ORIGINAL ARTICLE

Noninvasive Analysis of the Sputum Transcriptome Discriminates Clinical Phenotypes of Asthma

Xiting Yan[1,2], Jen-Hwa Chu[3], Jose Gomez[1], Maria Koenigs[1], Carole Holm[1], Xiaoxuan He[1], Mario F. Perez[1], Hongyu Zhao[2,4,5,6], Shrikant Mane[6], Fernando D. Martinez[7], Carole Ober[8], Dan L. Nicolae[8], Kathleen C. Barnes[9], Stephanie J. London[10], Frank Gilliland[11], Scott T. Weiss[3], Benjamin A. Raby[3], Lauren Cohn[1], and Geoffrey L. Chupp[1]

[1]Section of Pulmonary, Critical Care, and Sleep Medicine, Department of Internal Medicine, [2]Biostatistics Resource, Keck Laboratory, [4]Department of Epidemiology and Public Health, [5]Program in Computational Biology and Bioinformatics, and [6]Department of Genetics, Yale University School of Medicine, New Haven, Connecticut; [3]The Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts; [7]Arizona Respiratory Center and BIO5 Institute, University of Arizona, Tucson, Arizona; [8]Department of Human Genetics, University of Chicago, Chicago, Illinois; [9]Department of Medicine, Johns Hopkins University, Baltimore, Maryland; [10]National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, North Carolina; and [11]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California

Abstract

Rationale: The airway transcriptome includes genes that contribute to the pathophysiologic heterogeneity seen in individuals with asthma.

Objectives: We analyzed sputum gene expression for transcriptomic endotypes of asthma (TEA), gene signatures that discriminate phenotypes of disease.

Methods: Gene expression in the sputum and blood of patients with asthma was measured using Affymetrix microarrays. Unsupervised clustering analysis based on pathways from the Kyoto Encyclopedia of Genes and Genomes was used to identify TEA clusters. Logistic regression analysis of matched blood samples defined an expression profile in the circulation to determine the TEA cluster assignment in a cohort of children with asthma to replicate clinical phenotypes.

Measurements and Main Results: Three TEA clusters were identified. TEA cluster 1 had the most subjects with a history of intubation ($P = 0.05$), a lower prebronchodilator FEV_1 ($P = 0.006$), a higher bronchodilator response ($P = 0.03$), and higher exhaled nitric oxide levels ($P = 0.04$) compared with the other TEA clusters. TEA cluster 2, the smallest cluster, had the most subjects that were hospitalized for asthma ($P = 0.04$). TEA cluster 3, the largest cluster, had normal lung function, low exhaled nitric oxide levels, and lower inhaled steroid requirements. Evaluation of TEA clusters in children confirmed that TEA clusters 1 and 2 are associated with a history of intubation ($P = 5.58 \times 10^{-6}$) and hospitalization ($P = 0.01$), respectively.

Conclusions: There are common patterns of gene expression in the sputum and blood of children and adults that are associated with near-fatal, severe, and milder asthma.

Keywords: molecular endotyping; genomic; RNA; severe asthma; pathway analysis

Asthma is a chronic inflammatory disease of the airways that will likely afflict more than 10% of the U.S. population by the end of this decade (1). Differences in genetic susceptibility, environmental exposures, and medication compliance are known to contribute to the heterogeneous clinical manifestations of disease (2, 3). However, it is increasingly evident that pathobiologic

(Received in original form August 8, 2014; accepted in final form March 10, 2015)

Supported by National Institutes of Health (NIH) grants R01HL095390-04 and R01HL118346-01 (G.L.C.); NIH training grants T32HL007778-18 and T15LM007056-26 and FAMRI Young Clinical Scientist Award (J.G.); NIH/NHLBI grant K99HL114651-01A1 (J.-H.C.); Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (S.J.L.); the Mary Beryl Patch Turnbull Scholar Program (K.C.B.); and NCATS grant UL1 3 TR000142 (X.Y.).

Author Contributions: G.L.C. conceived of and designed experiments. X.H., C.H., J.G., M.K., M.F.P., L.C., and S.M. collected the samples and performed the experiments. X.Y. and H.Z. analyzed the data. J.-H.C., F.D.M., C.O., D.L.N., K.C.B., S.J.L., F.G., S.T.W., and B.A.R. provided the Asthma BioRepository for Integrative Genomic Exploration cohort data. X.Y. and G.L.C. wrote the manuscript.

Correspondence and requests for reprints should be addressed to Geoffrey L. Chupp, M.D., 300 Cedar Street, S441 TAC, New Haven, CT 06520-8057. E-mail: geoffrey.chupp@yale.edu

This article has an online supplement, which is accessible from this issue's table of contents at www.atsjournals.org

Am J Respir Crit Care Med Vol 191, Iss 10, pp 1116–1125, May 15, 2015
Copyright © 2015 by the American Thoracic Society
Originally Published in Press as DOI: 10.1164/rccm.201408-1440OC on March 12, 2015
Internet address: www.atsjournals.org

At a Glance Commentary

Scientific Knowledge on the Subject: Asthma is a chronic inflammatory disease of the airways that is clinically and physiologically heterogeneous. Patterns that can be captured by analyzing gene expression levels in the airway and circulation could resolve genes and pathways that contribute to this heterogeneity. Comprehensive studies that examine disease heterogeneity by measuring gene expression in the sputum and associate it with important clinical features of asthma are limited.

What This Study Adds to the Field: This study is the first to use noninvasive analysis of sputum gene expression to identify transcriptomic endotypes of asthma clusters that correlate with clinical characteristics of severe disease including a history of hospitalization and near-fatal asthma attack. The transcriptomic endotypes of asthma clusters are associated with a gene signature in the blood and are evident in both children and adults with asthma, which suggests that there are common patterns of gene expression among children and adults with asthma.

alterations in asthma are also heterogeneous and that differences in the expression of many biologic pathways underlie differences in the phenotypic expressions of the disease (4). Therefore, asthma could be considered as a collection of airway diseases, each driven by a different set of biologic networks with unique but overlapping genomic, transcriptomic, inflammatory, physiologic, and clinical features of disease. In keeping with this paradigm shift, asthma research efforts have moved to defining subgroups of patients with asthma that have different clinical and physiologic manifestations of disease that may be driven by novel biologic mechanisms or relative differences in the expression of known pathways, such as those driven by IL-13 and IL-5 (5, 6). Ultimately, dissecting these subgroups of disease will enable pathogenesis research, therapeutic development, and clinical management to focus on distinct subsets of asthma and their associated clinical phenotypes, leading to a more personalized approach to disease management (7).

To date, most efforts to define asthma subgroups have relied on clustering individuals by clinical features, such as atopic history, age of onset, lung function, or symptoms of severity. These studies, including the Severe Asthma Research Program and the Childhood Asthma Management Program characterizations of asthma clusters, have generated novel insights, but are driven by analytical approaches that are based on differences in parameters that may be distal to many molecular perturbations associated with the disease (8, 9). In contrast to these clinically biased approaches, unsupervised integrative functional transcriptomics has the potential to discriminate asthma subtypes at a level that is reflective of patterns in gene expression, pathobiology, and common clinical and physiologic features of disease: transcriptional endotypes of asthma (TEA) clusters (10, 11). To this end, we conducted an unsupervised clustering analysis of gene expression in the induced sputum of adults and children with asthma and identified three TEA clusters and their associated clinical features of disease. Some of the results of these studies have been previously reported in the form of an abstract (12).

Methods

Yale Center for Asthma and Airway Diseases Cohort

A cross-sectional analysis was conducted on sputum RNA samples collected from subjects with asthma and control subjects that completed the Yale Center for Asthma and Airway Diseases (YCAAD) phenotyping protocol between September 2009 and June 2012. Subjects were greater than or equal to 12 years of age, nonsmokers, and with less than or equal to 10 pack-years of smoking history. Inclusion criteria for asthma included a history, physical examination, and physiologic testing consistent with a diagnosis of asthma based on National Asthma Education and Prevention Program guidelines. Exclusion criteria included smoking within the past year, a history chronic lung disease other than asthma (i.e., chronic obstructive pulmonary disease, allergic bronchopulmonary aspergillosis, Churg-Strauss syndrome, pulmonary vascular disease, or interstitial lung disease); other severe chronic conditions including congestive heart failure, chronic kidney disease, liver disease, or viral infection; or inability to safely undergo the studies required for participation. The protocol was approved by the Yale University School of Medicine Human Investigation Committee, and all patients or their parents provided informed consent.

YCAAD Phenotyping Protocol

An asthma questionnaire was administered and whole blood was collected in RNA isolation tubes (Life Technologies or Applied Biosystems, Grand Island, NY). Exhaled nitric oxide (F_{ENO}) was measured and spirometry was conducted in adherence with the American Thoracic Society guidelines before and after short-acting bronchodilator administration (13). Sputum induction was performed with hypertonic saline, as previously described (14–17). Mucus plugs were dissected from the sputum sample using a microscope, and the cellular and aqueous compartments separated. Total cell count, viability, and differential were determined by hemocytometer, trypan blue exclusion, and Wright-Giemsa stain, respectively, and cell pellets were stored in All-in-One RNA stabilization buffer (Norgen Biotek, Thorhold, Canada).

Genomic Analysis

Sputum cell pellets were processed using the All-in-One purification kit (Norgen Biotek), checked on an Agilent bioanalyzer (Agilent Technologies, Santa Clara, CA), and, if needed, treated again to remove DNA contamination (Qiagen, Gaithersburg, MD). A total of 10 ng of sputum RNA was amplified using the WT-Ovation Pico RNA amplification System (NuGen, San Carlos, CA) and processed per Affymetrix (Santa Clara, CA) protocols, as previously described (18). Total RNA from the blood was isolated using a column-based system (total-RNA kit; Norgen), and if needed, DNA contamination was removed using a DNA clear kit (Qiagen). Hemoglobin reduction of blood samples was used to remove hemoglobin gene transcripts

(GLOBINclear Kit; Ambion, Austin, TX) and samples were checked by Agilent bioanalyzer. Purified RNA from the sputum or blood was processed for gene expression using the Affymetrix HuGene 1.0 ST gene arrays following manufacturer's protocols as previously described. Samples with RNA integrity numbers less than 4.0 were rejected from the analysis. The data can be obtained from GEO database (http://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE56396.

Computational Analyses

An overview of the computational analysis flow can be found in Figure 1. Raw microarray intensity data were processed using R packages for normalization, quality check, batch effect adjustment, and RIN adjustment. Distances between samples were evaluated using a pathway-based method and three clusters were selected to minimize the connectivity criteria (*see* online supplement for details) (19). K-means clustering was applied to assign samples into the selected three TEA clusters. Kruskal-Wallis tests were used to assess differences in continuous clinical, physiologic, and inflammatory asthma phenotypes between the clusters, and the chi-square or Cochran-Armitage test was used to assess differences in categorical phenotypes. False discovery rate (FDR) was estimated using a permutation-based method to adjust for the multiple testing error (20). Differentially expressed genes (DEGs) between each TEA cluster and control subjects were identified as genes with an FDR less than 0.05 using the Student *t* test (21). The online supplement provides more details on computational analyses.

Validation Studies

A TEA cluster classifier was built using an L1 regularized logistic regression model in the YCAAD cohort to predict TEA cluster using blood gene expression and visualized by principal component analysis (*see* online supplement for details) (22, 23). This classifier was applied to 870 whole blood samples selected from the Asthma BioRepository for Integrative Genomic Exploration (Asthma BRIDGE) (24, 25). Cross-platform replication of the TEA cluster clinical phenotypes was conducted (24, 25).

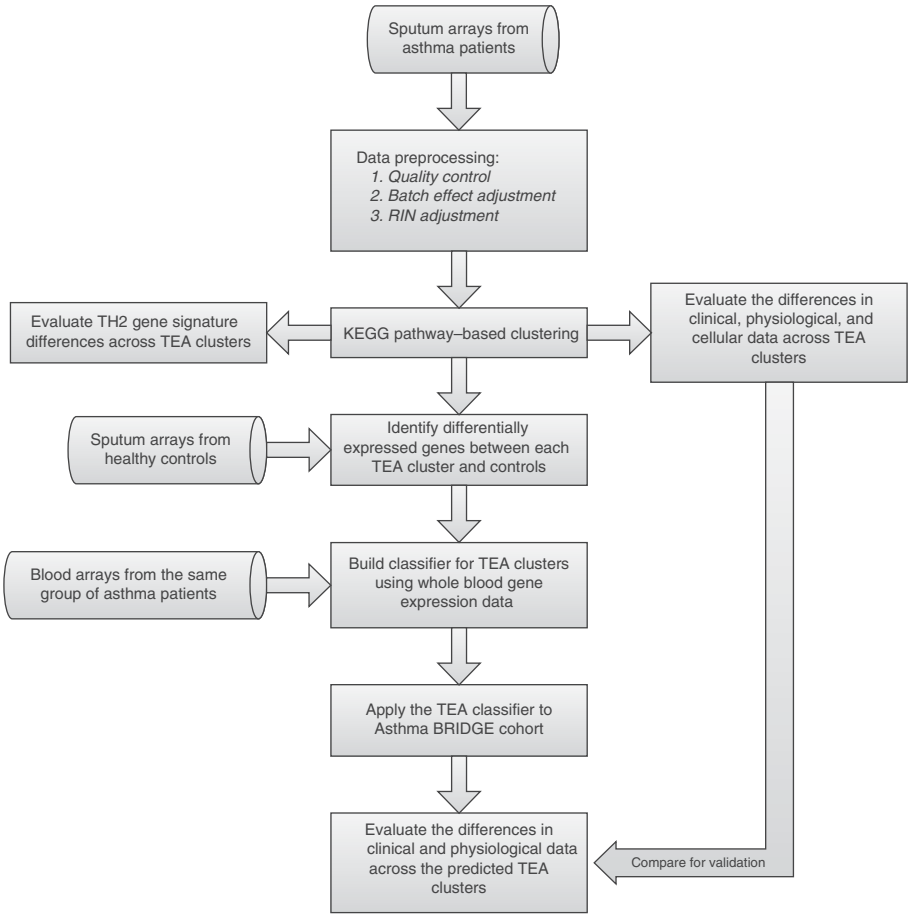

Figure 1. Identification of transcriptomic endotypes of asthma (TEA) clusters. Diagram showing an overview of the computational analysis flow. BRIDGE = BioRepository for Integrative Genomic Exploration; KEGG = Kyoto Encyclopedia of Genes and Genomes; RIN = RNA integrity number.

Results

Identification of Sputum TEA Clusters in YCAAD Cohort

The sputum expression levels of 5,500 genes from 186 Kyoto Encyclopedia of Genes and Genomes pathways were used to assess the pathway-based distance between samples followed by unsupervised K means clustering to define sputum TEA clusters (Figure 2). Three clusters were selected based on the connectivity criteria (*see* Figure E5 in the online supplement). The "relatedness" of the samples within each cluster was evident on a sample distance matrix (Figure 3). This demonstrated that samples within TEA cluster 3 are the most strongly related (*darkest red*) and most homogeneous, followed by TEA cluster 1, and then TEA cluster 2, the smallest cluster. The clusters are not associated with differences in the sputum inflammatory cell populations among the clusters. Therefore, using only sputum gene expression as a discriminator, distinct subgroups of disease can be defined within a heterogeneous group of individuals with asthma.

Phenotypic Characteristics of TEA Clusters

To determine whether the defined TEA clusters correspond to distinct phenotypes of asthma, the clinical, physiologic, and biologic parameters were compared among the TEA clusters (Table 1). The subjects in TEA cluster 1 required a higher daily dose of inhaled steroids

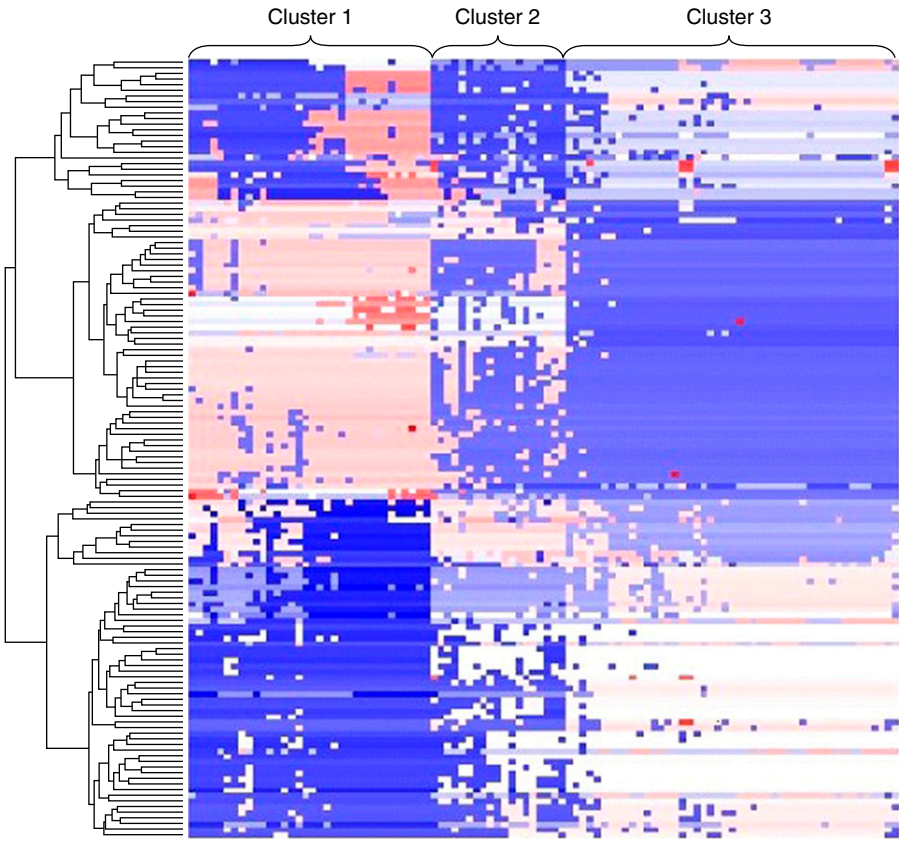

Figure 2. Heatmap showing the clustering results by Kyoto Encyclopedia of Genes and Genomes pathways using MCLUST. The color represents the clustering assignment of each sample by the Kyoto Encyclopedia of Genes and Genomes pathways.

(mean daily inhaled corticosteroid [ICS] dose, 617 μg/day; $P = 0.04$) and were more likely to have a history of an intubation for asthma, compared with the other TEA clusters (18%; $P = 0.05$). In addition, TEA cluster 1 had the lowest prebronchodilator and post-bronchodilator FEV_1 (prebronchodilator percent predicted FEV_1, 73 ± 24%; $P < 0.01$), more bronchodilator reversibility (12 ± 12%; $P = 0.03$), and elevated F_{ENO} levels (mean, 53 ± 43 ppb; $P = 0.03$), compared with the other TEA clusters (Table 2). The fewest number of individuals (n = 19) were in TEA cluster 2. Compared with the other clusters, TEA cluster 2 had the highest percentage of subjects with no atopy (26%; $P = 0.02$), subjects of Hispanic origin ($P = 0.04$), and the highest percentage of individuals that were hospitalized for asthma (68%; $P = 0.03$). TEA cluster 3, the largest cluster (47%), demonstrated the "mildest" phenotypic characteristics of asthma. This TEA cluster had the lowest percentage of subjects with a history of hospitalization or intubation for asthma and the lowest daily ICS dose. In addition, subjects in TEA cluster 3 had preserved prebronchodilator and post-bronchodilator lung function, minimal bronchodilator reversibility, and the lowest F_{ENO} (mean, 38 ± 27 ppb; $P = 0.04$) compared with the other TEA clusters (Table 2). There were no between-cluster differences in sputum cell counts, cell differentials, viability, DNase treatment percentage, percentage of patients on ICS, or RNA integrity number, suggesting that the TEA clustering was not related to a particular cell population, sample processing, or treatment (Table 3).

DEGs in the Airway among TEA Clusters

We considered each TEA cluster as a unique pathobiologic process and compared the gene expression in the sputum of each TEA cluster with control subjects without asthma to determine the genes associated with each of the TEA clusters. Using an FDR threshold of 0.05, there were 31 significant DEGs in TEA cluster 1, a total of 0 DEGs in TEA cluster 2 (15 DEGs had an FDR <0.25), and 27 DEGs in TEA cluster 3 compared with control subjects without asthma (the top 10 most significant DEGs are shown in Tables 4 and 5). In TEA cluster 1, expression of L-histidine decarboxylase, an enzyme in the histamine metabolism pathway that converts histidine to histamine, was increased in individuals with asthma compared with control subjects with no asthma (26). Two DEGs (*EXOSC9* and *SNAPC5*) that code for proteins involved in RNA processing were down-regulated compared with control subjects with no asthma (27). Three DEGs (*NRCAM*, *PCLO*, and *SLC4A4*) that are associated with neuron function were significantly increased in TEA cluster 1 compared with control subjects with no asthma (28–30). In TEA cluster 3, all of the 27 DEGs are up-regulated compared with control subjects. These included *DNAH17*, a force-generating dynein heavy chain motor protein in respiratory epithelium, and defensin β1 (*DEFB1*), an antimicrobial peptide. Both genes have been previously associated with asthma and primary ciliary dyskinesia, respectively (31–33). Gene set enrichment analysis of DEGs using GeneGO MetaCore for TEA cluster 1 and 3 shows that the strongest pathway enrichment was in TEA cluster 3 (*see* Table E2).

Validation of Sputum TEA Clusters Using Blood Gene Expression

Because additional, sufficiently powered datasets of genome-wide sputum gene expression were not available for validation of the TEA cluster model, we turned to a second compartment, the peripheral blood, to validate the TEA clusters. First, using 76 YCAAD subjects for whom both sputum and peripheral blood expression data were available, we identified 53 gene expression signatures in the peripheral blood using L1 regularized logistic regression that predicts an individual's sputum TEA cluster assignment (*see* Table E3). Principal components analysis (PCA) of the peripheral blood expression levels of these 53 genes shows that the first two components separate the population into clusters that closely recapitulate those defined with the sputum data (Figure 4A).

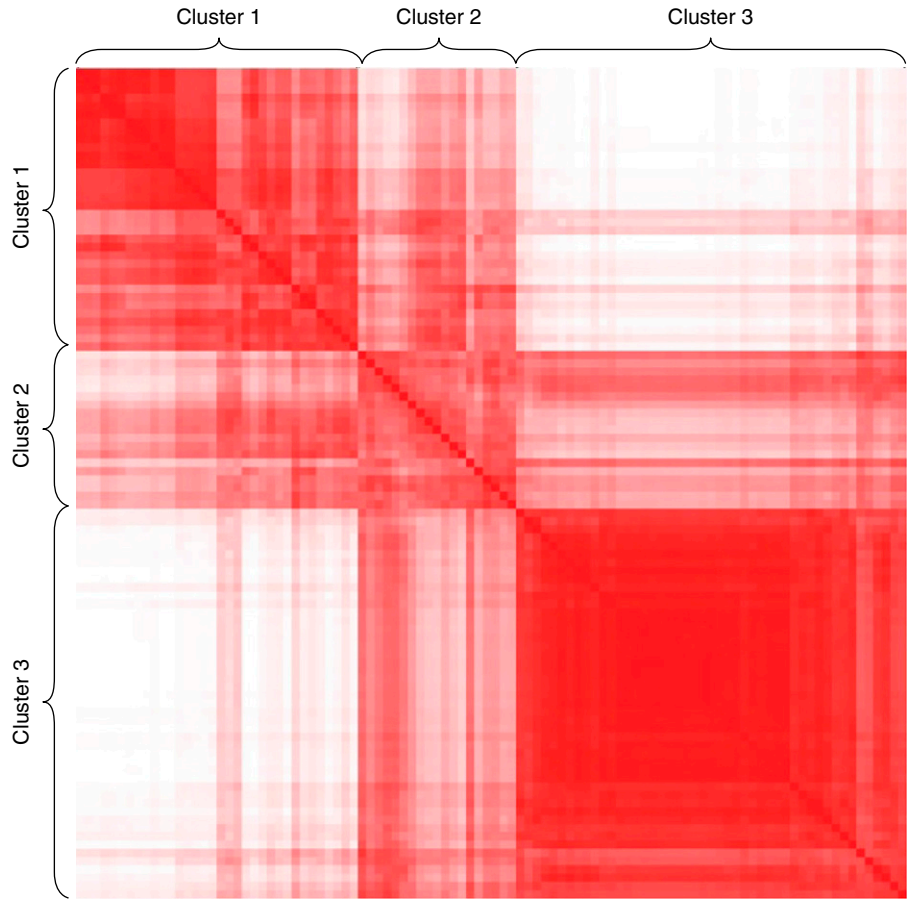

Figure 3. Pathway-based distance matrix among the clusters. The color of entry represents the pathway-based distance between the corresponding two samples. *Red* represents a small distance (samples are strongly related), and *white* represents a longer distance, showing the strength of the clusters (samples are weakly related). Samples within transcriptomic endotypes of asthma cluster 3 are the most strongly related and most homogeneous, followed by clusters 1 and 2, respectively.

Next, to validate the clinical phenotypes of the TEA clusters in a separate cohort, we applied the blood TEA classifier to 870 whole blood samples from the Asthma BRIDGE cohort using the 53 transcripts common to both microarray platforms (*see* online supplement for details) (24, 25). Similar to the patterns observed in the YCAAD cohort, the PCA plot of the blood arrays from the Asthma BRIDGE cohort in Figure 4B shows strong separation between TEA 1 and TEA 3. TEA 2 in the Asthma BRIDGE cohort, however, mingled with TEA 3. The prevalence of each TEA cluster in the Asthma BRIDGE cohort was similar to the YCAAD cohort (31%, 12%, 57% for TEA cluster 1, 2, and 3, respectively; $P = 2.2 \times 10^{-16}$). This demonstrated that a blood signature was able to discriminate subgroups of children with asthma (Asthma BRIDGE) with a prevalence similar to the adults with asthma (YCAAD).

To determine if the childhood and adult TEA clusters have similar clinical features, differences among the clinical phenotypes were evaluated in the TEA clusters in the Asthma BRIDGE cohort (Table 6). Consistent with the clinical phenotypes of the YCAAD cohort, TEA cluster 1 subjects in the Asthma BRIDGE cohort were significantly more likely to have a history of intubation for asthma (8%; $P = 5.58 \times 10^{-6}$) and TEA cluster 2 subjects were the most likely to have a history of hospitalization for asthma (35%; $P = 0.01$), compared with the other TEA clusters. Therefore, in the Asthma BRIDGE cohort, the prevalence of the clusters, the association of TEA cluster 1 with near-fatal disease, and TEA cluster 2 with severe asthma was the same as the YCAAD cohort. Although TEA cluster 1 in the Asthma BRIDGE cohort did not have lower lung function compared with the other TEA clusters (data not shown), 8% of these children had a history of near-fatal asthma attacks.

Discussion

Noninvasive analysis of the sputum transcriptome conducted in these studies identified three TEA clusters with different clinical and physiologic characteristics of disease. Two TEA clusters are associated with phenotypes of severe disease: a history of a near-fatal asthma attack and a history of hospitalization for asthma. These phenotypes were replicated in the TEA clusters of a second cohort of children with asthma that were determined using a unique 53-gene transcriptomic profile in whole blood that is associated with the sputum transcriptome. Taken together, these data suggest that there are common, stable patterns of gene expression in individuals with asthma that are independent of age, age of disease onset, or duration. These TEA clusters are associated with severe phenotypes of asthma and gene signatures in the blood. This indicates that there are systemic alterations in the gene expression that link tissue compartments in patients with severe asthma that can be used to identify subgroups of disease that could be clinically relevant.

The generalizable clinical features associated with each TEA cluster suggest that unsupervised transcriptomic clustering generates disease subgroups that overlap with guideline-defined or Th2 gene level–defined disease severity (*see* Figure E6). Although this shows that there is some biologic overlap between the TEA clusters and Th2 biology, the link is relatively weak because allergic inflammation Kyoto Encyclopedia of Genes and Genomes pathway was not a pathway that significantly contributed to the clustering result. This suggests that the TEA clusters are driven by biologic phenomena that are upstream or possibly parallel to Th2 inflammation (34).

TEA cluster 1 has the highest percentage of subjects with a history of near-fatal asthma in both the YCAAD and Asthma BRIDGE cohorts, despite a large

Table 1. Phenotypic Characteristics of TEA Clusters in the YCAAD Cohort

	Control Subjects (*n* = 12)	Cluster 1	Cluster 2	Cluster 3	*P* Value
Prevalence, n		34	19	47	0.003
Age at visit, yr	37 ± 14	51 ± 13	49 ± 16	45 ± 17	0.32
Female sex, n (%)	5 (42)	23 (68)	15 (79)	39 (83)	0.26
Race					0.58
White, n (%)	12 (100)	22 (65)	14 (74)	37 (79)	
Black, n (%)	0 (0)	10 (29)	4 (21)	7 (15)	
Other, n (%)	0 (0)	1 (3)	1 (5)	3 (6)	
Hispanic origin, n (%)	0 (0)	1 (3)	4 (21)	7 (15)	0.04
BMI, kg/m^2	23.6 ± 2.8	30.0 ± 7.2	30 ± 7.3	29.3 ± 8.0	0.84
History of atopy, n (%)	7 (58)	33 (97)	14 (74)	43 (92)	0.02
Age of symptom onset	NA	25.8 ± 19.1	29.3 ± 20.4	20.7 ± 20.9	0.17
Disease duration, yr	NA	25.2 ± 17.5	20.7 ± 16.9	24.2 ± 17.3	0.70
History of hospitalization, n (%)	NA	13 (38.1)	13 (68.4)	16 (34.0)	0.03
History of intubations, n (%)	NA	6.0 (18)	2.0 (11)	2.0 (4)	0.05
OCS tapers in past year, n (%)	NA	19 (55.9)	12 (63.2)	24 (51.1)	0.67
ACT score	NA	16 ± 6.4	14 ± 6.6	18 ± 5.1	0.22
ICS dose, μg/d	NA	617 ± 448	530 ± 449	396 ± 356	0.04
ICS use, yes, n (%)	NA	27 (79)	17 (89)	31 (66)	0.10
Chronic OCS use, n (%)	NA	4 (11.8)	2 (10.5)	3 (6.4)	0.68

Definition of abbreviations: ACT = Asthma Control Test; BMI = body mass index; ICS = inhaled corticosteroids; NA = not applicable; OCS = oral corticosteroids; TEA = transcriptomic endotypes of asthma; YCAAD = Yale Center for Asthma and Airway Diseases.
Data are means ± SD, except where indicated. *P* values for comparisons among TEA clusters were determined using Kruskal-Wallis or Cochran-Armitage test. The false discovery rate estimated by the permutation-based method is 11%.

difference in age among the individuals. This cluster also has the lowest baseline lung function, the highest bronchodilator reversibility, the highest F_{ENO} levels, and the highest doses of ICS: all characteristics that are associated with near-fatal asthma (35). Given that the adults and children in TEA cluster 1 are linked by a common transcriptomic signature in the airway that is associated with epithelial cell differentiation (*EXOSC9* and *SNAPC5*), neurohumoral hemostasis (*NRCAM* and *PCLO*), and histamine synthesis (*DNAH17* and *DEFB1*), it is plausible that these genes contribute to a greater risk of severe bronchospasm and near-fatal asthma associated with this cluster (26–30).

TEA cluster 2 is the least common TEA cluster in both cohorts (19% of YCAAD and 12% of Asthma BRIDGE) and has the most within-cluster heterogeneity, compared with the other TEA clusters (color heterogeneity seen in Figure 3). These individuals also have severe disease and are more likely to have been hospitalized for asthma. Although the evaluation of larger cohorts is required to further define this cluster and its associated DEGs (no significant genes associated with this TEA cluster were identified by an FDR cutoff of 0.05 in part because of the small number of individuals in this cluster), the transcriptomic discrimination of this cluster suggests that there are distinct

Table 2. Pulmonary Function of TEA Clusters in YCAAD Cohort

	Control Subjects (*n* = 12)	Cluster 1 (*n* = 34)	Cluster 2 (*n* = 19)	Cluster 3 (*n* = 47)	*P* Value
FEV_1, % of predicted value					
Pre–β_2-agonist use	96 ± 11	73 ± 24	76 ± 22	86 ± 22	0.006
Post–β_2-agonist use	98 ± 14	80 ± 24	81 ± 20	91 ± 22	0.493
FVC, % of predicted value					
Pre–β_2-agonist use	86 ± 22	85 ± 22	86 ± 18	96 ± 19	0.09
Post–β_2-agonist use	91 ± 20	90 ± 20	88 ± 18	97 ± 18	0.28
FEV_1/FVC					
Pre–β_2-agonist use	0.77 ± 0.62	0.67 ± 0.13	0.70 ± 0.11	0.72 ± 0.10	0.13
Post–β_2-agonist use	0.79 ± 0.50	0.70 ± 0.13	0.72 ± 0.12	0.80 ± 0.10	0.06
BDR, %	2 ± 6.2	12 ± 12	9 ± 13	6 ± 7	0.03
F_{ENO}, ppb	20 ± 9.7	53 ± 43	52 ± 42	38 ± 27	0.04

Definition of abbreviations: BDR = bronchodilator response; F_{ENO} = exhaled nitric oxide; TEA = transcriptomic endotypes of asthma; YCAAD = Yale Center for Asthma and Airway Diseases.
Data are means ± SD. *P* values for comparisons among TEA clusters were determined using Kruskal-Wallis or Cochran-Armitage test. The false discovery rate estimated by the permutation-based method is 5.6%.

Table 3. Sputum Characteristics of TEA Clusters in the YCAAD Cohort

	Control Subjects (*n* = 12)	Cluster 1 (*n* = 34)	Cluster 2 (*n* = 19)	Cluster 3 (*n* = 47)	*P* Value
Mucus cell concentration*	40.86 ± 20.98	83.02 ± 105.75	89.23 ± 143.61	73.72 ± 62.48	0.63
Squamous, %	8.2 ± 6.7	7.9 ± 7.0	8.0 ± 5.9	9.2 ± 6.9	0.60
Viability, %	58.1 ± 9.6	56.5 ± 16.1	64.4 ± 11.9	61.7 ± 17.8	0.14
Neutrophils, %	34.6 ± 10.0	41.5 ± 13.0	41.9 ± 15.2	37.8 ± 14.6	0.34
Eosinophil, %	1.5 ± 1.8	5.8 ± 6.7	4.7 ± 5.9	5.2 ± 7.7	0.91
Macrophage, %	61.3 ± 11.8	50.9 ± 13.0	50.9 ± 16.0	55.4 ± 15.4	0.31
Lymphocyte, %	1.0 ± 0.9	1.3 ± 1.5	1.2 ± 1.0	1.3 ± 1.4	0.90
Bronchial epithelial cell, %	1.6 ± 4.3	0.8 ± 1.5	1.3 ± 3.3	0.4 ± 1.0	0.26
RIN, mean	7.6 ± 1.1	7.4 ± 1.2	7.5 ± 1.0	7.7 ± 1.4	0.1

Definition of abbreviations: RIN = RNA integrity number; TEA = transcriptomic endotypes of asthma; YCAAD = Yale Center for Asthma and Airway Diseases. Data are means ± SD. *P* values for comparisons among TEA clusters were determined using Kruskal-Wallis or Cochran-Armitage test.
*Cells per microliter × 10^4.

pathobiologic differences between patients that have near-fatal asthma attacks and those that have severe exacerbations requiring hospitalization. Consistent with this concept is the fact that patients in TEA cluster 2 have the highest levels of YKL-40 in the sputum among the clusters ($P = 0.03$, data not shown), findings consistent with a possible YKL-40 endotype/phenotype with increased risk of exacerbations and abnormal post-bronchodilator FEV_1 (36, 37).

Individuals in TEA cluster 3 have clinical features most consistent with mild disease including increased expression of DEFB1, a gene that has been associated with mild asthma in multiple studies (31). Compared with the other TEA clusters, these individuals have preserved lung function, the lowest ICS dose, and the lowest F_{ENO} level and are less likely to have been hospitalized or intubated for asthma. TEA cluster 3 is also the most strongly related with the least within-cluster heterogeneity (consistent *red* color of this cluster in Figure 3). It is also the most common cluster with a prevalence of approximately 50% in children and adults. Interestingly, PCA analysis of the blood transcriptome shows that in children, TEA cluster 3 overlaps with cluster 1 and 2 (Asthma BRIDGE cohort, Figure 4B), but is

Table 4. Top 10 Differentially Expressed Genes between TEA Cluster 1 and Control Subjects

Gene Name	Gene Symbol	Biologic Processes	Function	Fold Change	*P* Value	FDR	PubMed References
Hemogen	*HEMGN*	Apoptosis	Transcription factor	1.25	2.79×10^{-7}	4.01×10^{-3}	16
Piccolo (presynaptic cytomatrix protein)	*PCLO*	Cytoskeleton organization, regulation of exocytosis	Protein	1.21	6.26×10^{-7}	4.01×10^{-3}	26
Cartilage-associated protein	*CRTAP*	Extracellular matrix organization	Scaffolding protein	−1.48	7.87×10^{-7}	4.01×10^{-3}	25
Exosome component 9	*EXOSC9*	RNA processing	RNase complex component	−2.10	1.11×10^{-6}	4.01×10^{-3}	32
Chromosome 9 open reading frame 173	*C9orf173*	NA	NA	1.36	1.23×10^{-6}	4.01×10^{-3}	2
Small nuclear RNA activating complex, polypeptide 5, 19 kD	*SNAPC5*	DNA-dependent regulation of transcription	Transcription factor	−1.08	1.23×10^{-6}	4.01×10^{-3}	9
Impact RWD domain protein	*IMPACT*	Negative regulation of protein phosphorylation	Enzyme	−2.12	1.44×10^{-6}	4.01×10^{-3}	11
Solute carrier family 4, sodium bicarbonate cotransporter, member 4	*SLC4A4*	Sodium ion transport	Membrane transporter protein	1.22	1.45×10^{-6}	4.01×10^{-3}	61
Histidine decarboxylase	*HDC*	Histamine biosynthesis	Enzyme	1.39	1.94×10^{-6}	4.78×10^{-3}	48
Neuronal cell adhesion molecule	*NRCAM*	Neuron migration	Cell adhesion molecules	1.19	2.29×10^{-6}	5.08×10^{-3}	40

Definition of abbreviations: FDR = false discovery rate; NA = not applicable; TEA = transcriptomic endotypes of asthma.

Table 5. Top 10 Differentially Expressed Genes between TEA Cluster 3 and Control Subjects

Gene Name	Gene Symbol	Biologic Processes	Functions	Fold Change	*P* Value	FDR	PubMed References
Heart development protein with EGF-like domains 1	*HEG1*	Endothelial cell angiogenesis	Cell adhesion molecules	1.78	1.09×10^{-7}	2.42×10^{-3}	7
Small nucleolar RNA, C/D box 104	*SNORD104*	RNA modification	Noncoding RNA	2.58	2.46×10^{-7}	2.73×10^{-3}	2
Dynein, axonemal, heavy chain 17	*DNAH17*	Ciliary motility	Microtubule-associated motor protein	1.88	2.61×10^{-6}	1.18×10^{-2}	7
Cbl proto-oncogene, E3 ubiquitin protein ligase B	*CBLB*	Regulation of T-cell anergy	Ubiquitin protein ligase	2.04	2.95×10^{-6}	1.18×10^{-2}	93
Defensin, β1	*DEFB1*	Innate immune response	Antimicrobial peptide	2.85	2.97×10^{-6}	1.18×10^{-2}	133
Nonprotein coding RNA 204	*NCRNA204*	NA	Noncoding RNA	2.20	3.55×10^{-6}	1.18×10^{-2}	1
Transcription elongation factor A (SII) N-terminal and central domain containing	*TCEANC*	RNA elongation	Transcription elongation factor	2.22	6.58×10^{-6}	1.82×10^{-2}	10
Radical *S*-adenosyl methionine domain containing 2	*RSAD2*	Regulation of Toll-like receptor 9 signaling pathway	Antiviral protein	1.57	1.04×10^{-5}	2.07×10^{-2}	28
Purinergic receptor P2Y, G-protein coupled, 14	*P2RY14*	Regulation of inflammation	UDP-glucose receptor	5.14	1.04×10^{-5}	2.07×10^{-2}	31
Malignant fibrous histiocytoma amplified sequence	*MFHAS1*	Cell cycle	Protein	4.14	1.11×10^{-5}	2.07×10^{-2}	11

Definition of abbreviations: EGF = epidermal growth factor; FDR = false discovery rate; NA = not applicable; TEA = transcriptomic endotypes of asthma.

distinct from the other TEA clusters in adults (Figure 4A). This suggests that the transcriptome (and clinical phenotype) of the individuals in this TEA cluster could change over time.

These studies also demonstrate that the unsupervised clustering analysis of gene expression in the sputum and/or blood has the capacity to discriminate subgroups of asthma that are independent of clinical characteristics typically used to study severe asthma (i.e., body mass index, atopy, FEV_1). This is caused, in part, by the clustering approach we developed that is distinctly different compared with conventional approaches that use clinical features, and large differences in gene expression compared with subjects with no asthma to select genes for the clustering analysis (5, 9–11). In contrast, the analytical approach used herein was not biased by

Table 6. Phenotypic Characteristics of TEA Clusters in the Asthma BRIDGE Cohort

	Cluster 1 (*n* = *266*)	Cluster 2 (*n* = *105*)	Cluster 3 (*n* = *499*)	*P* Value
Prevalence in cohort	31%	12%	57%	$<2.2 \times 10^{-16}$
Age at visit, yr	14.8 ± 8	10.1 ± 5	12.6 ± 7	3.27×10^{-8}
Sex, n (%) female	128 (48)	45 (43)	212 (43)	0.31
Race				1.68×10^{-7}
White, n (%)	102 (38)	79 (75)	284 (57)	
Black, n (%)	105 (40)	9 (9)	124 (25)	
Other, n (%)	29 (11)	7 (7)	37 (7)	
Hispanic origin, n (%)	30 (11)	10 (10)	54 (11)	0.09
History of atopy, n (%)	69 (26)	40 (38)	113 (23)	0.0013
Age of symptom onset	3.50 ± 3.21	3.32 ± 2.89	3.48 ± 2.80	0.86
History of hospitalization, n (%)	91 (34)	37 (35)	128 (26)	0.011
History of intubations, n (%)	21 (8)	0 (0)	9 (2)	5.58×10^{-6}
ACT score	14 ± 4	12 ± 3	13 ± 3	8.79×10^{-7}

Definition of abbreviations: ACT = Asthma Control Test; BRIDGE = BioRepository for Integrative Genomic Exploration; TEA = transcriptomic endotypes of asthma. Data are means ± SD, except where indicated.

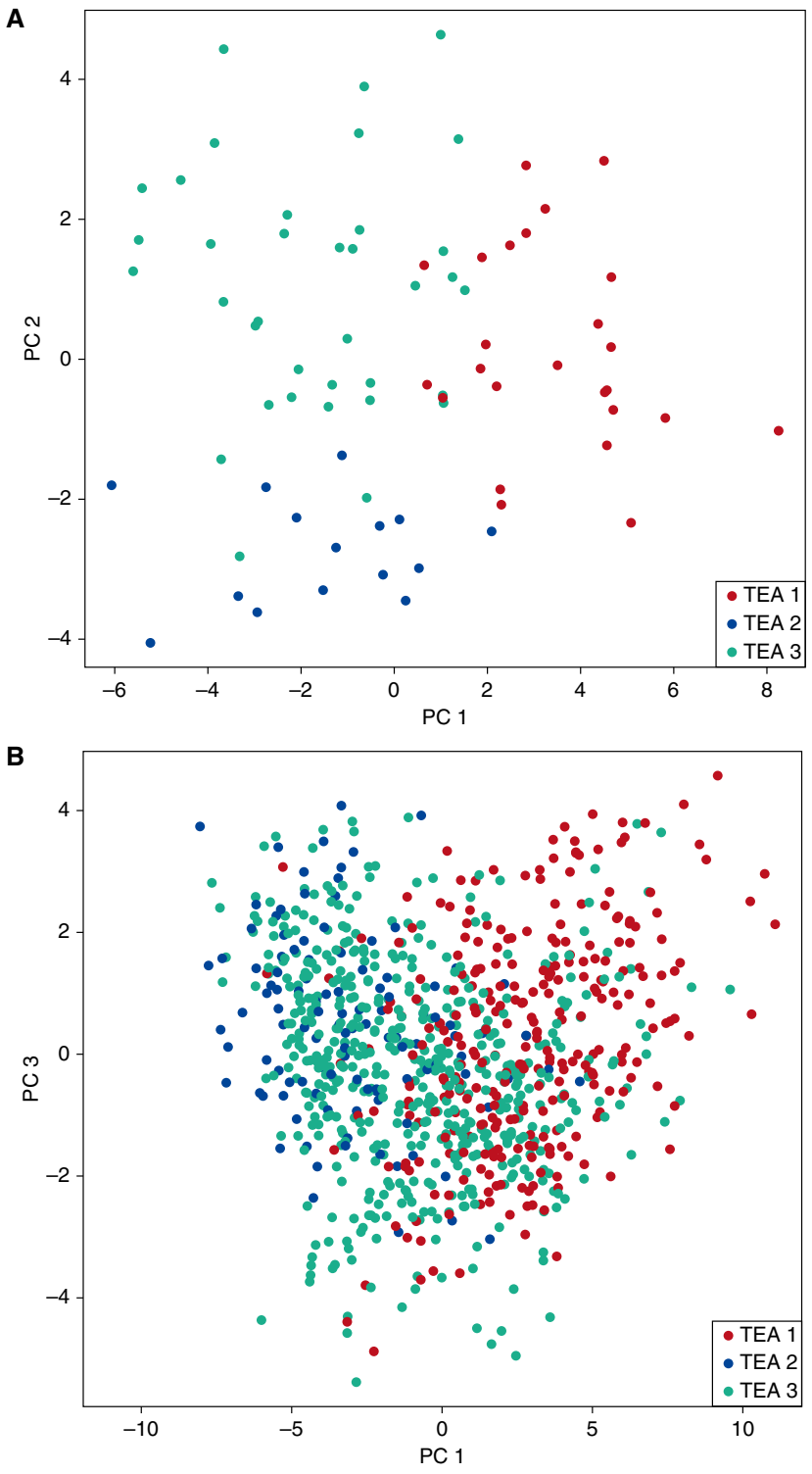

Figure 4. Data visualization of the transcriptomic endotypes of asthma (TEA) clusters using the 53 blood expression. (*A*) Principal components analysis plot of the 76 matched blood arrays in the Yale Center for Asthma and Airway Diseases cohort. (*B*) Principal components analysis plot of the blood arrays from the Asthma BioRepository for Integrative Genomic Exploration cohort. TEA cluster assignment was predicted using the TEA cluster classifier built in the Yale Center for Asthma and Airway Diseases cohort. PC = principal component.

clinical phenotypes and is solely based on ontologically derived, pathway-based gene expression. This approach results in reduced background from random statistical events interfering with the clustering algorithm and the overwhelming effects of analyzing gene expression that is different between patients with asthma and control subjects: eliminating gene expression signals that are associated with disease heterogeneity. Ultimately we found two different sets of genes in the blood and sputum that are associated with the same or similar clinical phenotypes. Because clustering analysis using only blood gene expression identified clusters without unique clinical features (data not shown), we believe that the evaluation of gene expression in the primary organ of involvement is essential to dissect disease heterogeneity of chronic inflammatory airway disease.

Although TEA clusters are clearly distinguished, there remains heterogeneity within each cluster, especially within TEA clusters 1 and 2. This suggests that analysis of larger populations of patients will define additional TEA clusters that are biologically similar within the clusters we have defined. These may reveal additional novel molecular phenomena that further define the heterogeneity of disease. In addition, longitudinal studies of patients with asthma currently underway will determine how stable and robust the TEA clusters are over time, and will define the potential of unsupervised transcriptomic analysis of the blood and sputum to identify patients at risk of adverse outcomes, such as near-fatal and severe asthma exacerbations early in the course of their disease. Ultimately, these studies will determine if transcriptomic signatures in the blood and/or airway have the capacity to personalize approaches to the management of asthma, enhance outcomes and selection for existing and emerging treatments, or will be most useful to advance the pathogenesis research in asthma and other complex diseases. ■

Author disclosures are available with the text of this article at www.atsjournals.org.

Acknowledgment: The authors thank all the subjects who participated in this study, Donna Cook for data management and support enrolling subjects, and Arron Mitchell for development of the Yale Center for Asthma and Airway Diseases database.

References

1. Moorman JE, Zahran H, Truman BI, Molla MT; Centers for Disease Control and Prevention (CDC). Current asthma prevalence - United States, 2006-2008. *MMWR Surveill Summ* 2011;60:84–86.
2. Wenzel SE. Asthma phenotypes: the evolution from clinical to molecular approaches. *Nat Med* 2012;18:716–725.
3. American Thoracic Society. Proceedings of the ATS workshop on refractory asthma: current understanding, recommendations, and unanswered questions. *Am J Respir Crit Care Med* 2000;162: 2341–2351.
4. Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, Heath S, von Mutius E, Farrall M, Lathrop M, Cookson WO; GABRIEL Consortium. A large-scale, consortium-based genomewide association study of asthma. *N Engl J Med* 2010;363:1211–1221.
5. Woodruff PG, Modrek B, Choy DF, Jia G, Abbas AR, Ellwanger A, Koth LL, Arron JR, Fahy JV. T-helper type 2-driven inflammation defines major subphenotypes of asthma. *Am J Respir Crit Care Med* 2009; 180:388–395.
6. Peters MC, Mekonnen ZK, Yuan S, Bhakta NR, Woodruff PG, Fahy JV. Measures of gene expression in sputum cells can identify TH2-high and TH2-low subtypes of asthma. *J Allergy Clin Immunol* 2014;133:388–394.
7. Moore WC, Bleecker ER, Curran-Everett D, Erzurum SC, Ameredes BT, Bacharier L, Calhoun WJ, Castro M, Chung KF, Clark MP, *et al*.; National Heart, Lung, Blood Institute's Severe Asthma Research Program. Characterization of the severe asthma phenotype by the National Heart, Lung, and Blood Institute's Severe Asthma Research Program. *J Allergy Clin Immunol* 2007;119:405–413.
8. Howrylak JA, Fuhlbrigge AL, Strunk RC, Zeiger RS, Weiss ST, Raby BA; Childhood Asthma Management Program Research Group. Classification of childhood asthma phenotypes and long-term clinical responses to inhaled anti-inflammatory medications. *J Allergy Clin Immunol* 2014;133:1289–1300, e1–e12.
9. Moore WC, Meyers DA, Wenzel SE, Teague WG, Li H, Li X, D'Agostino R Jr, Castro M, Curran-Everett D, Fitzpatrick AM, *et al*.; National Heart, Lung, and Blood Institute's Severe Asthma Research Program. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. *Am J Respir Crit Care Med* 2010;181:315–323.
10. Haldar P, Pavord ID, Shaw DE, Berry MA, Thomas M, Brightling CE, Wardlaw AJ, Green RH. Cluster analysis and clinical asthma phenotypes. *Am J Respir Crit Care Med* 2008;178:218–224.
11. Baines KJ, Simpson JL, Wood LG, Scott RJ, Gibson PG. Transcriptional phenotypes of asthma defined by gene expression profiling of induced sputum samples. *J Allergy Clin Immunol* 2011; 127:153–60, e1–e9.
12. Chupp GL, Perez MF, Gomez-Arroyo JG, He S, Holm C, Cook D, Zhao H, Cohn LE, Yan X. Determination of asthma phenotypes by unsupervised transcriptomic cluster analysis of induced sputum [abstract]. *Am J Respir Crit Care Med* 2013:187:A2328.
13. American Thoracic Society. Standardization of spirometry, 1994 update. *Am J Respir Crit Care Med* 1995;152:1107–1136.
14. Covar RA, Spahn JD, Martin RJ, Silkoff PE, Sundstrom DA, Murphy J, Szefler SJ. Safety and application of induced sputum analysis in childhood asthma. *J Allergy Clin Immunol* 2004;114:575–582.
15. Gelder CM, Thomas PS, Yates DH, Adcock IM, Morrison JF, Barnes PJ. Cytokine expression in normal, atopic, and asthmatic subjects using the combination of sputum induction and the polymerase chain reaction. *Thorax* 1995;50:1033–1037.
16. Grootendorst DC, van den Bos JW, Romeijn JJ, Veselic-Charvat M, Duiverman EJ, Vrijlandt EJ, Sterk PJ, Roldaan AC. Induced sputum in adolescents with severe stable asthma: safety and the relationship of cell counts and eosinophil cationic protein to clinical severity. *Eur Respir J* 1999;13:647–653.
17. Green RH, Brightling CE, Woltmann G, Parker D, Wardlaw AJ, Pavord ID. Analysis of induced sputum in adults with asthma: identification of subgroup with isolated sputum neutrophilia and poor response to inhaled corticosteroids. *Thorax* 2002;57:875–879.
18. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AM, Pletikos M, Meyer KA, Sedmak G, *et al*. Spatio-temporal transcriptome of the human brain. *Nature* 2011;478:483–489.
19. Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis. *Bioinformatics* 2005;21:3201–3212.
20. Millstein J, Volfson D. Computationally efficient permutation-based confidence interval estimation for tail-area FDR. *Front Genet* 2013;4: 179.
21. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* 1995;57:289–300.
22. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat* 2004;32:407–451.
23. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Series B Stat Methodol* 1996;58:267–288.
24. Raby BA BK, Beaty TH, Bosco A, Carey VJ, Castro M, Cheadle C, Gilliland FD, Islam KTS, Salam MT, Kelly R, *et al*. Asthma BRIDGE: the Asthma BioRepository for Integrative Genomic Exploration. *Am J Respir Crit Care Med* 2010;183:A6189.
25. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* 2008;24:1547–1548.
26. Koarai A, Ichinose M, Ishigaki-Suzuki S, Yamagata S, Sugiura H, Sakurai E, Makabe-Kobayashi Y, Kuramasu A, Watanabe T, Shirato K, *et al*. Disruption of L-histidine decarboxylase reduces airway eosinophilia but not hyperresponsiveness. *Am J Respir Crit Care Med* 2003;167:758–763.
27. Mistry DS, Chen Y, Sen GL. Progenitor function in self-renewing human epidermis is maintained by the exosome. *Cell Stem Cell* 2012;11: 127–135.
28. Aoki T, Matsumoto Y, Hirata K, Ochiai K, Okada M, Ichikawa K, Shibasaki M, Arinami T, Sumazaki R, Noguchi E. Expression profiling of genes related to asthma exacerbations. *Clin Exp Allergy* 2009;39: 213–221.
29. Laprise C. The Saguenay-Lac-Saint-Jean asthma familial collection: the genetics of asthma in a young founder population. *Genes Immun* 2014;15:247–255.
30. Dorfman R, Li W, Sun L, Lin F, Wang Y, Sandford A, Paré PD, McKay K, Kayserova H, Piskackova T, *et al*. Modifier gene study of meconium ileus in cystic fibrosis: statistical considerations and gene mapping results. *Hum Genet* 2009;126:763–778.
31. Levy H, Raby BA, Lake S, Tantisira KG, Kwiatkowski D, Lazarus R, Silverman EK, Richter B, Klimecki WT, Vercelli D, *et al*. Association of defensin beta-1 gene polymorphisms with asthma. *J Allergy Clin Immunol* 2005;115:252–258.
32. Park YD, Lyou YJ, Lee KJ, Lee DY, Yang JM. Towards profiling the gene expression of fibroblasts from atopic dermatitis patients: human 8K complementary DNA microarray. *Clin Exp Allergy* 2006;36:649–657.
33. Chodhari R, Mitchison HM, Meeks M. Cilia, primary ciliary dyskinesia and molecular genetics. *Paediatr Respir Rev* 2004;5:69–76.
34. National Asthma Education and Prevention Program. Expert Panel Report 3 (EPR-3): guidelines for the diagnosis and management of asthma—summary report 2007. *J Allergy Clin Immunol* 2007;120(5 Suppl):S94–S138.
35. Wenzel SE, Szefler SJ, Leung DY, Sloan SI, Rex MD, Martin RJ. Bronchoscopic evaluation of severe asthma. Persistent inflammation associated with high dose glucocorticoids. *Am J Respir Crit Care Med* 1997;156:737–743.
36. Chupp GL, Lee CG, Jarjour N, Shim YM, Holm CT, He S, Dziura JD, Reed J, Coyle AJ, Kiener P, *et al*. A chitinase-like protein in the lung and circulation of patients with severe asthma. *N Engl J Med* 2007; 357:2016–2027.
37. Cunningham J, Basu K, Tavendale R, Palmer CN, Smith H, Mukhopadhyay S. The CHI3L1 rs4950928 polymorphism is associated with asthma-related hospital admissions in children and young adults. *Ann Allergy Asthma Immunol* 2011;106: 381–386.

Non-invasive Analysis of the Sputum Transcriptome Discriminates Clinical Phenotypes of Asthma

Xiting Yan, Jen-Hwa Chu, Jose Gomez, Maria Koenigs, Carole Holm, Xiaoxuan He, Mario F. Perez, Hongyu Zhao, Shrikant Mane, Fernando D. Martinez, Carole Ober, Dan L. Nicolae, Kathleen C. Barnes, Stephanie J. London, Frank Gilliland, Scott T. Weiss, Benjamin A. Raby, Lauren Cohn and Geoffrey L Chupp

Online Data Supplement

Table of Contents

Non-invasive Analysis of the Sputum Transcriptome Discriminates Clinical Phenotypes of Asthma 1

Microarray Data Preprocessing 3

Sample Screening 3

Normalization and Quality Control 3

Batch Effect and RIN Number Adjustment 3

Identifying the Sputum TEA Clusters 5

Pathway based distance between samples 5

Pathways driving each TEA cluster 6

Pathways enriched for DEGs of each TEA cluster 7

TH2 gene expression among the TEA clusters 8

Sputum TEA classifier using blood gene expression 8

Validation of clinical phenotypes associated with TEA clusters in a second cohort 11

Asthma BRIDGE Cohort 11

Sputum TEA cluster identification 11

REFERENCES 12

Microarray Data Preprocessing

Sample Screening

We have collected 190 visits by subjects including both asthma patients and healthy controls. All these subjects have total pack year ≤10, squamous cell %≤25% and have at least one array hybridized. Some of these visits were replicated visits by the same patient. Therefore, for the same subject, we only kept the visit with the highest RNA quality (RIN number) which resulted in 171 visits by subjects including 12 healthy normal controls and 159 asthma patients. For the 159 asthma patients, 108 were measured for the sputum gene expression profile using the Affymetrix HuGene 1.0 ST arrays. A flow chart showing the sample size for the sample screening steps is shown in Figure E1.

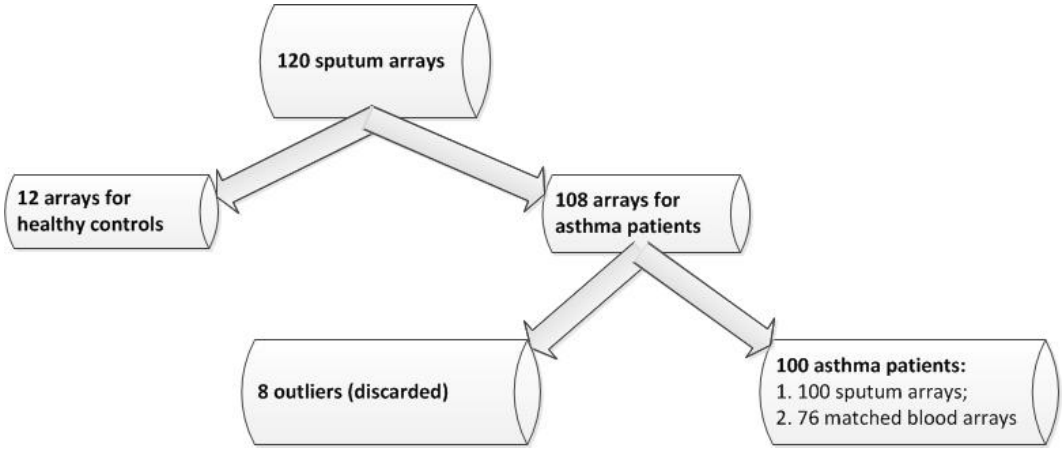

Figure E1. Flow chart showing the sample size for the sample screening steps.

Normalization and Quality Control

The raw microarray data (CEL files) were corrected for background signal, quantile normalized and summarized using robust multiarray average[E1] (RMA) implemented by the XPS R package to obtain the log2 transformed normalized gene expression levels in both blood and sputum. Then principal component analysis (PCA) was applied to visualize the data and remove outlying arrays. To make sure the data has enough information for the batch effect adjustment, array batches with less than 3 arrays were removed from further analysis. In total, 8 arrays were removed based resulting in 112 sputum arrays for 100 asthma patients and 12 healthy controls.

Batch Effect and RIN Number Adjustment

Visualization of the array data using PCA showed existed batch effect in the data (Figure E2A). ComBat[E2] was applied to adjust the data for batch effect (Figure E2B). The batch effect adjusted data was further adjusted for RIN number using a linear regression model to remove the effect of RNA quality on the expression data. The three dimensional PCA plot of all the 112 sputum arrays after both the batch effect and RIN number adjustment is shown in Figure E3.

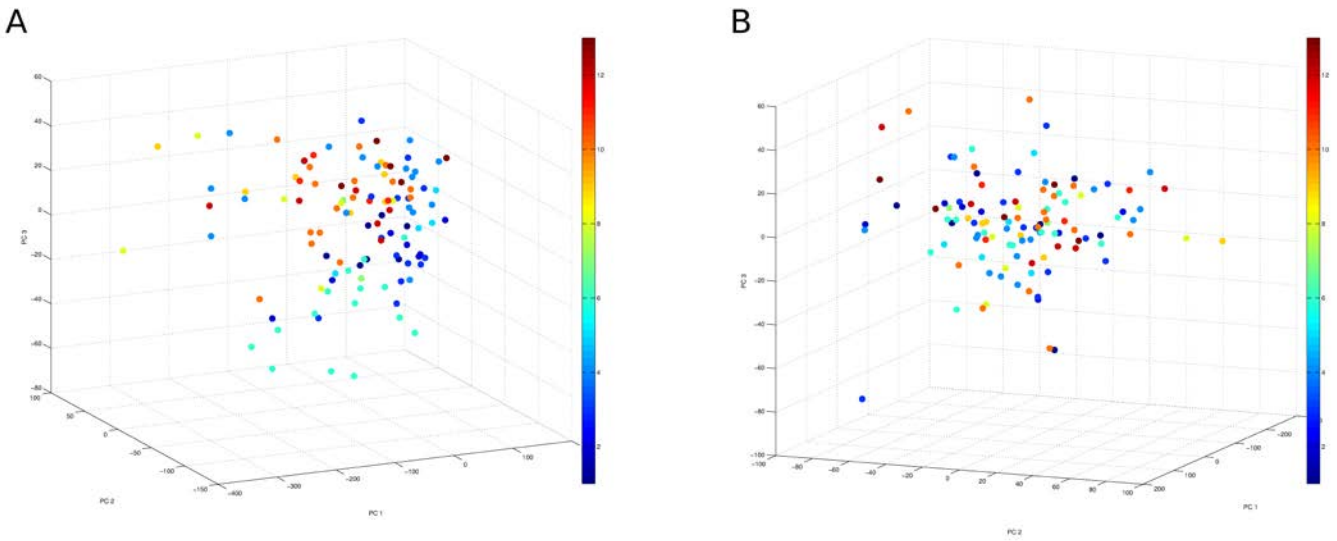

Figure E2: PCA plots for all the 112 sputum arrays (A) before the batch effect adjustment and (B) after the batch effect adjustment. Arrays were colored based on their batches so that arrays scanned at closer time have closer colors from the color bar.

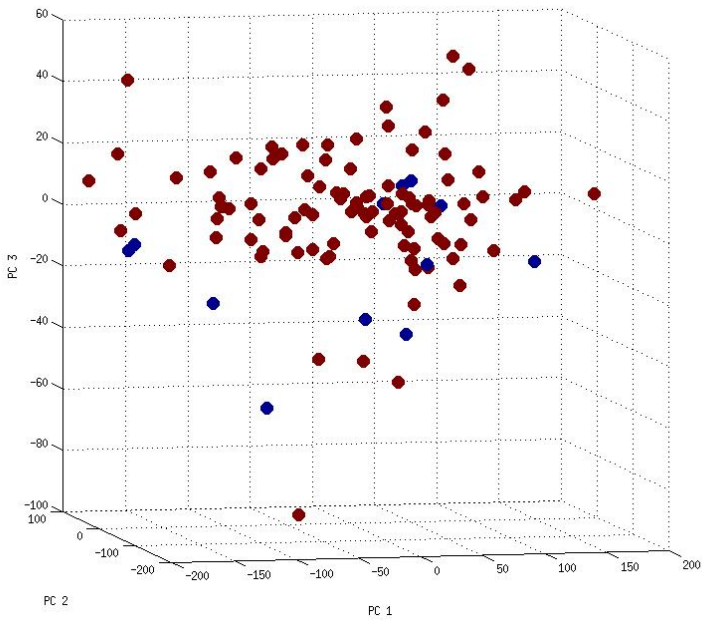

Figure E3: Three dimensional PCA plot for all the 112 sputum arrays. Red dots are asthma patients and blue dots are controls.

The principal component analysis was applied to the batch and RIN adjusted data and labelled the arrays based on the self-reported race (Figure E4). No significant population stratification effect was observed.

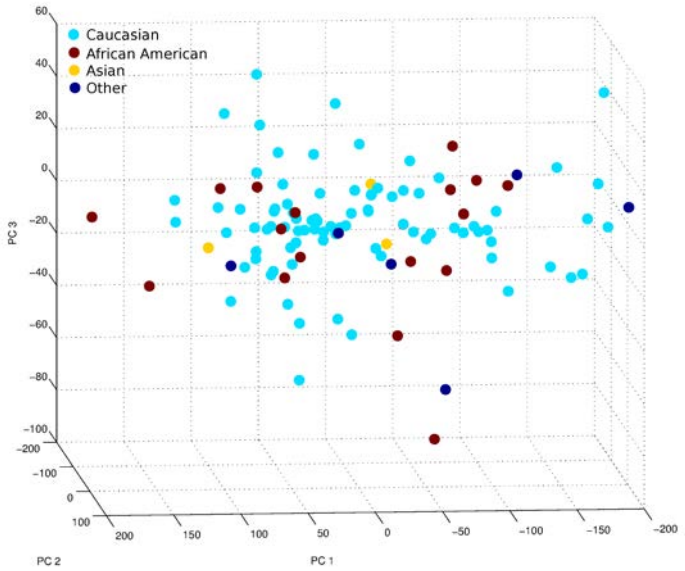

Figure E4: PCA plot for all the 112 sputum arrays labelled using the self-reported race.

Identifying the Sputum TEA Clusters

Pathway based distance between samples

To cluster the samples in an unbiased way, the distance between samples was calculated. One traditional way is to calculate the Euclidean distance between samples using genes with the largest variance. However, the Euclidean distance applies equal weights to all the genes regardless of the high variation in their expression levels. This may result in inaccurate measurement of the distance and errors in accurately identifying subtypes of asthma.

To overcome the methodological limitations of the traditional Euclidean distance, we developed a novel method to measure distances between samples that implicitly applies different weights by integrating genes in each pathway and calculating the distance by summarizing across different pathways. Known pathways were downloaded from MsigDB[E3,4] to provide list of member genes in each pathway. Using the expression levels of genes in each pathway, a mixture Gaussian model (MCLUST[E5]) is fitted to identify any clusters or subtypes in the samples. Suppose the i-th pathway identified k_i clusters and the cluster assignment of sample j is denoted as C_{ij}, where $C_{ij} = 1,2, \dots, k_i$. The pathway-based distance between sample l and m is then calculated as

$$d_{lm} = \frac{\#\{i: C_{il} \neq C_{im}\}c}{N_2},$$

where $\#\{i: C_{il} \neq C_{im}\}$ is the number of pathways that assign the two samples into different clusters and N_2 is the total number of pathways that identified at least two clusters. Visualization of both the distance matrix (Figure 1B) and the connectivity criteria[E6] (Figure E5) showed that there seems to be three clusters. To identify the actual clusters, K means was then applied with the number of clusters

set to 3. Compared to the clustering results by Euclidean distance, the pathway based distance matrix achieved lower internal criteria (data not shown).

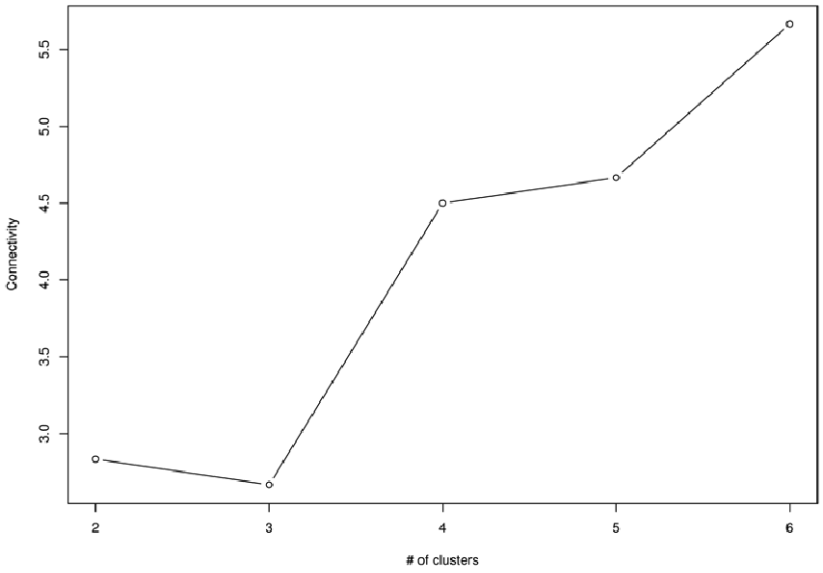

Figure E5: The connectivity of the clustering results versus different numbers of clusters.

Pathways driving each TEA cluster

For each of the KEGG pathway that identified more than one cluster using the Gaussian mixture model, a score was calculated to show how strong the given KEGG pathway was driving each TEA cluster. For a given TEA cluster i and a given pathway j, we first align the TEA clusters with the clusters identified by pathway j. Suppose cluster C_{ij} has the largest number of overlapping samples with TEA cluster i among all the clusters identified by pathway j. Then the score is defined as the percentage of samples in TEA cluster i that are not in cluster C_{ij} plus the percentage of samples in cluster C_{ij} that are not in TEA cluster i. All the KEGG pathways were ranked based on this score and the top 5 KEGG pathways for each of the TEA clusters are shown in Table E1.

Table E1: The top 5 KEGG pathways that drive each TEA cluster. The score that was used to rank the pathways are also shown.

TEA 1		TEA 2		TEA 3	
Pathways	***Score***	***Pathways***	***Score***	***Pathways***	***Score***
KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	0.021	KEGG_PROGESTERONE_MEDIATED_OOCYTE_MATURATION	0.128	KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	0.410
KEGG_PATHWAYS_IN_CANCER	0.043	KEGG_VIBRIO_CHOLERAE_INFECTION	0.128	KEGG_WNT_SIGNALING_PATHWAY	0.457
KEGG_ADHERENS_JUNCTION	0.060	KEGG_EPITHELIAL_CELL_SIGNALING_IN_HELICOBACTER_PYLORI_INFECTION	0.138	KEGG_INSULIN_SIGNALING_PATHWAY	0.500

KEGG_TASTE_TRANSDUCTION	0.063	KEGG_ALANINE_ASPARTATE_AND_GLUTAMATE_METABOLISM	0.150	KEGG_NEUROTROPHIN_SIGNALING_PATHWAY	0.513
KEGG_TIGHT_JUNCTION	0.063	KEGG_CHEMOKINE_SIGNALING_PATHWAY	0.150	KEGG_PROXIMAL_TUBULE_BICARBONATE_RECLAMATION	0.513

Pathways enriched for DEGs of each TEA cluster

We identified the differentially expressed genes (DEGs) between each TEA cluster and controls to better understand the underlying biological mechanisms of each TEA cluster. Functional analysis of these DEGs was performed using GeneGo MetaCore which assesses the significant of enrichment of a collection of pathways in the DEGs. The list of top 10 enriched pathways can be found in Table E2. Since there are no significant DEGs under FDR<0.2 for TEA2, this analysis was not be performed for TEA2.

Table E2. List of the top 10 enriched pathways for TEA1 and TEA3 by GeneGO MetaCore.

TEA 1		
Maps	**Total**	**pValue**
Histidine-glutamate-glutamine and proline metabolism/ Rodent version	44	0.0044
Colorectal cancer (general schema)	29	0.0095
Histamine metabolism	10	0.0097
Phospholipid metabolism p.2	12	0.0139
Leukotriene 4 biosynthesis and metabolism	14	0.0188
DNA damage_NHEJ mechanisms of DSBs repair	19	0.0336
Plasmalogen biosynthesis	19	0.0336
G-protein signaling_Cross-talk between Ras-family GTPases	20	0.0370
Cell adhesion_Gap junctions	22	0.0441
Huntington's disease (general schema)	3	0.0454
TEA 3		
Maps	**Total**	**pValue**
Protein folding and maturation_POMC processing	30	5.8E-25
Cytoskeleton remodeling_TGF, WNT and cytoskeletal remodeling	111	5.7E-08
Development_PIP3 signaling in cardiac myocytes	47	1.2E-06
Cytoskeleton remodeling_Cytoskeleton remodeling	102	1.4E-06

Signal transduction_AKT signaling	43	9.0E-05
Development_Thrombopoietin-regulated cell processes	45	1.2E-04
Development_IGF-1 receptor signaling	52	2.7E-04
Apoptosis and survival_Role of PKR in stress-induced apoptosis	53	2.9E-04
Cell cycle_Influence of Ras and Rho proteins on G1/S Transition	53	2.9E-04
Development_WNT signaling pathway. Part 2	53	2.9E-04

TH2 gene expression among the TEA clusters

Previous studies have identified the TH-2 inflammation high group of asthma patients based on the Th2 signature gene (IL-4, IL-5 and IL-13) expression levels in both sputum and blood. [E7-9] We evaluated the mean Th2 signature gene expression levels of the TEA clusters in the sputum (Figure E6). As can be seen in Figure E6, there is a significant difference in the Th2 gene expression level among the TEA clusters, with TEA 1 showing the highest Th2 gene expression level (p value=4.2e-8). This is consistent with TEA cluster 1 having the highest FeNO level, ICS requirement, and risk of intubations. This shows consistent clinical and biological characteristics among the TEA clusters.

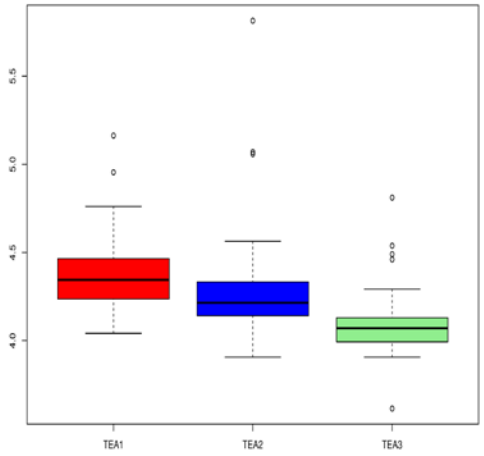

Figure E6. Boxplots of the mean TH2 gene expression in the three TEA clusters.

Sputum TEA classifier using blood gene expression

A logistic regression model was built to predict the TEA cluster of samples using blood gene expression data from validation cohort. Among the 100 samples included in the sputum TEA cluster analysis, 76 of them also have matched blood arrays, which were used as a training data set to build the logistic regression model. First, the sputum TEA cluster assignment of the 76 samples was obtained and the blood gene expression profiles of samples from each TEA cluster were compared to each other using the moderate Student's t test. Genes with a nominal p values<0.001 were considered to be significant. Second, we merged the genes differentially expressed between each pair of TEA clusters and compare this merged gene set to the 19,601 genes probed on both the Affymetrix HuGene1.0 ST array and the Illumina HumanHT12 BeadChips. About 301 overlapping

genes were overlapped. Finally, the data was standardized so that each gene has a mean of 0 and standard deviation of 1. Then a L1 regularized logistic regression model was fitted to the blood expression levels of these 301 genes which further selected 53 genes for the sputum TEA cluster classifying model using blood gene expression data. The list of these 53 genes and the top 10 enriched pathways by GeneGO MetaCore are shown in Table E3 and Table E4. The tuning parameter in LASSO was chosen using leave-one-out cross validation. The sputum TEA cluster classifier demonstrated an internal prediction accuracy of 89%.

Table E3. List of the 53 genes selected by LASSO for classifying patients with whole blood gene expression data.

transcript_cluster_id	Chromosome	Gene Symbol	Gene Name
7898594	chr1	HTR6	5-hydroxytryptamine (serotonin) receptor 6
7907396	chr1	C1orf10	chromosome 1 open reading frame 105
7911017	chr1	SDCCAG8	serologically defined colon cancer antigen 8
7915385	chr1	EDN2	endothelin 2
7923131	chr1	DENND1B	DENN/MADD domain containing 1B
7925728	chr1	OR6F1	olfactory receptor, family 6, subfamily F, member 1
7940114	chr11	OR1S1	olfactory receptor, family 1, subfamily S, member 1
7940989	chr11	ESRRA	estrogen-related receptor alpha
7948155	chr11	OR5AP2	olfactory receptor, family 5, subfamily AP, member 2
7949021	chr11	RCOR2	REST corepressor 2
7950321	chr11	UCP3	uncoupling protein 3 (mitochondrial, proton carrier)
7950332	chr11	C2CD3	C2 calcium-dependent domain containing 3
7950374	chr11	P4HA3	prolyl 4-hydroxylase, alpha polypeptide III
7951521	chr11	C11orf65	chromosome 11 open reading frame 65
7967337	chr12	ABCB9	ATP-binding cassette, sub-family B (MDR/TAP), member 9
7980213	chr14	PROX2	prospero homeobox 2
8000899	chr16	ZNF768	zinc finger protein 768
8000974	chr16	ZNF668	zinc finger protein 668
8007794	chr17	ARL17A	ADP-ribosylation factor-like 17A
8012028	chr17	ASGR2	asialoglycoprotein receptor 2
8024934	chr19	ZNRF4	zinc and ring finger 4
8028552	chr19	NFKBIB	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, beta
8030416	chr19	SCAF1	SR-related CTD-associated factor 1
8037283	chr19	PSG4	pregnancy specific beta-1-glycoprotein 4
8045208	chr2	POTEE	POTE ankyrin domain family, member E
8053341	chr2	REG3A	regenerating islet-derived 3 alpha
8061564	chr20	ID1	inhibitor of DNA binding 1, dominant negative helix-loop-helix protein
8062108	chr20	PROCR	protein C receptor, endothelial
8066411	chr20	C20orf62	chromosome 20 open reading frame 62
8076046	chr22	TMEM184B	transmembrane protein 184B
8076219	chr22	SNORD83B	small nucleolar RNA, C/D box 83B

8083605	chr3	RSRC1	arginine/serine-rich coiled-coil 1
8083794	chr3	MYNN	Myoneurin
8086048	chr3	TMPPE	transmembrane protein with metallophosphoesterase domain
8091537	chr3	IGSF10	immunoglobulin superfamily, member 10
8093852	chr4	MSX1	msh homeobox 1
8097626	chr4	LOC441046	glucuronidase, beta pseudogene
8100519	chr4	TXNDC9	thioredoxin domain containing 9
8104463	chr5	MARCH6	membrane-associated ring finger (C3HC4) 6
8105842	chr5	CENPH	centromere protein H
8114572	chr5	HBEGF	heparin-binding EGF-like growth factor
8115871	chr5	FLJ16171	FLJ16171 protein
8124588	chr6	GPX6	glutathione peroxidase 6 (olfactory)
8125748	chr6	LYPLA2P1	lysophospholipase II pseudogene 1
8136863	chr7	TMEM139	transmembrane protein 139
8140909	chr7	ERVWE1	endogenous retroviral family W, env(C7), member 1
8148655	chr8	TIGD5	tigger transposable element derived 5
8150920	chr8	CYP7A1	cytochrome P450, family 7, subfamily A, polypeptide 1
8150978	chr8	CA8	carbonic anhydrase VIII
8153043	chr8	ZFAT	zinc finger and AT hook domain containing
8155696	chr9	FAM122A	family with sequence similarity 122A
8166127	chrX	GLRA2	glycine receptor, alpha 2
8174648	chrX	CXorf61	chromosome X open reading frame 61

Table E4. The top 10 pathways enriched for the 53 gene profile by GeneGO MetaCore.

Maps	Total	Genes	pValue	Min FDR
Immune response_Lipoxins and Resolvin E1 inhibitory action on neutrophil functions	2/20	I-kB, NFKBIB	1.209E-03	1.050E-01
Impaired inhibitory action of lipoxins and Resolvin E1 on neutrophil functions in CF	2/23	I-kB, NFKBIB	1.603E-03	1.050E-01
Development_ERBB-family signaling	2/36	I-kB, HB-EGF	3.914E-03	1.343E-01
Immune response_HMGB1/RAGE signaling pathway	2/50	I-kB, NFKBIB	7.453E-03	1.343E-01
Apoptosis and survival_Role of PKR in stress-induced apoptosis	2/51	I-kB, NFKBIB	7.745E-03	1.343E-01
Development_Role of IL-8 in angiogenesis	2/53	I-kB, HB-EGF	8.346E-03	1.343E-01
Immune response_Role of PKR in stress-induced antiviral cell response	2/55	I-kB, NFKBIB	8.967E-03	1.343E-01
Immune response_Gastrin in inflammatory response	2/60	I-kB, HB-EGF	1.061E-02	1.343E-01

Development_EGFR signaling pathway	2/65	I-kB, HB-EGF	1.237E-02	1.343E-01
Development_Role of nicotinamide in G-CSF-induced granulopoiesis	1/8	ID1	2.111E-02	1.343E-01

Validation of clinical phenotypes associated with TEA clusters in a second cohort

Asthma BRIDGE Cohort

1,542 subjects from the EVE Consortium of asthma genome-wide association studies were invited to participate in Asthma BRIDGE by completing a detailed phenotype questionnaire and providing tissue samples for gene expression profiling, including a subset of 981 who provided whole blood samples. Gene expression profiling was conducted using HumanHT12 BeadChips (Illumina, San Diego CA), employing a randomized-array allocation strategy to minimize potential confounding by technical batch effects. Following filtering of poor quality samples, 870 high quality whole blood expression profiles were available for analysis. Quantlie normalization and log2 transformation were conducted with the *lumi* (R package) (see https://biolincc.nhlbi.nih.gov/home/).[E10] Illumina gene probes (Asthma BRIDGE cohort) were selected that matched Affymetrix HuGene 1.0 ST array probe sets (YCAAD cohort). Comparable information is available for 19,601 of the 22,148 probe sets on the Affymetrix HuGene 1.0 ST arrays.

Sputum TEA cluster identification

The log2 transformed gene expression levels of the 53 genes chosen for the sputum TEA classifier model were obtained and standardized for each sample in the Asthma BRIDGE cohort. The expression levels were imported into the sputum TEA cluster classifier to predict the sputum TEA cluster assignment for each sample in the Asthma BRIDGE cohort. The predicted probability of a sample belonging to a particular TEA cluster was compared among the three TEA clusters (Figure E6). The comparison showed that samples in each TEA cluster are significantly more likely to belong to that cluster compared to the samples assigned to the other clusters (p value<2.2e-16 for all three TEA clusters). This quantitatively demonstrates that the three TEA clusters are well separated from each other in the Asthma BRIDGE cohort.

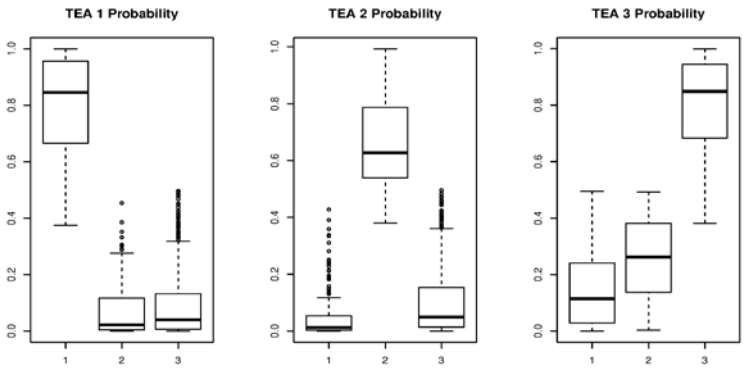

Figure E7. Boxplot of the predicted probability of samples belonging to each TEA cluster 1 (left panel), TEA cluster 2 (middle panel) and TEA cluster 3 (right panel), respectively. X-axis of the boxplots show the TEA cluster assignment of the samples.

REFERENCES

E1. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic acids research 2003;31:e15.
E2. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007;8:118-27.
E3. Mootha VK, Lindgren CM, Eriksson KF, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature genetics 2003;34:267-73.
E4. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 2005;102:15545-50.
E5. Fraley C, Raftery AE. MCLUST: Software for model-based cluster analysis. J Classif 1999;16:297-306.
E6. Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis. Bioinformatics 2005;21:3201-12.
E7. Peters MC, Mekonnen ZK, Yuan S, Bhakta NR, Woodruff PG, Fahy JV. Measures of gene expression in sputum cells can identify TH2-high and TH2-low subtypes of asthma. The Journal of allergy and clinical immunology 2014;133:388-94.
E8. Woodruff PG, Boushey HA, Dolganov GM, et al. Genome-wide profiling identifies epithelial cell genes associated with asthma and with treatment response to corticosteroids. Proceedings of the National Academy of Sciences of the United States of America 2007;104:15858-63.
E9. Woodruff PG, Modrek B, Choy DF, 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.
E10. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics 2008;24:1547-8.