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

**Xiting Yan1,2, Jen-Hwa Chu6, Jose Gomez1, Maria Koenigs1, Carole Holm1, Xiaoxuan He1, Mario F. Perez1, Hongyu Zhao2,3,4,5, Shrikant Mane5, Fernando D. Martinez7, Carole Ober8, Dan L. Nicolae8, Kathleen C. Barnes9, Stephanie J. London10, Frank Gilliland11, Scott T. Weiss6, Benjamin A. Raby6,Lauren Cohn1 and Geoffrey L. Chupp1**

1. Section of Pulmonary, Critical Care, and Sleep Medicine and Department of Internal Medicine, Yale University School of Medicine, New Haven, CT 06520, USA
2. Biostatistics Resource, Keck Laboratory, Yale University School of Medicine, New Haven, CT 06520, USA
3. Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, USA
4. Program in Computational Biology and Bioinformatics, Yale University School of Medicine, New Haven, CT 06520, USA
5. Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA
6. The Channing Division of Network Medicine, Department of Medicine, Brigham and Women’s Hospital, and Harvard Medical School, Boston, MA 02115, USA
7. Arizona Respiratory Center and BIO5 Institute, University of Arizona, Tucson, AZ 85724, USA
8. Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA
9. Department of Medicine, Johns Hopkins University, Baltimore, MD 21224, USA
10. National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, NC 27709, USA
11. Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA

**Corresponding Author**: Geoffrey L. Chupp

300 Cedar Street (S441 TAC), New Haven, CT 06520-8057

Phone (203) 785-3627 / Fax (203) 785-3826 / Email geoffrey.chupp@yale.edu

**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.C., F.D., C.O., D.L.N., K.C.B., S.J.L., F.G., S.T.W. and B.A.R. provided the Asthma BRIDGE cohort data. X.Y. and G.L.C wrote the manuscript.

**Sources of support:**

- NIH Training grants T32HL007778-18 and T15LM007056-26. R01HL095390-04.

- Jose Gomez is supported through a FAMRI Young Clinical Scientist Award.

- NIH/NHLBI grant K99HL114651-01A1

- K.C.B. was supported in part by the Mary Beryl Patch Turnbull Scholar Program

- X.Y. is supported in part by the NCATS grant UL1 TR000142

**Running head:** Characterizing Transcriptomic Endotypes of Asthma

**Descriptor number:** 1.6 Airway Remodeling: Asthma Genes and Gene Product

**Manuscript body word count:** 3271

**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 non-invasive analysis of sputum gene expression to identify transcriptomic endotypes of asthma (TEA Clusters) that correlate with clinical characteristics of severe disease including a history of hospitalization and near fatal asthma attack. The TEA 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 that among children and adults with asthma.

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

**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 asthma patients 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 for validation.

**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 pre-bronchodilator FEV1 (P = 0.006), a higher bronchodilator response (P = 0.03), and higher exhaled nitric oxide levels (P = 0.04), compared to 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 x 10-06) 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.

**Word count:** 250

**Keywords:** molecular endotyping, genomic, RNA, severe asthma, pathway analysis

**INTRODUCTION**

Asthma is a chronic inflammatory disease of the airways that will likely afflict over 10% of the U.S. population by the end of this decade.[1](#_ENREF_1) Differences in genetic susceptibility, environmental exposures and medication compliance are known to contribute to the heterogeneous clinical manifestations of disease.[2](#_ENREF_2),[3](#_ENREF_3) However, it is increasingly evident that pathobiologic 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](#_ENREF_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 asthmatics 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](#_ENREF_5),[6](#_ENREF_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](#_ENREF_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 (SARP) and the Childhood Asthma Management Program (CAMP) 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](#_ENREF_8),[9](#_ENREF_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](#_ENREF_10),[11](#_ENREF_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 3 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](#_ENREF_12)

**METHODS**

***Yale Center for Asthma and Airway Diseases (YCAAD) Cohort.*** A cross-sectional analysis was conducted on sputum RNA samples collected from asthmatic and control subjects that completed the Yale Center for Asthma and Airway Diseases (YCAAD) phenotyping protocol between September 2009 and June 2012. Subjects were > 12 years of age, non-smokers, and with < 10 pack years of smoking history. Inclusion criteria for asthma included a history, physical exam, and physiologic testing consistent with a diagnosis of asthma based on NAEPP 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). Exhaled nitric oxide (FeNO) was measured and spirometry was conducted in adherence with the American Thoracic Society guidelines before and after short-acting bronchodilator administration.[13](#_ENREF_13) Sputum induction was performed with hypertonic saline, as previously described.[14-17](#_ENREF_14) 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-Geimsa stain, respectively, and cell pellets were stored in All-in-One RNA stabilization buffer (Norgen Biotek, Thorhold, CAN).

***Genomic analysis.*** Sputum cell pellets were processed using the All-in-One purification kit (Norgen Biotek, Thorhold, CAN), checked on an Agilent bioanalyzer and, if needed, treated again to remove DNA contamination (Qiagen, Gaithersberg, MD). 10 ng of sputum RNA was amplified using the WT-Ovation Pico RNA amplification System (NuGen, San Carlos, CA) and processed per Affymetrix protocols, as previously described.[18](#_ENREF_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 1A. Raw microarray intensity data was 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 data supplement for details).[19](#_ENREF_19) K-means clustering was applied to assign samples into the selected 3 TEA clusters. Kruskal-Wallis tests were used to assess differences in continuous clinical, physiologic, and inflammatory asthma phenotypes between the clusters, and the Chi-squared or Cochran-Armitage test was used to assess differences in categorical phenotypes. False discovery rate was estimated using a permutation based method to adjust for the multiple testing error.[20](#_ENREF_20) Differentially expressed genes (DEGs) between each TEA cluster and controls were identified as genes with a false discovery rate (FDR)<0.05 using the Student’s t test.[21](#_ENREF_21) See online data supplement for more details on computational analyses.

 ***Validation Studies.*** A TEA cluster classifier was built using 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 data supplement for details).[22](#_ENREF_22),[23](#_ENREF_23) This classifier was applied to 870 whole blood samples selected from the Asthma BioRepository for Integrative Genomic Exploration (Asthma BRIDGE).[24](#_ENREF_24),[25](#_ENREF_25) Cross-platform replication of the TEA cluster clinical phenotypes was conducted.[24](#_ENREF_24),[25](#_ENREF_25)

**RESULTS**

***Identification of Sputum TEA clusters in YCAAD Cohort.*** The sputum expression levels of 5,500 genes from 186 KEGG pathways were used to assess the pathway-based distance between samples followed by unsupervised K means clustering to define sputum TEA clusters (Figure 1B). Three clusters were selected based on the connectivity criteria (Figure E5 in supplement). The “relatedness” of the samples within each cluster was evident on a sample distance matrix (Figures 1C). This demonstrated that samples within TEA cluster 3 are the most strongly related (darkest red) and most homogeneous, followed by TEA cluster 1, 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 1A). The subjects in TEA cluster 1 required a higher daily dose of inhaled steroids (mean daily ICS dose 617 g/day, P = 0.04) and were more likely to have a history of an intubation for asthma, compared to the other TEA clusters (18%, P = 0.05). In addition, TEA cluster 1 had the lowest pre- and post- bronchodilator FEV1 (pre-bronchodilator percent predicted FEV1, 73% ± 24, P < 0.01), more bronchodilator reversibility (12% ±12, P = 0.03), and elevated FeNO levels (mean 53 ± 43 ppb. P = 0.03), compared to the other TEA clusters (Table 1B). The fewest number of individuals (N=19) were in TEA cluster 2. Compared to the other clusters, TEA cluster 2 had the highest percentage of non-atopic subjects (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), compared to the other clusters. 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 pre- and post- bronchodilator lung function, minimal bronchodilator reversibility, and the lowest FeNO (mean, 38 ppb ± 27, P = 0.04) compared to the other TEA clusters (Table 1B). 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 or, sample processing, or treatment (Table 1C).

***Differentially Expressed Genes 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 to non-asthmatic controls to determine the genes associated with each of the TEA clusters. Using an FDR threshold of 0.05, there were 31 significantly differentially expressed genes (DEGs) in TEA cluster 1, 0 DEGs in TEA cluster 2 (15 DEGs had an FDR <0.25), and 27 DEGs in TEA cluster 3 compared to non-asthmatic control subjects (the top 10 most significant DEGs are shown in Table 3A and 3B). In TEA cluster 1, expression of L-histidine decarboxylase (HDC), an enzyme in the histamine metabolism pathway that converts histidine to histamine, was increased in asthmatic individuals compared to non-asthmatic control subjects.[26](#_ENREF_26) Two DEGs, EXOSC9 and SNAPC5, which code for proteins that are involved in RNA processing, were down-regulated compared to non-asthmatic control subjects.[27](#_ENREF_27) Three DEGs, NRCAM, PCLO, and SLC4A4, which are associated with neuron function, were significantly increased in TEA cluster 1 compared to non-asthmatics controls.[28-30](#_ENREF_28) In TEA cluster 3, all of the 27 DEGs are up-regulated compared to control subjects. These included DNAH17, a force generating dynein heavy chain motor protein in respiratory epithelium and defensin DEFB1), an antimicrobial peptide. Both genes have been previously associated with asthma and primary ciliary dyskinesia, respectively.[31-33](#_ENREF_31) 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 (Table E2 in the online data supplement).

***Validation of Sputum TEA clusters using blood gene expression.*** Since 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 was 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 in the online data supplement). 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 2A).

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 data supplement for details).[24](#_ENREF_24),[25](#_ENREF_25) Similar to the patterns observed in the YCAAD cohort, the PCA plot of the blood arrays from Asthma BRIDGE cohort in Figure 2B 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 x 10-16). This demonstrated that a blood signature was able to discriminate subgroups of children with asthma (ABRIDGE) with a similar prevalence 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 2). 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 x 10-06) and TEA cluster 2 subjects were the most likely to have a history of hospitalization for asthma (35%, P = 0.01), compared to 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 were the same as the YCAAD cohort. Although TEA cluster 1 in the ABRIGE cohort did not have lower lung function compared to the other TEA clusters (data not shown), 8% of these children had a history of near fatal asthma attacks.

**DISCUSSION**

Non-invasive 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 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 the 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.

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). While this shows that there is pathobiological overlap between the TEA clusters and Th2 biology, the link is relatively weak as allergic inflammation KEGG pathways was not a pathway that drove the clustering. This suggests that the TEA clusters are driven by biologic phenomena that are upstream or possibly parallel to Th2 inflammation.[34](#_ENREF_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 difference in age among the individuals. This cluster also has the lowest baseline lung function, the highest bronchodilator reversibility, the highest FeNO levels and the highest doses of ICS: all characteristics that are associated with near fatal asthma.[35](#_ENREF_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 this cluster.[26-30](#_ENREF_26)

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 to the other TEA clusters (color heterogeneity seen in Figure 1C). These individuals also have severe disease and are more likely to have been hospitalized for asthma. While the evaluation of larger cohorts will be 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 due to the small number of individuals in this cluster), the transcriptomic discrimination of this cluster suggests that there are distinct 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 has the highest levels of YKL-40 in the sputum among the clusters (P value = 0.03, data not shown), findings consistent with a possible YKL-40 endo/phenotype with increased risk of exacerbations and abnormal post-bronchodilator FEV1.[36](#_ENREF_36),[37](#_ENREF_37)

Individuals in TEA cluster 3 have clinical features most consistent with mild disease including increased expression of defensin $β1$(DEFB1), a gene that has been associated with mild asthma in multiple studies.[31](#_ENREF_31) Compared to the other TEA clusters, these individuals have preserved lung function, the lowest inhaled corticosteroid dose, the lowest FeNO 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 1C). 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 2B), but is distinct from the other TEA clusters in adults (Figure 2A). 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 that are typically used to study severe asthma (i.e. BMI, atopy, FEV1). This is due, in part, to the clustering approach we developed that is distinctly different compared to conventional approaches that use clinical features, and large differences in gene expression compared to non-asthmatics to select genes for the clustering analysis.[5](#_ENREF_5),[9-11](#_ENREF_9) In contrast, the analytical approach used herein was not biased by 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 asthmatics and controls - 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. Since 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 asthmatics 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.

**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 YCAAD database.

**REFERENCES**

1. Moorman JE, Zahran H, Truman BI, Molla MT, Centers for Disease C, Prevention. Current asthma prevalence - United States, 2006-2008. Morbidity and mortality weekly report Surveillance summaries 2011;60 Suppl:84-6.

2. Wenzel SE. Asthma phenotypes: the evolution from clinical to molecular approaches. Nature medicine 2012;18:716-25.

3. Proceedings of the ATS workshop on refractory asthma: current understanding, recommendations, and unanswered questions. American Thoracic Society. American journal of respiratory and critical care medicine 2000;162:2341-51.

4. Moffatt MF, Gut IG, Demenais F, et al. A large-scale, consortium-based genomewide association study of asthma. The New England journal of medicine 2010;363:1211-21.

5. 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.

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. The Journal of allergy and clinical immunology 2014;133:388-94.

7. Moore WC, Bleecker ER, Curran-Everett D, et al. Characterization of the severe asthma phenotype by the National Heart, Lung, and Blood Institute's Severe Asthma Research Program. The Journal of allergy and clinical immunology 2007;119:405-13.

8. Howrylak JA, Fuhlbrigge AL, Strunk RC, Zeiger RS, Weiss ST, Raby BA. Classification of childhood asthma phenotypes and long-term clinical responses to inhaled anti-inflammatory medications. J Allergy Clin Immun 2014;In press.

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

10. Haldar P, Pavord ID, Shaw DE, et al. Cluster analysis and clinical asthma phenotypes. American journal of respiratory and critical care medicine 2008;178:218-24.

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

12. Geoffrey LC, Mario FP, Jose GG-A, et al. Determination Of Asthma Phenotypes By Unsupervised Transcriptomic Cluster Analysis Of Induced Sputum. B14 ASTHMA PHENOTYPING: UNDERSTANDING OF PATHOGENESIS AND RATIONALE FOR NEW THERAPIES: American Thoracic Society; 2013:A2328-A.

13. Standardization of Spirometry, 1994 Update. American Thoracic Society. American journal of respiratory and critical care medicine 1995;152:1107-36.

14. Covar RA, Spahn JD, Martin RJ, et al. Safety and application of induced sputum analysis in childhood asthma. The Journal of allergy and clinical immunology 2004;114:575-82.

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;50:1033-7.

16. Grootendorst DC, van den Bos JW, Romeijn JJ, et al. Induced sputum in adolescents with severe stable asthma. Safety and the relationship of cell counts and eosinophil cationic protein to clinical severity. The European respiratory journal 1999;13:647-53.

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;57:875-9.

18. Kang HJ, Kawasawa YI, Cheng F, et al. Spatio-temporal transcriptome of the human brain. Nature 2011;478:483-9.

19. Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis. Bioinformatics 2005;21:3201-12.

20. Millstein J, Volfson D. Computationally efficient permutation-based confidence interval estimation for tail-area FDR. Frontiers in genetics 2013;4:179.

21. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met 1995;57:289-300.

22. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Ann Stat 2004;32:407-51.

23. Tibshirani R. Regression shrinkage and selection via the Lasso. J Roy Stat Soc B Met 1996;58:267-88.

24. Raby BA BK, Beaty TH, Bosco A, Carey VJ, Castro M, Cheadle C, Gilliland FD, Islam KTS, Salam MT, Kelly R, Klanderman BJ, Lemanske RL, Guilbert TW, London SJ, Liu A, Martinez F, Mathias RA, Nicolae D, Ober C, Qiu W, Solway J, Strunk R, Sylvia J, Tubbs D, Weiss ST. Asthma BRIDGE: The Asthma BioRepository for Integrative Genomic Exploration. Am J Respir Crit Care Med 183: A6189 2010;183:A6189.

25. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics 2008;24:1547-8.

26. Koarai A, Ichinose M, Ishigaki-Suzuki S, et al. Disruption of L-histidine decarboxylase reduces airway eosinophilia but not hyperresponsiveness. American journal of respiratory and critical care medicine 2003;167:758-63.

27. Mistry DS, Chen YF, Sen GL. Progenitor Function in Self-Renewing Human Epidermis Is Maintained by the Exosome. Cell Stem Cell 2012;11:127-35.

28. Aoki T, Matsumoto Y, Hirata K, et al. Expression profiling of genes related to asthma exacerbations. Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology 2009;39:213-21.

29. Laprise C. The Saguenay-Lac-Saint-Jean asthma familial collection: the genetics of asthma in a young founder population. Genes and immunity 2014.

30. Dorfman R, Li W, Sun L, et al. Modifier gene study of meconium ileus in cystic fibrosis: statistical considerations and gene mapping results. Human genetics 2009;126:763-78.

31. Levy H, Raby BA, Lake S, et al. Association of defensin beta-1 gene polymorphisms with asthma. The Journal of allergy and clinical immunology 2005;115:252-8.

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. Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology 2006;36:649-57.

33. Chodhari R, Mitchison HM, Meeks M. Cilia, primary ciliary dyskinesia and molecular genetics. Paediatric respiratory reviews 2004;5:69-76.

34. Expert Panel Report 3 (EPR-3): Guidelines for the Diagnosis and Management of Asthma-Summary Report 2007. The Journal of allergy and clinical immunology 2007;120:S94-138.

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. American journal of respiratory and critical care medicine 1997;156:737-43.

36. Chupp GL, Lee CG, Jarjour N, et al. A chitinase-like protein in the lung and circulation of patients with severe asthma. The New England journal of medicine 2007;357:2016-27.

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. Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology 2011;106:381-6.

**FIGURE LEGENDS**

**Figure 1. Identification of TEA clusters.** A). Diagram showing an overview of the computational analysis flow. B). Heatmap showing the clustering results by KEGG pathways using MCLUST. The color represents the clustering assignment of each sample by the KEGG pathways. C). 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 longer distance showing the strength of the clusters (samples are weakly related). Samples within TEA cluster 3 are the most strongly related and most homogeneous, followed by cluster 1 and 2, respectively.

**Figure 2. Data visualization of the TEA clusters using the 53 blood expression.** A). PCA plot of the 76 matched blood arrays in YCAAD cohort. B). PCA plot of the blood arrays from Asthma BRIDGE cohort. TEA cluster assignment was predicted using the TEA cluster classifier built in YCAAD cohort.

**Table 1A: Phenotypic Characteristics of TEA Clusters in the YCAAD Cohort**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Controls (N=12) | Cluster 1 | Cluster 2 | Cluster 3 | P Value |
| Prevalence (N) |  | 34 | 19 | 47 | 0.003 |
| Age at Visit (years) | 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/m2) | 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 (years) | 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 per day (µg)**  | **NA** | **617** ± **448** | **530** ± **449** | **396** ± **356** | **0.04** |
| **ICS use yes or no – N (%)** | **NA** | **27 (79)** | **17 (89)** | **31 (66)** | **0.10** |
| Chronic OCS use (%) | NA | 4 (11.8) | 2 (10.5) | 3 (6.4) | 0.68 |

Means ± Standard Deviation are shown. P values for comparisons among TEA clusters were determined using Kruskal-Wallis or Cochran-Armitage test. BMI=Body Mass Index, ICS=Inhaled Corticosteroids, OCS=oral corticosteroids, FEV1 = forced expiratory volume in 1 second, FVC=forced vital capacity. NA = not applicable. The false discovery rate estimated by the permutation based method is 11% for Table 1A and 5.6% for Table 1B.

**Table 1B: Pulmonary Function of TEA Clusters in YCAAD Cohort**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Controls (N=12) | Cluster 1 (N=34) | Cluster 2 (N=19) | Cluster3 (N=47) | P Value |
| **FEV1- % of predicted value** |  |  |  |  |  |
| **Pre β2 agonist use** | **96 ± 11** | **73 ± 24** | **76 ± 22** | **86 ± 22** | **0.006** |
| Post β2 agonist use | 98 ± 14 | 80 ± 24 | 81 ± 20 | 91 ± 22 | 0.493 |
| FVC- % of predicted value |  |  |  |  |  |
| Pre β2 agonist use | 86 ± 22 | 85 ± 22 | 86 ± 18 | 96 ± 19 | 0.09 |
| Post β2 agonist use | 91 ± 20 | 90±20 | 88 ± 18 | 97 ± 18 | 0.28 |
| FEV1/FVC- % of predicted value |  |  |  |  |  |
| Pre β2 agonist use | 0.77 ± 0.62 | 0.67 ± 0.13 | 0.70 ± 0.11 | 0.72 ± 0.10 | 0.13 |
| Post β2 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** |
| **FENO (ppb)** | **20 ± 9.7** | **53 ± 43** | **52 ± 42** | **38 ± 27** | **0.04** |

**Table 1C: Sputum Characteristics of TEA Clusters in the YCAAD Cohort**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Controls(N=12) | Cluster 1 (N=34) | Cluster 2 (N=19) | Cluster 3 (N=47) | P Value |
| Mucus Cell Concentration | 40.86±20.98Ρ | 83.02±105.75Ρ | 89.23±143.61Ρ | 73.72±62.48Ρ | 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 |

 Ρ Cells/Microliter x 104

**Table 2: 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%** | **<2.2x10-16** |
| **Age at Visit (years)** | **14.8 ± 8** | **10.1 ± 5** | **12.6 ± 7** | **3.27 x 10-08** |
| Sex, N (%) Female | 128 (48) | 45 (43) | 212 (43) | 0.31 |
| **Race** |  |  |  | **1.68 x 10-07** |
| 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 x 10-06** |
| **ACT Score** | **14 ± 4** | **12 ± 3** | **13 ± 3** | **8.79 x 10-07** |

**Table 3A: Top 10 differentially expressed genes between TEA cluster 1 and controls**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Gene Name | Gene Symbol | Biological Processes | Function | Fold Change | Pvalue | FDR | PubmedReferences |
| hemogen | HEMGN | apoptosis | Transcription Factor | 1.25 | 2.79x10-7 | 4.01x10-3 | 16 |
| piccolo (presynaptic cytomatrix protein) | PCLO | cytoskeleton organizationregulation of exocytosis | Protein | 1.21 | 6.26x10-7 | 4.01x10-3 | 26 |
| cartilage associated protein | CRTAP | extracellular matrix organization | Scaffolding Protein | -1.48 | 7.87x10-7 | 4.01x10-3 | 25 |
| exosome component 9 | EXOSC9 | RNA processing | RNase Complex Component | -2.10 | 1.11x10-6 | 4.01x10-3 | 32 |
| chromosome 9 open reading frame 173 | C9orf173 | N/A | N/A | 1.36 | 1.23x10-6 | 4.01x10-3 | 2 |
| small nuclear RNA activating complex, polypeptide 5, 19kDa | SNAPC5 | DNA-dependent regulation of transcription | Transcription Factor | -1.08 | 1.23x10-6 | 4.01x10-3 | 9 |
| impact RWD domain protein | IMPACT | negative regulation of protein phosphorylation | Enzyme | -2.12 | 1.44x10-6 | 4.01x10-3 | 11 |
| solute carrier family 4, sodium bicarbonate cotransporter, member 4 | SLC4A4 | sodium ion transport | Membrane Transporter Protein | 1.22 | 1.45x10-6 | 4.01x10-3 | 61 |
| histidine decarboxylase | HDC | histamine biosynthesis | Enzyme | 1.39 | 1.94x10-6 | 4.78x10-3 | 48 |
| neuronal cell adhesion molecule | NRCAM | neuron migration | Cell Adhesion Molecules | 1.19 | 2.29x10-6 | 5.08x10-3 | 40 |

**Table 3B: Top 10 differentially expressed genes between TEA cluster 3 and controls**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Gene Name | Gene Symbol | Biological Processes | Functions | Fold Change | Pvalue | FDR | PubmedReferences |
| heart development protein with EGF-like domains 1 | HEG1 | endothelial cell angilgenesis | Cell adhesion molecules | 1.78 | 1.09x10-7 | 2.42x10-3 | 7 |
| small nucleolar RNA, C/D box 104 | SNORD104 | RNA modification | non-coding RNA | 2.58 | 2.46x10-7 | 2.73x10-3 | 2 |
| dynein, axonemal, heavy chain 17 | DNAH17 | ciliary motility | microtubule-associated motor protein | 1.88 | 2.61x10-6 | 1.18x10-2 | 7 |
| Cbl proto-oncogene, E3 ubiquitin protein ligase B | CBLB | regulation of T cell anergy | ubiquitin protein ligase | 2.04 | 2.95x10-6 | 1.18x10-2 | 93 |
| defensin, beta 1 | DEFB1 | Innate Immune Response | antimicrobial peptide | 2.85 | 2.97x10-6 | 1.18x10-2 | 133 |
| non-protein coding RNA 204 | NCRNA204 | N/A | non-coding RNA | 2.20 | 3.55x10-6 | 1.18x10-2 | 1 |
| transcription elongation factor A (SII) N-terminal and central domain containing | TCEANC | RNA elongation | transcription elongation factor | 2.22 | 6.58x10-6 | 1.82x10-2 | 10 |
| radical S-adenosyl methionine domain containing 2 | RSAD2 | regulation of toll-like receptor 9 signaling pathway | antiviral protein | 1.57 | 1.04x10-5 | 2.07x10-2 | 28 |
| purinergic receptor P2Y, G-protein coupled, 14 | P2RY14 | regulation of inflammation | UDP-glucose receptor | 5.14 | 1.04x10-5 | 2.07x10-2 | 31 |
| malignant fibrous histiocytoma amplified sequence | MFHAS1 | cell cycle | Protein | 4.14 | 1.11x10-5 | 2.07x10-2 | 11 |