospective lung function decline rather than associated with existing lung damage, thus yielding more prognostic value. For a biomarker that can predict future risk of rapid lung function loss, a more invasive diagnostic method such as bronchoscopy might be justified as this also is a better representation of the activity of all the cells involved in the disease process, whereas analysis of white blood cells is limited to (off-site) inflammatory cells only.

Amongst the 20 most significantly higher and lower expressed genes in CAGE2 compared to CAGE1, the following were previously found to be associated with obstructive pulmonary disease: lactotransferrin (LTF) and Chromosome 6 Open Reading Frame 58 (C6orf58) were higher, while Cytochrome P450 Family 3 Subfamily A Member 5 (CYP3A5) and Serpin Family B Member 2 (SERPINB2) were lower expressed.

The LTF gene encodes lactotransferrin (also called lactoferrin), a globular glycoprotein which is widely present in secretory body fluids, has both direct and indirect antimicrobial effects [27], and is found to be elevated in non-typeable Haemophilus influenzae and seasonal influenza A virus infections [28]. In virus-exposed human bronchial epithelial cells, pre-treated with concentrations of budesonide, levels of LTF expression were significantly higher [28]. $CD8^+$ T-cell subsets in influenza A infected mice express LTF as well [29].

The function of the C6orf58 gene, apart from being involved in liver development in zebra fish, is unknown [30,31]. Nevertheless, C6orf58 was identified as a differentially expressed protein in sputum supernatant of COPD patients compared to asymptomatic smokers [31].

The CYP3A5 protein is a well-known member of the cytochrome P450 superfamily of enzymes, which metabolize (inhaled) toxicants, such as tobacco smoke, resulting activation or inactivation [32,33]. Therefore, altered expression of CYP3A5 may contribute to the risk of developing lung diseases [32]. While a CYP3A5 gene polymorphism was associated with a faster FEV_1 and FVC decline in current smokers [33], Hukkanen et al. found a decreased CYP3A5 expression level in alveolar macrophages from current-smoking patients with respiratory diseases [32]. Our data suggest that lower expression of the CYP3A5 gene in CAGE2 might be linked to a more rapid lung function decline in COPD.

The SERPINB2 gene enables transcription of the plasminogen activator inhibitor-2, a coagulation factor that is present in most cells, especially in monocytes and macrophages. A negative correlation was found between SERPINB2 expression by PCR in respiratory epithelial cells and FEV_1/FVC ratio, FEV_1 % predicted, and disease severity in asthmatic adults [34]. The high expression of this gene is also included in a signature that enabled identification of the TH2-high, ICS responsive asthma endotype [35]. In line with this, CAGE2 subjects, who have a lower expression of the SERPINB2 gene compared to CAGE1, were unresponsive to ICS treatment.

GSEA pathway analysis of the genes down-regulated in CAGE2 showed enrichment for genes associated with packaging of telomere ends and telomere maintenance, including a number of histone related genes including many members of the HIST1H2 family. There is evidence that COPD is a disease of accelerated lung aging [36]; the progression of the disease and lung function decline is associated to the failure of the lung to repair DNA damage by oxidative stress and from telomere shortening, caused by tobacco smoke [37]. In agreement with this, factors linked to telomere shortening may play a role in the accelerated lung function decline in the CAGE2 subjects.

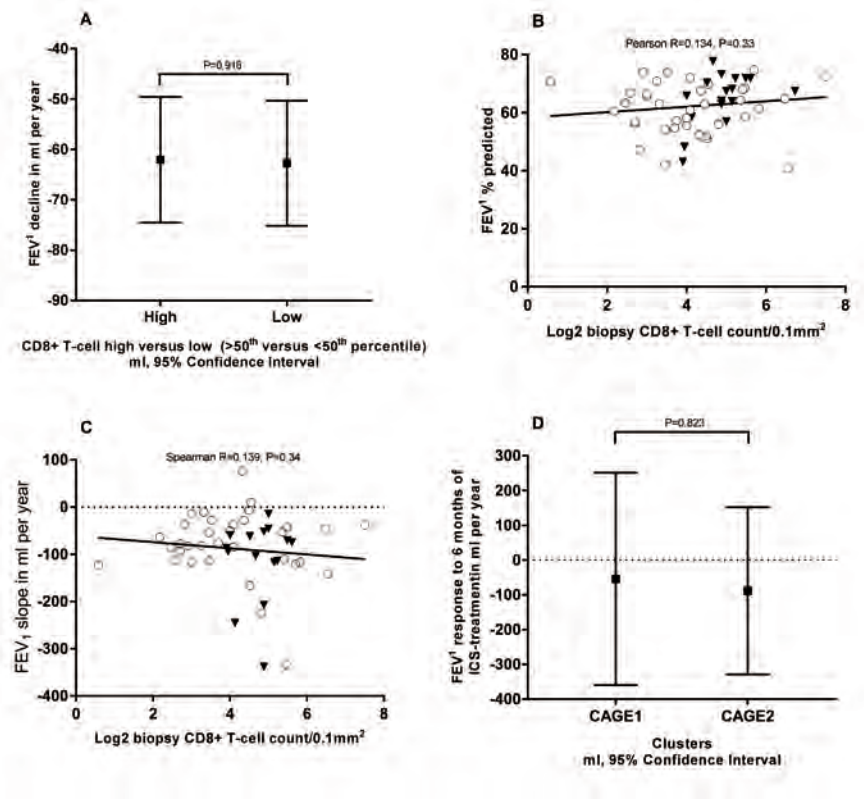
In the validation cohort, we saw a trend of more severe COPD in CAGE2 GSVA compared to CAGE1. However, GSVA for genes increased and decreased with CAGE2 were not associated with lung function decline in COPD. The latter may be due the fact that we used a bronchial biopsy gene signature including a wider variety of (inflammatory) cells and replicated this on a gene set that contains predominantly epithelial cells. Another possible explanation could be that COPD patients with less severe obstruction have

more room to deteriorate and experience more rapid decline expressed in ml per year, as described in previous studies [39,40].

This is the first study to apply an unsupervised clustering method on gene expression data derived from bronchial biopsies in COPD patients. The validation was limited as it is difficult to replicate our findings as there is a lack of studies investigating bronchial gene expression profiles associated with severity of COPD, let alone analysing longitudinal outcomes. We investigated a composite score of genes up- and down-regulated in our CAGE2 signature in the British Columbia Lung Health Study, which were derived from epithelial brushings. Not all genes were present on the array of expressed genes that was used for the validation study, which could be due to the different cell populations represented in bronchial brushings versus biopsies.

Conclusion

In conclusion, our gene expression profile based clustering approach we identified a new COPD endotype, i.e. CAGE2, that is characterized by a higher bronchial $CD8^+$ and $CD4^+$ T-cell level, and most notably, a more rapid lung function decline, compared to the CAGE1 endotype. The identity of genes linked to the expression signatures of the CAGE2 and CAGE1 endotypes confirm underlying immune responses driving COPD, but as such do not explain the difference in lung function decline. These results show that unsupervised clustering analysis based on an existing COPD gene expression signature enables the identification of a new COPD endotype, which in turn could be relevant for diagnosis, staging and treatment of this complex and heterogeneous disease.



Supplementary figure 1: S1A: effect of $CD8^+$ T-cell high/low level (i.e. cut-off at median) on FEV_1 decline per year (95% CI) between 6 months and 7.5 years of follow-up, corrected for smoking status, **S1B:** correlation between baseline FEV_1 % predicted and baseline log2-transformed biopsy $CD8^+$ T-cell count (white circle is CAGE1 and black triangle is CAGE2), **S1C:** correlation between individual FEV_1 ml slope per year between 6 and 7.5 years and baseline log2-transformed biopsy $CD8^+$ T-cell count (white circle is CAGE1 and black triangle is CAGE2), **S1D:** FEV_1 ml change per year (95% CI) between 0 and 6 months of ICS treatment per cluster, corrected for smoking status and log2-transformed $CD8^+$ T-cell biopsy count variance from geometric mean.

Supplementary table 1: baseline demographics of the validation cohort

	BCLHS	GLUCOLD
N	87	56
Mean age	65 ±6	61 ±8
Mean pack years	51 ±25	45 ±18
Current/former smoking ratio (n, %)	30 (34.5%)	38 (67.9%)
Male ratio (n, %)	35 (40.2%)	50 (89.3%)
Inhaled medications (n, %)	23 (26.4%)	Placebo: 14 (25.0%) Fluticasone/salmeterol: 16 (28.6%) Fluticasone 6 months: 13 (23.2%) Fluticasone 30 months: 13 (23.2%)
Baseline FEV_1 % predicted	60.3% ±13.8	62.5% ±8.9
ΔFEV_1 (ml/year)	-40.0 ±50.0	-66.1 [-59.6 - -72.7]
Follow up time (years)	4	5.7 ±2.6

References

1. Global Initiative for Chronic Obstructive-Lung Disease: Global Strategy for the Diagnossis, Management, and Prevention of Chronic Obstructive Pulmonary Disease (2018 Report). 2018.
2. O'Donnell R, Breen D, Wilson S, et al. Inflammatory cells in the airways in COPD. Thorax 2006;61:448–54.
3. Mendy A, Forno E, Niyonsenga T, et al. Blood biomarkers as predictors of long-term mortality in COPD. Clin Respir J Published Online First: 5 January 2018.
4. Finch D, Lange P, Halpin D, et al. Diagnosis, assessment, and phenotyping of COPD: beyond FEV1. Int J Chron Obstruct Pulmon Dis 2016;11 Spec Iss:3.
5. Bossé Y, Lamontagne M, Gaudreault N, et al. Early-onset emphysema in a large French-Canadian family: a genetic investigation. Lancet Respir Med 2019;7:427–36.
6. Agustí A, Celli B, Faner R. What does endotyping mean for treatment in chronic obstructive pulmonary disease? Lancet 2017;390:980–7.
7. Baines KJ, Simpson JL, Wood LG, et al. Transcriptional phenotypes of asthma defined by gene expression profiling of induced sputum samples. J Allergy Clin Immunol 2011;127:153-160.e9.
8. Baines KJ, Simpson JL, Wood LG, et al. Sputum gene expression signature of 6 biomarkers discriminates asthma inflammatory phenotypes. J Allergy Clin Immunol 2014;133:997–1007.
9. Loza MJ, Djukanovic R, Chung KF, et al. Validated and longitudinally stable asthma phenotypes based on cluster analysis of the ADEPT study. Respir Res 2016;17:165.
10. Kuo C-HS, Pavlidis S, Loza M, et al. A Transcriptome-driven Analysis of Epithelial Brushings and Bronchial Biopsies to Define Asthma Phenotypes in U-BIOPRED. Am J Respir Crit Care Med 2017;195:443–55.
11. Chang Y, Glass K, Liu Y-Y, et al. COPD subtypes identified by network-based clustering of blood gene expression. Genomics 2016;107:51–8.
12. Steiling K, van den Berge M, Hijazi K, et al. A dynamic bronchial airway gene expression signature of chronic obstructive pulmonary disease and lung function impairment. Am J Respir Crit Care Med 2013;187:933–42.
13. Lapperre TS, Snoeck-Stroband JB, Gosman MME, et al. Effect of fluticasone with and without salmeterol on pulmonary outcomes in chronic obstructive pulmonary disease: a randomized trial. Ann Intern Med 2009;151:517–27.
14. Kunz LIZ, Ten Hacken NH, Lapperre TS, et al. Airway inflammation in COPD after long-term withdrawal of inhaled corticosteroids. Eur Respir J 2017;49:1700848.

15. Kunz LIZ, Postma DS, Klooster K, et al. Relapse in FEV1 Decline After Steroid Withdrawal in COPD. Chest 2015;148:389–96.
16. van den Berge M, Steiling K, Timens W, et al. Airway gene expression in COPD is dynamic with inhaled corticosteroid treatment and reflects biological pathways associated with disease activity. Thorax 2014;69:14–23.
17. Tasena H, Faiz A, Timens W, et al. microRNA–mRNA regulatory networks underlying chronic mucus hypersecretion in COPD. Eur Respir J 2018;52:1701556.
18. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 2010;26:1572–3.
19. Faiz A, Donovan C, Nieuwenhuis MA, et al. Latrophilin receptors: novel bronchodilator targets in asthma. Thorax 2017;72:74–82.
20. Hodge G, Nairn J, Holmes M, et al. Increased intracellular T helper 1 proinflammatory cytokine production in peripheral blood, bronchoalveolar lavage and intraepithelial T cells of COPD subjects. Clin Exp Immunol 2007;150:22–9.
21. Saetta M, Baraldo S, Corbino L, et al. CD8+ve cells in the lungs of smokers with chronic obstructive pulmonary disease. Am J Respir Crit Care Med 1999;160:711–7.
22. Hogg JC, Chu F, Utokaparch S, et al. The Nature of Small-Airway Obstruction in Chronic Obstructive Pulmonary Disease. N Engl J Med 2004;350:2645–53.
23. Barnes PJ, Shapiro SD, Pauwels RA. Chronic obstructive pulmonary disease: molecular and cellular mechanisms. Eur Respir J 2003;22:672–88.
24. Cosio MG, Majo J, Cosio MG. Inflammation of the airways and lung parenchyma in COPD: role of T cells. Chest 2002;121:160S-165S.
25. Cosio MG, Saetta M, Agusti A. Immunologic Aspects of Chronic Obstructive Pulmonary Disease. N Engl J Med 2009;360:2445–54.
26. Imai K, Mercer BA, Schulman LL, et al. Correlation of lung surface area to apoptosis and proliferation in human emphysema. Eur Respir J 2005;25:250–8.
27. Vogel L, Schoonbrood D, Geluk F, et al. Iron-binding proteins in sputum of chronic bronchitis patients with Haemophilus influenzae infections. Eur Respir J 1997;10:2327–33.
28. van den Berge M, Jonker MR, Miller-Larsson A, et al. Effects of fluticasone propionate and budesonide on the expression of immune defense genes in bronchial epithelial cells. Pulm Pharmacol Ther 2018;50:47–56.
29. Yoshizawa A, Bi K, Keskin DB, et al. TCR-pMHC encounter differentially regulates transcriptomes of tissue-resident CD8 T cells. Eur J Immunol 2018;48:128–50.
30. Chang C, Hu M, Zhu Z, et al. liver-enriched gene 1a and 1b Encode Novel Secretory Proteins Essential for Normal Liver Development in Zebrafish. PLoS One 2011;6:e22910.

31. Titz B, Sewer A, Schneider T, et al. Alterations in the sputum proteome and transcriptome in smokers and early-stage COPD subjects. J Proteomics 2015;128:306–20.
32. Hukkanen J, Väisänen T, Lassila A, et al. Regulation of CYP3A5 by Glucocorticoids and Cigarette Smoke in Human Lung-Derived Cells. J Pharmacol Exp Ther 2003;304:745–52.
33. Seo T, Pahwa P, McDuffie HH, et al. Association between cytochrome P450 3A5 polymorphism and the lung function in Saskatchewan grain workers. Pharmacogenet Genomics 2008;18:487–93.
34. ELBadawy NE, Abdel-Latif RS, El-Hady HA. Association between SERPINB2 Gene Expression by Real Time PCR in Respiratory Epithelial Cells and Atopic Bronchial Asthma Severity. Egypt J Immunol 2017;24:165–81.
35. Woodruff PG, Modrek B, Choy DF, et al. T-helper Type 2–driven Inflammation Defines Major Subphenotypes of Asthma. Am J Respir Crit Care Med 2009;180:388–95.
36. Ito K, Barnes PJ. COPD as a disease of accelerated lung aging. Chest 2009;135:173–80.
37. Valdes AM, Andrew T, Gardner JP, et al. Obesity, cigarette smoking, and telomere length in women. Lancet (London, England) 2005;366:662–4.