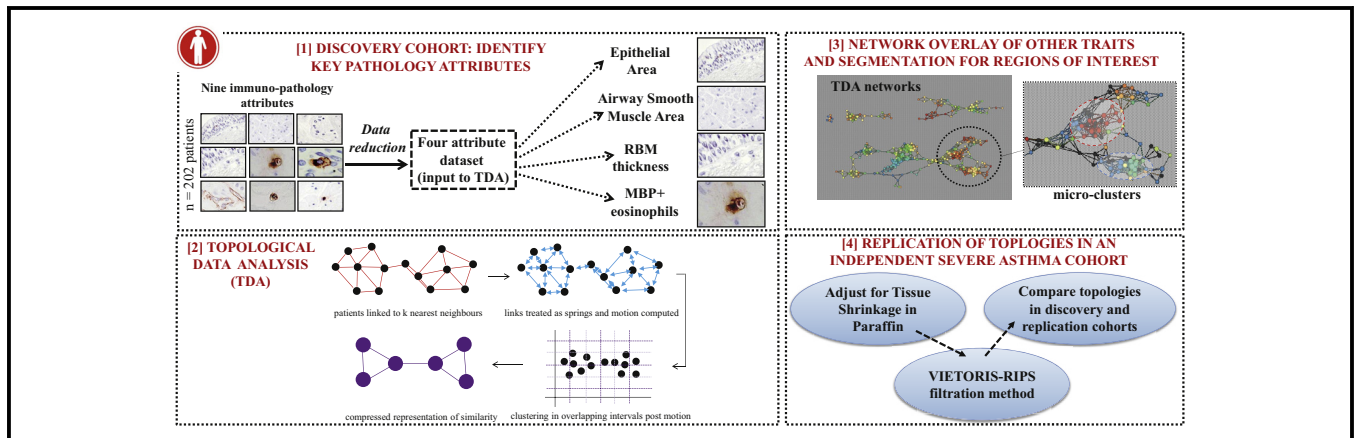


Airway pathological heterogeneity in asthma: Visualization of disease microclusters using topological data analysis

Salman Siddiqui, MRCP, PhD,^a Aarti Shikotra, BSc, PhD,^a Matthew Richardson, BSc, PhD,^a Emma Doran, PhD,^b David Choy, BSc,^b Alex Bell, MMath,^{a,d} Cary D. Austin, MD, PhD,^b Jeffrey Eastham-Anderson, MS,^b Beverley Hargadon, RGN,^a Joseph R. Arron, MD, PhD,^b Andrew Wardlaw, FRCP, PhD,^a Christopher E. Brightling, FRCP, PhD,^a Liam G. Heaney, FRCP, MD,^c and Peter Bradding, FRCP, DM^a *Leicester and Belfast, United Kingdom, and South San Francisco, Calif*

GRAPHICAL ABSTRACT



Background: Asthma is a complex chronic disease underpinned by pathological changes within the airway wall. How variations in structural airway pathology and cellular inflammation contribute to the expression and severity of asthma are poorly understood. **Objectives:** Therefore we evaluated pathological heterogeneity using topological data analysis (TDA) with the aim of visualizing disease clusters and microclusters.

Methods: A discovery population of 202 adult patients (142 asthmatic patients and 60 healthy subjects) and an external replication population (59 patients with severe asthma) were evaluated. Pathology and gene expression were examined in bronchial biopsy samples. TDA was applied by using pathological variables alone to create pathology-driven visual networks.

From ^athe Department of Infection Immunity and Inflammation, Institute for Lung Health, University of Leicester, Glenfield Hospital, Leicester; ^bGenentech, South San Francisco; ^cthe Centre for Infection and Immunity, Health Sciences Building, Queens University Belfast; and ^dthe Department of Mathematics, University of Leicester.

Supported by the National Institute for Health Research (NIHR) Leicester Respiratory Biomedical Research Unit and project grants (to S.S., P.B., J.R.A., and L.G.H.) from Asthma UK (AUK-PG-2013-208 [project title: Endotyping asthma for the benefit of patients]) and from the Jules Thorne Trust Clinical Senior Lecturer Award ("Systems medicine: novel mathematical approaches to personalised care in asthma patients"; to S.S.). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health. Additional funding was received from Genentech (to P.B.) to support this study. Biopsy samples and additional funding from the GlaxoSmithKline-sponsored studies RES100767 ([clinicaltrials.gov](https://clinicaltrials.gov/ct2/show/study?term=RES100767): NCT00331058) and RES100769 ([clinicaltrials.gov](https://clinicaltrials.gov/ct2/show/study?term=RES100769): NCT00327197) were provided by GlaxoSmithKline (to A.J.W.) to support this study. In addition, biopsy samples were provided by C.E.B. to support this study from studies funded by an MRC intermediate and a Wellcome Trust senior fellowship award.

Disclosure of potential conflict of interest: S. Siddiqui declares grants from the Jules Thorne Trust Grant for Clinical Senior Lecturers, an Asthma UK Project Grant, a grant from the NIHR Biomedical Research Centre and consultancy with AstraZeneca, GlaxoSmithKline, Owlstone, Mundipharma, and Boehringer Ingelheim. D. Choy is employed by Genentech, a member of the Roche Group; holds patents (planned, pending or issued) with Genentech, a member of the Roche Group; and has stock/stock options with Roche. C. D. Austin is employed by, has stock/stock options with, and has received travel/accommodation/meeting expenses unrelated to activities listed from Genentech. J. R. Arron is employed by Genentech and has stock/stock options with

Roche Holdings. A. Wardlaw has received a grant from GlaxoSmithKline; holds board membership with TEVA; has been a consultant for GlaxoSmithKline, Pulmocide, Boehringer, and KNOPP; and has received grants from/has grants pending with Pfizer. C. E. Brightling has received grants from the Wellcome Trust, Asthma UK, and the MRC; has consultant arrangements with GlaxoSmithKline, AstraZeneca/MedImmune, Chiesi, Novartis, Roche/Genentech, Theravance, Sanofi/Regeneron, Vectura, and Prep; and has grants from/grants pending with Roche/Genentech, Novartis, GlaxoSmithKline, AstraZeneca/MedImmune, Mologic, and Pfizer. L. G. Heaney has received grants from Northern Ireland Chest, Heart & Stroke Association and Genentech; has consultant arrangements with Hoffman la Roche, AstraZeneca, Novartis, GlaxoSmithKline, and TEVA; has received grants from/has grants pending with Novartis UK, Roche/Genentech, MedImmune, and GlaxoSmithKline; has received payment for lectures, including service on speakers' bureaus, from Hoffman la Roche, Novartis, and TEVA; and has received travel/accommodation/meeting expenses unrelated to activities listed from AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, TEVA, Chiesi, and Napp. P. Bradding has received a grant from Genentech and has Consultant arrangements with Roche. The rest of the authors declare that they have no relevant conflicts of interest.

Received for publication April 28, 2017; revised November 16, 2017; accepted for publication December 4, 2017.

Corresponding author: Salman Siddiqui, MRCP, PhD, Institute for Lung Health, Respiratory Biomedical Research Unit, University Hospitals of Leicester, Leicester LE3 9QP, United Kingdom. E-mail: ss338@le.ac.uk.

0091-6749/\$36.00

Crown Copyright © 2018 Published by Elsevier Inc. on behalf of the American Academy of Allergy, Asthma & Immunology. All rights reserved

<https://doi.org/10.1016/j.jaci.2017.12.982>

Results: In the discovery cohort TDA identified 4 groups/networks with multiple microclusters/regions of interest that were masked by group-level statistics. Specifically, TDA group 1 consisted of a high proportion of healthy subjects, with a microcluster representing a topological continuum connecting healthy subjects to patients with mild-to-moderate asthma. Three additional TDA groups with moderate-to-severe asthma (Airway Smooth Muscle^{High}, Reticular Basement Membrane^{High}, and Remodeling^{Low} groups) were identified and contained numerous microclusters with varying pathological and clinical features. Mutually exclusive T_H2 and T_H17 tissue gene expression signatures were identified in all pathological groups. Discovery and external replication applied to the severe asthma subgroup identified only highly similar “pathological data shapes” through analyses of persistent homology.

Conclusions: We have identified and replicated novel pathological phenotypes of asthma using TDA. Our methodology is applicable to other complex chronic diseases. (J Allergy Clin Immunol 2018;■■■:■■■-■■■.)

Key words: Severe asthma, remodeling, airway inflammation, topological data analysis, phenotyping

Asthma is a complex chronic disease that causes considerable morbidity¹ and is often suboptimally controlled with current therapies.² It is characterized by the presence of airway inflammation, which can be associated with the presence or absence of type 2 inflammatory cytokines.³ In addition to inflammation, airway remodeling⁴ is a common feature of asthma and encompasses a variety of structural changes, including reticular basement membrane thickening, increased airway smooth muscle (ASM) mass, increased glandular mass, goblet cell hyperplasia,⁵ and vascular remodeling.⁶

The use of endobronchial biopsies to investigate the pathology of asthmatic airways has greatly improved our understanding of asthma pathophysiology.^{4,7} However, studies to date describing airway pathology in asthmatic patients have been limited by (1) relatively small sample sizes, (2) variability in the pathological features described, and (3) variability in the techniques used. In particular, there has been no comprehensive evaluation of airway pathology in asthma using a single set of analytical features with a view to identifying pathological subtypes of asthma. Such a study is important because emerging asthma therapies can target specific features of remodeling, cellular inflammation, or both⁸⁻¹⁰ and their associated molecular pathways.

We hypothesized that (1) patients with asthma comprise distinct pathological phenotypes and (2) the phenotypes observed in asthma could be replicated by using mathematical techniques suitable for replication of high-dimensional topologies. We used unbiased statistical approaches combining data topology and visualization using topological data analysis (TDA) to generate pathological phenotypes¹¹⁻¹³ of patients with asthma across the spectrum of treatment intensity and healthy age-matched subjects. An independent severe asthma external replication cohort was used to evaluate topology replication. Specifically, we selected TDA as an analytic tool because of the ability to interrogate complex data in high-dimensional (>3 dimensions) space and create visualizations (networks) of the data. Unlike classical approaches, such as some forms of cluster analysis, TDA does not require any *a priori* specification of the number of clusters and has the added advantage of enhanced visualization of putative phenotypes and microscopic clusters identified as connected nodes of individual patients sharing

Abbreviations used

ASM: Airway smooth muscle
BOBCAT: Biomarkers in Corticosteroid-refractory Asthma
GINA: Global Initiative for Asthma
GMA: Glycol methacrylate
IQR: Interquartile range
TDA: Topological data analysis

similar pathological attributes. Additionally, with classical clustering methods, data points that can be naturally connected can be broken apart, but TDA preserves natural connections and reflects the “gradual” separation of points in a real data set. If well-separated groups exist within a data set, then TDA will identify these groups but not at the expense of introducing artificial separation between data points.

METHODS

Patient populations

We evaluated 2 independent populations (see Fig E1 in this article's Online Repository at www.jacionline.org): (1) a discovery population (discovery cohort) of asthmatic patients across the spectrum of asthma treatment intensity and age-matched healthy volunteers and (2) an independent replication population (replication cohort) of patients with severe asthma at Global Initiative for Asthma (GINA) treatment steps 4 and 5.¹ The replication cohort was evaluated and compared with the GINA step 4 and 5 population in the discovery cohort.

Discovery cohort. Two hundred two patients (142 asthmatic patients and 60 healthy volunteers) were recruited prospectively at Glenfield Hospital, Leicester, United Kingdom, and Queen's University, Belfast, United Kingdom. Some pathological features presented here have been reported previously from subcohorts of patients.^{6,14-20}

Replication cohort. Bronchial biopsy specimens were evaluated in 59 patients from the multicenter Bronchoscopic Exploratory Research Study of Biomarkers in Corticosteroid-refractory Asthma (BOBCAT) study,²¹ a 3-visit observational study of patients with uncontrolled severe asthma taking high-dose inhaled or oral corticosteroids (GINA treatment steps 4-5).

Asthmatic/healthy volunteers: definition

All patients with asthma were current nonsmokers, with a 10 pack year or less smoking history, and met the following diagnostic criteria: an expert physician's diagnosis of asthma with objective evidence (defined as ≥ 1 of the following: peak flow variation $\geq 20\%$ over a 2-week period, bronchodilator reversibility $\geq 12\%$ and 200 mL, or airway hyperresponsiveness [methacholine PC₂₀ ≤ 8 mg/mL]). Asthma treatment intensity was used as a marker of disease severity and defined according to GINA treatment steps with the following severity strata (GINA 1, $n = 30$ [mild], GINA 2-3, $n = 31$ [moderate], and GINA 4-5 [severe], $n = 81$).¹ Healthy volunteers in the discovery cohort ($n = 60$) were identified from local advertising and had no prior history of respiratory disease.

Ethical approvals

All studies were approved by locally relevant ethics committees and institutional review boards. A detailed statement on ethical approvals is provided in the Methods section in this article's Online Repository at www.jacionline.org.

Patient characterization

Discovery cohort. Patients underwent extensive evaluation at baseline, including a full medical history and lung function testing with bronchial challenge using methacholine where appropriate and sputum inflammometry. Tables of data completeness for clinical metadata are reported in Table E1 in this article's Online Repository at www.jacionline.org.

Replication cohort. Characterization of patients in the BOBCAT cohort has been described previously, and results have been compared with those in the discovery severe asthma cohort (GINA treatment steps 4-5).²¹

Further details of patient characterization and bronchoscopy in both cohorts are provided in the [Methods](#) section in this article's Online Repository.

Immunohistochemistry

Discovery cohort. Two-micrometer glycol methacrylate (GMA) sections were cut and immunostained, as described previously.²² The following mouse primary antibodies were used: anti-neutrophil elastase clone NP57 (0.1 µg/mL; Dako, Ely, United Kingdom), anti-mast cell tryptase clone AA1 (0.1 µg/mL; Dako), anti-eosinophil major basic protein clone BMK-13 (0.4 µg/mL; Monosan, Uden, The Netherlands), anti-EN4 (5 µg/mL; Monosan), anti- α -smooth muscle actin (1A4, 0.7 µg/mL; Dako), and appropriate isotype controls (Dako).

Replication cohort. Tissues were fixed with 10% formalin and embedded in paraffin wax before cutting. Tissue sections (5 µm) were immunostained for a specific eosinophil marker (eosinophil major basic protein) by using the alkaline phosphatase technique. Immunoreactivity was detected with Fast Red and visualized under light microscopy.

Assessment of pathology using immunohistochemical staining

Discovery cohort. The following 9 previously validated pathological features were evaluated in the discovery cohort (see [Fig E2](#) in this article's Online Repository at www.jacionline.org). For features 1 to 3, percentages of epithelial, smooth muscle, and glandular areas in sections were identified¹⁶ and measured by using a computer analysis system (Cell F; Olympus, Hamburg, Germany). The mean of 2 sections at least 10 µm apart was taken for each analysis. For features 4 to 7, numbers of positively stained nucleated eosinophils, neutrophils, and mast cells in the lamina propria and mast cells within the ASM bundle areas were counted and expressed per square millimeter. ASM areas of less than 0.1 mm² were considered insufficient to quantify mast cell numbers.^{7,15} For feature 8, reticular basement membrane thickness was measured by using the technique reported by Sullivan et al.²³ In addition, for feature 9, the mean Chalkley count, a validated unbiased measure of vascular size and density, was evaluated in a subset of patients (n = 87 asthmatic and 32 healthy volunteers), as previously reported by us.⁶

Tables of data completeness for pathological features are reported in [Tables E1, B-D](#).

All pathological data were assessed by observers blinded to the identity of the patient. A single blinded observer, A.S. at Leicester, performed cross-counts in a random selection of asthmatic and healthy cases between F.S. and S.S. at Leicester (within-center repeatability) and E.D. at Belfast (between-center repeatability). Good to excellent between- and within-center repeatability was observed for all of the pathological features (intraclass correlation coefficient > 0.75). Further details relating to intraobserver/interobserver repeatability of pathology measurements are provided in the [Methods](#) and [Table E2](#) in this article's Online Repository at www.jacionline.org.

Replication cohort. J.E. and C.A. (Genentech) performed immunohistochemistry analysis with the help of J.R., who identified regions of interest for analysis and quantification. Because of the known effect of paraffin embedding, causing tissue shrinkage by approximately 30% in length and 50% in area when compared with GMA,²⁴ all biopsy measures in paraffin were normalized to account for this effect. We assumed that tissue shrinkage was isotropic and uniform within biopsy specimens. Further details on normalization for paraffin-embedded tissue shrinkage are provided in the [Methods](#) section in this article's Online Repository.

Lung function decrease cohort

Discovery cohort. A subgroup of asthmatic patients (67/142; see [Fig E1](#)) in the discovery cohort had longitudinal lung function (prebronchodilator FEV₁) data acquired after bronchial biopsy sampling. The median follow-up

duration was 2.73 years (Q1-Q3, 0.77-5.34 years). This population was used to evaluate the effect of remodeling phenotype on lung function decrease over time, although this substudy was not powered to detect interaction of TDA clusters and time.

Biopsy gene expression cohort

Discovery and replication cohorts. For a subset of biopsy specimens (see [Fig E1](#)), gene expression data were available from bronchial biopsy homogenates. RNA was isolated from homogenized bronchial biopsy specimens, as described previously.¹⁰ RNA was amplified (Ambion, Life Technologies, Grand Island, NY) for Agilent (Santa Clara, Calif) 2-color Whole Human Genome 4x44k gene expression microarray analysis. The gene expression profile of T_H2 or T_H17 asthma was defined, as previously reported.^{10,25,26} Additional analyses of genome-wide expression were performed in the severe asthma fraction (GINA treatment steps 4-5) of both cohorts and are outlined in the [Methods](#) section in this article's Online Repository.

Statistical analysis

Basic summary statistical analysis was performed with R 3.3.1 software (R Development Core Team) and SAS software (version 9.4; SAS Institute, Cary, NC). Parametric and nonparametric data are presented as means (SEMs) and medians (interquartile ranges [IQRs]), respectively, unless otherwise stated.

For comparison of multiple groups characterized by parametric and nonparametric variables, 1-way ANOVA or the Kruskal-Wallis test was used, and χ^2 /Fisher exact tests were used for proportions.

A 3-step statistical process was used and incorporated to generate data on pathological phenotypes in asthmatic patients. Step 1 was defined as factor analysis to reduce the dimensionality of 9 pathological features (leading to identification of 4 canonical pathological features highlighted with asterisks in [Fig E2](#)). Step 2 was defined as TDA with the input function derived from pathological factor analysis (described methodologically in [Fig E3](#) in this article's Online Repository at www.jacionline.org). Step 3 was defined as mapping of variables not used to generate TDA networks to the networks, such as FEV₁ decrease (by using linear mixed effects models; see [Table E3](#) in this article's Online Repository at www.jacionline.org) and biopsy gene expression (eg, T_H2 and T_H17 gene expression scores).

Microclusters were defined as regions within networks of patients (nodes) that were similar (connected by lines) and to avoid overinterpretation of the data represented visible substructures within networks, such as loops, tails, and large regions within networks separated by significant distances. Microclusters were used to understand within-network pathological heterogeneity within the discovery cohort.

Further details of these analyses are provided in the [Methods](#) section in this article's Online Repository.

Investigation of topological similarity in the discovery and replication severe asthma cohorts

An investigation of the topological equivalence of the severe asthma fraction (GINA treatment steps 4-5) within the discovery and replication data set was undertaken by using a specific analysis of persistence homology that evaluates the similarity between data shapes generated by using high-dimensional topology²⁷⁻³⁰ with JavaPlex³¹ in MATLAB R2016b.

Further details are provided in the [Methods](#) section and [Fig E4](#) in this article's Online Repository at www.jacionline.org.

RESULTS

Discovery cohort: Airway pathology and asthma severity

Clinical characteristics of the study population stratified by GINA treatment intensity are presented in [Tables E4, A and B](#), and pathological biopsy measurements stratified by GINA treatment

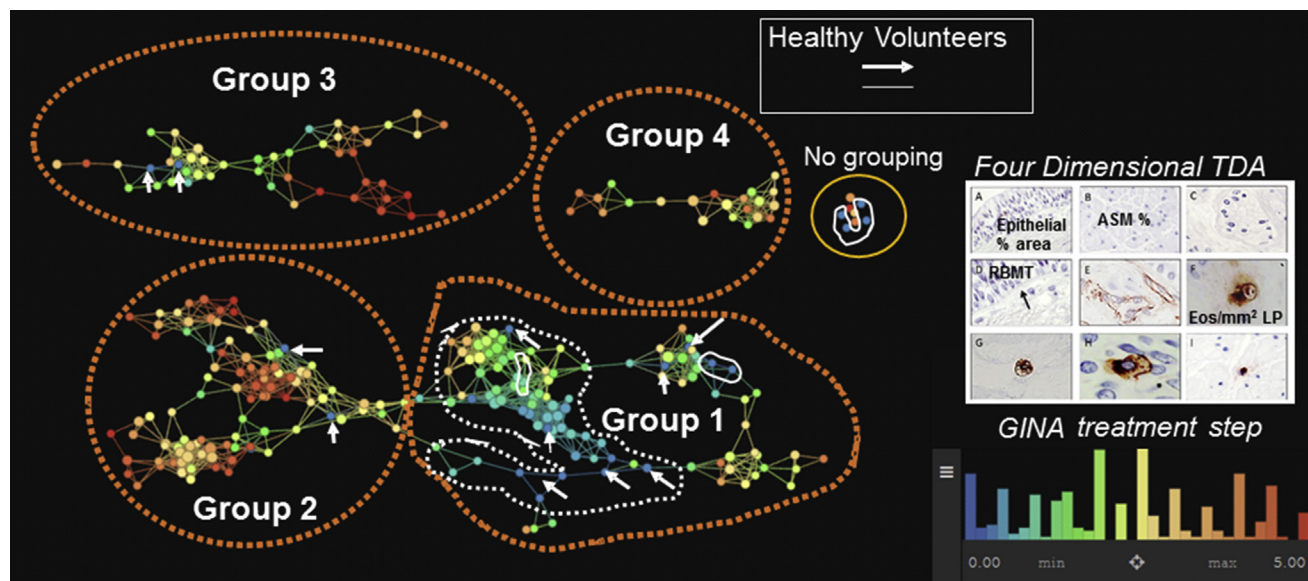


FIG 1. Discovery TDA pathology networks annotated by GINA treatment intensity. TDA networks were generated by using airway pathology markers in the discovery population of asthmatic and healthy volunteers. Networks have been colored by GINA treatment intensity (*bottom right-hand color key*) with GINA step 0 assigned to healthy volunteers to allow discrimination of asthmatic patients and nonasthmatic subjects, as well as asthma severity, both within and between networks. Four pathological traits derived after data reduction, as annotated in the *right-hand photomicrograph inset* (see Fig E2), were used to generate pathologically driven TDA networks. There are 4 overall groups of interest, with 1 additional set of cases who do not fall into any grouping. Patients with severe asthma fall primarily within groups 2 to 4. Healthy volunteers and patients with mild asthma (GINA steps 1-2) appear to be located primarily in group 1. The *white dotted loop* within group 1 identifies a group of patients who are healthy/have mild-to-moderate asthma without significant airway hyperresponsiveness and with no history of exacerbations that appear to represent a continuum between health and asthma. Additional healthy volunteers appear to be occasionally located within asthmatic networks (groups 2-4; *white arrows*) and in some cases are not associated with a core grouping (*orange circle*).

intensity are shown in [Tables E5](#) and [E6](#) (90th percentile analysis) and [Fig E5](#) in this article's Online Repository at www.jacionline.org.

Biopsy gene expression analysis and asthma treatment intensity

A total of 11 differentially expressed genes demonstrating increased (*FKBP5*, *KRT24*, *SLC25A16*, and *PHACTR3*) or reduced (*PHACTR1*, *RPL23AP32*, *YME1L1*, *ROBO3*, *TMC6*, and *IGLJ3*) expression between patients with severe asthma and healthy control subjects (after accounting for false discovery) were replicated in the BOBCAT severe asthma cohort (see [Fig E6](#) and [Table E7](#) in this article's Online Repository at www.jacionline.org). Based on these observations, T_H2 and T_H17 gene expression scores¹⁰ and a selection of the 11 replicated genes (see [Fig E6](#); *FKBP5*,³²⁻³⁴ *PHACTR3*,³⁵⁻³⁷ and *KRT24*³⁸) were used to subsequently annotate TDA networks and colocalize gene expression with various remodeling and clinical features.

Discovery cohort: Identification of pathological subtypes/phenotypes of asthma by using TDA

[Fig E3](#) demonstrates a conceptual and mathematical overview outlining the process used by TDA to generate pathological networks. [Fig 1](#) demonstrates TDA networks of airway pathology generated in a 4-dimensional data space. The networks have

been annotated by using GINA treatment intensity scores, with GINA step 0 assigned to healthy volunteers. The images represent 2-dimensional networks of airway pathology in asthmatic patients and healthy volunteers. Each network is a series of nodes (round circle) that represent small clusters of patients who share similar pathology and are connected by lines to other nodes in a network.

A total of 190 of 202 patients were identified within 4 discrete TDA groups ([Fig 1](#) and [Table I](#)). However, 10 of 202 did not demonstrate any clear grouping by using TDA ([Fig 1](#)).

Healthy volunteers ([Fig 1](#), white arrows and white boundary nodes) appeared to be primarily located in the group 1 network and within nodes that did not form any particular TDA grouping. In contrast, topological groups 2 to 4 were primarily composed of patients with moderate-to-severe asthma, although healthy volunteers were present within these groups (white arrows, [Table I](#)). Higher-resolution images of the TDA groups are reported in [Figs 2-5](#) with summaries for both clinical and immunopathologic data in the groups presented in [Tables I and II](#) and of discovery TDA groups in [Fig 6](#).

Discovery: High-resolution analysis groups 1 to 4

Group 1: Mild-to-Moderate Asthma/Healthy^{Continuum}

Group 1 ([Fig 2](#)) was characterized by patients with asthma ($n = 41/77$) and healthy volunteers ($n = 36/77$). Asthmatic patients within group 1 had a median GINA treatment intensity of

TABLE I. Clinical characteristics of the discovery populations stratified by TDA grouping

	Mild-to-Moderate Asthma/Healthy ^{Continuum} , group 1* (n = 77)	Reticular Basement Membrane ^{High} , group 2 (n = 60)	Airway Smooth Muscle ^{High} , group 3 (n = 36)	Remodeling ^{Low} , group 4 (n = 19)	P value†
Healthy subjects/group	36/77	8/60	6/36	6/19	<.0001‡
Age (y), mean ± SEM	39.49 ± 1.65	40.27 ± 1.82	44.58 ± 2.72	45.53 ± 3.01	.1575
Sex (% male)	55.56	58.33	48.48	42.11	.5994‡
BMI (kg/m ²)	27.25 (8.15)	26.30 (8.30)	27.55 (5.40)	27.09 (7.89)	.7850
GINA treatment steps	1.00 (3.00) ¶	4.00 (3.00)	3.00 (3.00)	4.00 (4.00)	<.0001
Age of onset (y) ± SEM*	28.11 ± 3.11	22.20 ± 2.64	29.41 ± 4.42	25.00 ± 6.04	.4447
Disease duration (y) ± SEM*	14.13 ± 2.40	19.19 ± 1.84	20.70 ± 2.72	18.62 ± 4.73	.0747
Atopic status (%)	55.56	77.36	64.52	70.00	.0186‡
Positive <i>Aspergillus</i> species skin test (%)	6.12	17.31	27.78	20.00	.0205‡
Total IgE (kU)	66.45 (129.35)	182.50 (262.90)	185.00 (425.00)	128.00 (210.20)	.0077
Smoking (pack years)	0.00 (0.00)	0.00 (1.00)	0.00 (5.00)	0.00 (1.00)	.5969
Annual exacerbation frequency*	0.00 (2.00)	1.00 (4.00)	1.50 (2.00)	2.00 (1.00)	.0575
BDP equivalent ICS dose/24 h*	500.00 (1000.00) ¶	1600.00 (1200.00)	1120.00 (1050.00)	900.00 (800.00)	.0102
Oral steroids (%)	12.82	23.08	25.00	30.77	.1816‡
Sputum eosinophils (%)	0.63 (4.25)	4.10 (16.75)	3.13 (11.45)	0.50 (2.25)	.0005
Blood eosinophils × 10 ⁹	0.19 (0.22)	0.25 (0.33)	0.27 (0.28)	0.27 (0.28)	.0714
FENO, 50 mL · s ⁻¹ (ppb)	17.60 (18.00)	38.95 (35.50)¶	17.00 (27.90)	15.00 (10.20)	<.0001
Sputum neutrophils (%)	54.75 (51.76)	61.73 (45.65)	41.10 (51.65)	74.00 (42.35)	.1469
FEV ₁ pre-BD (L) ± SEM	4.40 ± 1.17	2.95 ± 0.13	2.64 ± 0.17	2.83 ± 0.15	.0209
FEV ₁ pre-BD (%) ± SEM	90.70 ± 3.04	87.64 ± 2.82	83.42 ± 3.73	92.32 ± 3.96	.1221
FEV ₁ post-BD (L) ± SEM	4.67 ± 1.32	3.18 ± 0.11	2.85 ± 0.17	3.03 ± 0.15	.0894
FEV ₁ post-BD (%) ± SEM	94.93 ± 2.77	93.98 ± 2.82	89.10 ± 3.21	91.40 ± 4.38	.3063
FEV ₁ /FVC (%) post-BD ± SEM	78.16 ± 1.46	76.56 ± 1.35	73.53 ± 2.06	75.09 ± 3.07	.1300
PC ₂₀ methacholine (mg/mL)	15.73 (31.40) ¶	1.00 (7.90)	1.40 (3.40)	3.30 (31.50)	.0043
FEV ₁ BDR (%)	1.00 (6.00)	3.00 (8.00)	2.59 (10.71)	5.00 (7.37)	.4931

Values in boldface indicate statistical significance.

BD, Bronchodilator; BDP, beclomethasone dipropionate; BDR, bronchodilator responsiveness; FENO, fractional exhaled nitric oxide; FVC, forced vital capacity; ICS, inhaled corticosteroid.

*Group 1 patients include asthmatic and healthy volunteers and as a consequence between-group comparisons for asthma traits, only apply to the asthma population within group 1 (with the exception of GINA treatment step, which was set at 0 for healthy volunteers). All values not presented as means ± SEMs are presented as median (IQRs).

†Statistical analysis across all groups. Adjusted *P* values (*P* < .05) were denoted as follows: §versus 1; ||versus 2; ¶versus 3; and #versus 4. Adjusted *P* values were obtained by using Bonferroni correction. All tests are Kruskal-Wallis, unless indicated otherwise.

‡χ² test.

3 (Q1-Q3, 1-4), with 19 of 41 asthmatic patients demonstrating mild (GINA treatment step 1-2) disease. This population comprised the loop population and comprised a continuum with the healthy volunteers in group 1. The loop population did not have significant asthmatic-range airway hyperresponsiveness and did not experience any asthma exacerbations in the previous calendar year. Gene expression overlay identified low tissue expression of the DNA methylation gene *PHACTR3*³⁵⁻³⁷ in the loop population. In contrast, GINA step 4 to 5 asthmatic patients within group 1 either occupied the apex of the loop population or were a distinct population toward the right-hand side of the group 1 network.

Interestingly, a small microcluster within group 1 (MC1) was enriched markedly for patients with high submucosal glands. This patient population were T_H17 high and T_H2 low at the biopsy gene expression level and demonstrated a low postbronchodilator FEV₁ percentage and FEV₁/forced vital capacity ratio.

Group 2: Reticular Basement Membrane^{High} group.

Group 2 was characterized primarily by patients with moderate-to-severe asthma (n = 52/60), with 23% of patients receiving maintenance oral steroids. Pathologically, subjects in group 2 had reticular basement membrane thickening, with the topology demonstrating 2 nested subgroups with high and low epithelial percentage areas within the upper and lower poles of the network,

respectively (Fig 3). Two microclusters (Fig 3) within the topological data space of group 2 were identified. Microcluster 2 consisted of patients with primarily adult-onset severe asthma who had high tissue T_H17 gene expression, epithelial percentage area, cytokeratin 24 (KRT24) expression,³⁸ and tissue neutrophil counts. In contrast, microcluster 3 consisted of patients with early-onset moderate-to-severe asthma who demonstrated high sputum and tissue eosinophil counts with evidence of intermediate T_H2-high and low T_H17 gene expression.

Group 3: Airway Smooth Muscle^{High} group. Group 3 was primarily characterized by patients with moderate-to-severe asthma, with 25% of patients receiving maintenance oral corticosteroids. The most striking feature of group 3 was high airway percentage smooth muscle area and absolute area, with few other pathological features of structural remodeling. Overall, group 3 patients had significantly higher levels of airway hyperresponsiveness when compared with group 1 (median methacholine PC₂₀, 1.4 mg/mL [IQR, 3.4 mg/mL] vs 15.7 mg/mL [IQR, 31.4 mg/mL]; *P* = .0043; Table I). Group 3 subjects also demonstrated the lowest numeric prebronchodilator and postbronchodilator lung function of all the TDA groups. A microcluster of patients within the group 3 topology (microcluster 4) demonstrated the highest levels of mast cells in the lamina propria (and smooth muscle) and had concurrent tissue eosinophilia and

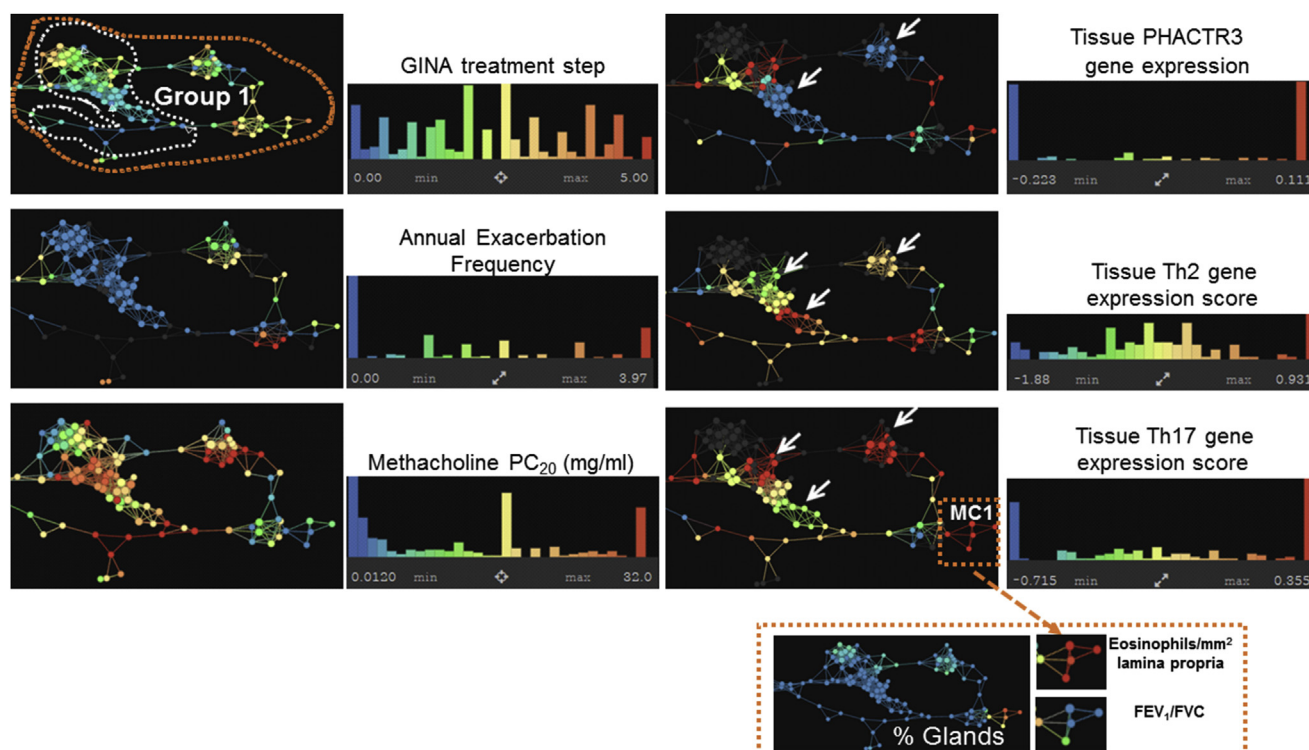


FIG 2. Mild-to-Moderate Asthma/Healthy^{Continuum} TDA network. Higher-magnification image of the group 1 TDA network (top left image). Color annotation keys are provided on the right-hand side of each corresponding TDA image. Gray nodes within the network indicate missing data. BASE and APEX refer to the bottom and top of the loop structure (white dotted outline). It can be seen that asthmatic patients from group 1 within the loop structure overall have low exacerbation frequency and low-grade airway hyperresponsiveness, with a gradual transition from T_H2 (base of the loop) to T_H17 (apex of the loop) tissue gene expression as corticosteroid exposure increases. The DNA hypermethylation marker of the *PHACTR3* gene was found to have low levels of expression in patients with mild asthma and healthy volunteers (top right image). The T_H2-high canonical gene demonstrated a reciprocal relationship with T_H17 gene expression (bottom right images) within tissue, suggesting that these phenotypes are mutually exclusive (white arrows). The dotted orange box (bottom right) indicates a microcluster (MC1) characterized by high submucosal percentage glands (this feature was not seen in any other TDA groups), a high T_H17 tissue gene expression score and low postbronchodilator FEV₁ percentage and FEV₁/forced vital capacity (FVC) ratio.

basement membrane remodeling with associated poor lung function and high exacerbations (Fig 4). Data from the severe asthma discovery and replication cohorts indicated that this population had a variable T_H2 and T_H17 tissue inflammatory phenotype (see Fig E9 in this article's Online Repository at www.jacionline.org).

Group 4: Remodeling^{Low} group. Group 4 was primarily characterized by patients with moderate-to-severe asthma (GINA step 4-5: 10/19), although 6 of 19 patients were healthy volunteers without a diagnosis of asthma. Asthmatic patients within this group had the highest numeric exacerbation frequency (median, 2 [IQR, 1.0] exacerbations/y; $P = .06$, comparison across asthmatic patients within TDA groups), despite having well-controlled eosinophilic inflammation in blood, sputum, and tissue (Tables I and II). Overall, group 4 had few features of airway remodeling.

However, a microcluster of patients was identified within this network, demonstrating a high epithelial percentage area, a non-T_H2 and variable high/low T_H17 tissue gene expression score, high tissue DNA hypermethylation marker *PHACTR3* gene expression,³⁵⁻³⁷ and a high frequency of exacerbations. This

group appeared to be enriched markedly for sputum neutrophilia (Fig 5).

Our observations of microclusters in all 4 discovery pathology phenotypes were masked group level statistics (Tables I and II) and are summarized in Fig 6.

Reciprocal tissue gene expression patterns

Overall, T_H2 and T_H17 gene expression in biopsy tissue was mutually exclusive across all TDA groups. In addition, the corticosteroid sensitivity gene *FKBP5*³²⁻³⁴ demonstrated reciprocal tissue expression when compared with *PHACTR3* across networks (Figs 2-5 and see Fig E6).³⁵⁻³⁷

Discovery: Prebronchodilator FEV₁ change analysis

Median decrease slopes in milliliters per year were centered at zero, with wide Q1-Q3 ranges across the TDA groups; see Table E3, B, and Fig E7 in this article's Online Repository at www.jacionline.org.

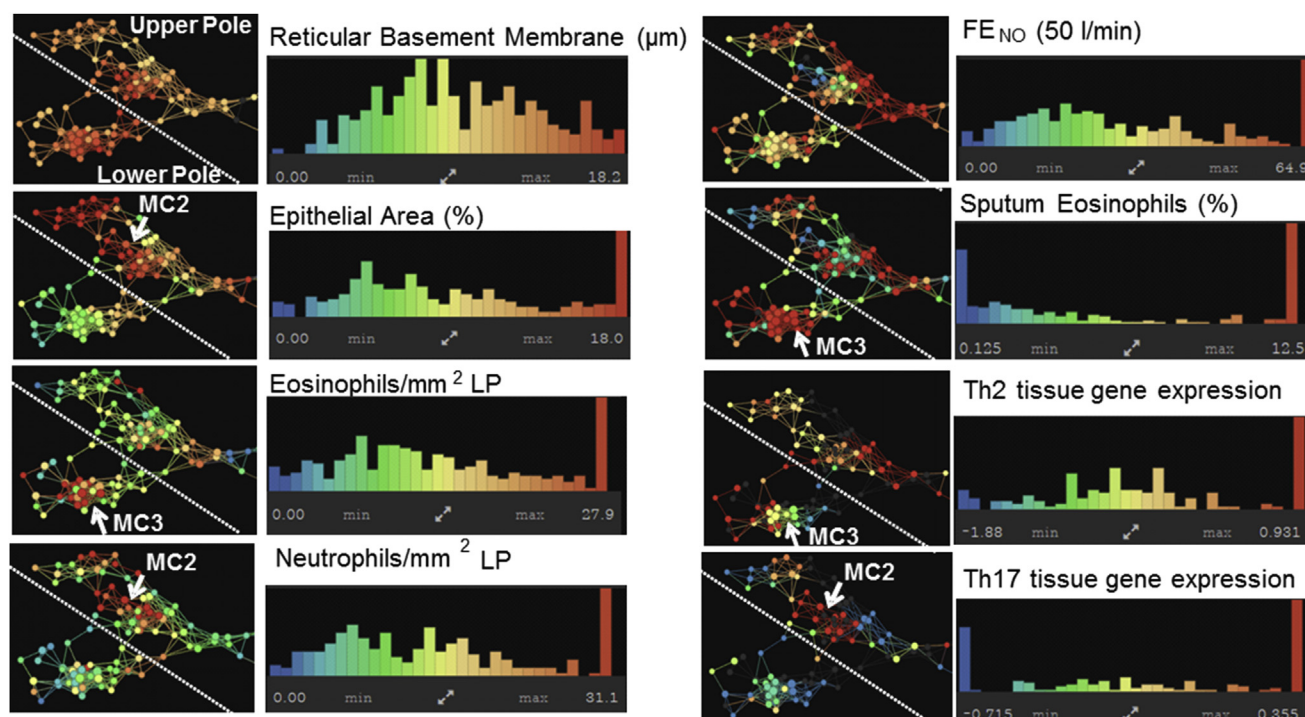


FIG 3. Reticular Basement Membrane^{High} TDA network. Higher-magnification image of the group 2 TDA network. The network is broadly divided into an upper pole and lower pole (correspondingly *above* and *below* the dotted diagonal line). Color annotations for fraction of exhaled nitric oxide (FE_{NO}), Sputum eosinophil values (as a percentage), and biopsy T_H2 and T_H17 gene expression scores are provided on the right-hand side of each corresponding TDA image. Gray nodes indicate missing data. It can be seen that basement membrane thickening is present in both the upper and lower poles of the network; however, the epithelial area is high in the upper pole and low in the lower pole. Arrows delineate regions of interest that are subsequently identified as microclusters 2 and 3 (MC2 and MC3, respectively). MC2 is characterized by adult-onset asthma, high tissue neutrophil counts, and a high epithelial percentage area and T_H17 gene expression score. MC3 is characterized by early-onset atopic refractory eosinophilic asthma, with high tissue and sputum eosinophil counts, a suppressed fraction of exhaled nitric oxide value, and intermediate T_H2 gene expression score. In keeping with previous reports, we identified pockets of T_H2 -high canonical gene expression with associated low T_H17 gene expression and *vice versa* (white arrows, bottom right images).

Replication of airway pathology phenotypes: Comparison of patients with severe asthma in the discovery and replication cohorts

Patients in the discovery and replication cohorts with severe asthma were matched for age and sex but differed with respect to the proportion receiving oral steroids (GINA step 5), bronchodilator responsiveness, and baseline spirometry (see Table E8, A-C, in this article's Online Repository at www.jacionline.org).

The results of computing persistent homology for the discovery and replication data sets are shown in barcode plots in Fig E8 in this article's Online Repository at www.jacionline.org. Both the discovery and replication populations demonstrated highly similar level 0 and level 1 homologies indicative of homology replication after accounting for 30% tissue shrinkage because of paraffin embedding. In contrast, simple visual inspection of networks (see Fig E9) using TDA did not identify topological similarities.

DISCUSSION

We have used novel methods to explore and visualize airway pathological heterogeneity in asthma and health. Specifically,

high-dimensional pathological objects and networks were generated by using TDA and identified 3 pathological phenotypes corresponding to visualized disease networks: (1) Mild-to-Moderate Asthma/Healthy^{Continuum}, (2) Reticular Basement Membrane^{High}, (3) Airway Smooth Muscle^{High}, and (4) Remodeling^{Low}.

We selected TDA as an analytical tool because of the ability to interrogate complex data in high-dimensional space based on the field of topology/shape analysis of data, machine learning, and data visualization.¹¹⁻¹³ Unlike classical approaches, such as cluster analysis, this tool does not require any *a priori* specification of the number of clusters and has the added advantage of enhanced visualization of putative phenotypes.

TDA group 1 (Mild-to-Moderate Asthma/Healthy^{Continuum}) comprised both patients without asthma and those with mild asthma, forming a distinct topological loop in the data space with a smooth transition of airway hyperresponsiveness from the nonresponsive range to low-grade hyperresponsiveness, few features of airway remodeling, and no exacerbations in the previous calendar year, which is indicative of a low risk phenotype. Nested within this pathological group, we identified a small microcluster of patients with T_H17 -high asthma with a high

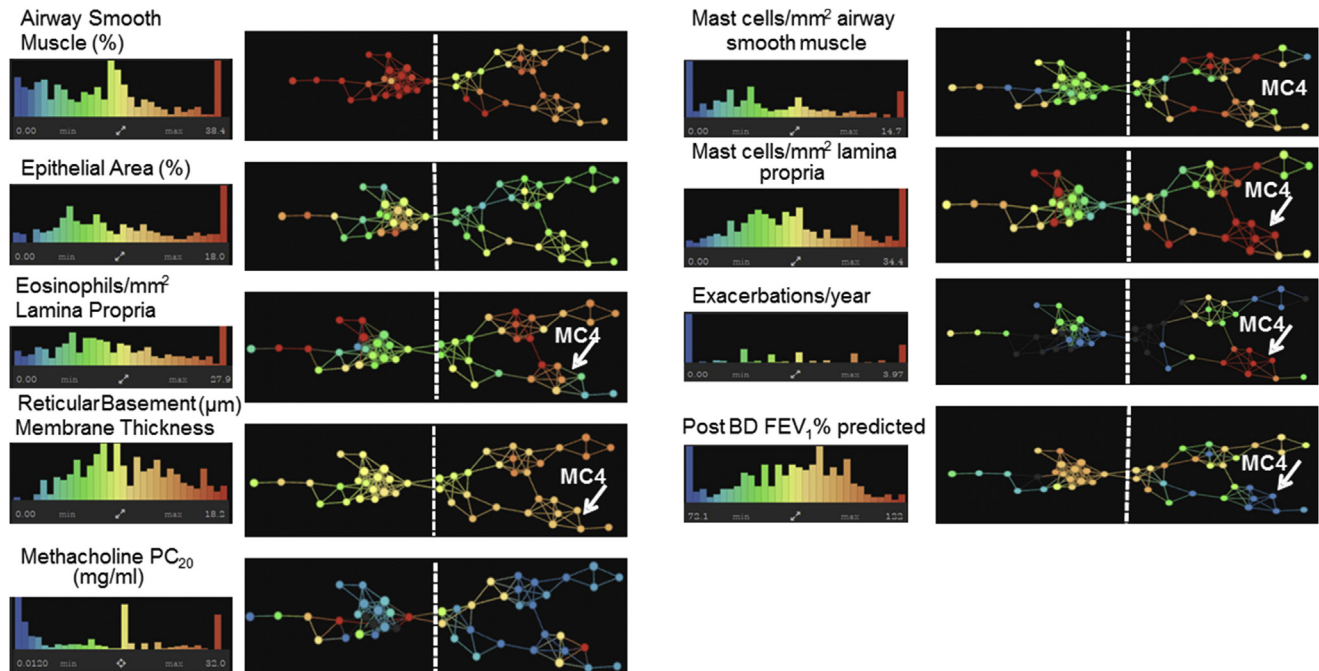


FIG 4. Airway Smooth Muscle^{High} and nested microcluster: group 3 TDA network. The network is characterized by patients with a high ASM percentage area (*bottom left image*) and airway hyperresponsiveness/low PC₂₀ (*top right image*) and demonstrated the lowest spirometric lung function (*bottom right image*). Patients with high tissue eosinophil and mast cell counts within both the ASM and lamina propria (LP) and basement membrane thickness are collocated to a region of the network on the *right-hand side of the dotted line*. Microcluster 4 (MC4; delineated by *white arrows*) of these patients is enriched for mast cells and eosinophils in the LP and appear to have low post bronchodilator lung function and a high annual exacerbation frequency. An adjacent collection of nodes, as well as selection of nodes within MC4, also demonstrated enrichment for mast cells in the ASM. We speculate that the group 3 overall population might be most amenable to therapies that potentially target the ASM in asthmatic patients (eg, bronchial thermoplasty), with those patients on the right of the *dotted line* being potentially amenable to both therapies that target airways smooth muscle and selectively target mast cells. T_H2/T_H17 status was highly variable in patients with severe asthma within this subgroup (see *Fig E9, A and B*).

percentage of submucosal glands and the lowest postbronchodilator lung function within the group 1 TDA network.

Similarly, in the moderate-to-severe asthma–predominant TDA groups (groups 2–4), we observed heterogeneity of phenotypes within networks, and the emergence of potentially biologically relevant microclusters of disease was evident.

Exemplars of observed microphenotypes generated by our analyses included 2 microclusters within TDA group 2 (Reticular Basement Membrane^{High}). The overall demonstration of increased reticular basement membrane thickness and variable epithelial area in group 2 might suggest that research evaluating extracellular matrix pathways,³⁹ epithelial wound repair (eg, EGFR receptor activation/signaling),⁴⁰ and epithelial metaplasia⁴¹ might shed a deeper insight into disease mechanisms in this group specially. The microcluster within group 2 was enriched for patients with late-onset disease and demonstrated marked airway hyperresponsiveness and tissue expression of a T_H17 tissue gene expression signature that was mutually exclusive with the T_H2 signature, as previously reported,¹⁰ and lamina propria neutrophilia.

Patients within this microcluster were taking high levels of both inhaled and, in many cases, oral corticosteroids, which have been proposed as a possible reason for the emergence of a T_H17 gene

signature and neutrophilic disease. However, against this hypothesis was the observed low level of tissue *PHACTR3* gene expression, which we observed consistently as being upregulated in patients taking high levels of corticosteroids. It is possible that this gene, which is associated with epigenetic modification (through DNA methylation) in patients with lung cancer,³⁵ might be unresponsive to corticosteroids in this population, rendering patients corticosteroid resistant or indeed that the gene and its product cross-talk with T_H17 inflammatory pathways, an assertion that would require *in vitro* evaluation. Nonetheless, we speculate that therapeutic strategies targeting IL-17 might be most fruitful in this microcluster.⁴² In contrast, the demonstration of an additional microcluster within the group 2 TDA network with severe early-onset atopic disease an persistent tissue and sputum eosinophilia despite high levels of corticosteroid exposure (GINA step 4–5, high tissue *PHACTR3* expression) might suggest a population that would benefit most from therapeutic approaches targeting the T_H2 axis.⁴³

Within TDA group 3 (Airway Smooth Muscle^{High}), we identified a microcluster with a high percentage ASM mass and increased numbers of tryptase-positive mast cells within both the ASM and lamina propria. This group had marked airway hyperresponsiveness and was enriched for patients with

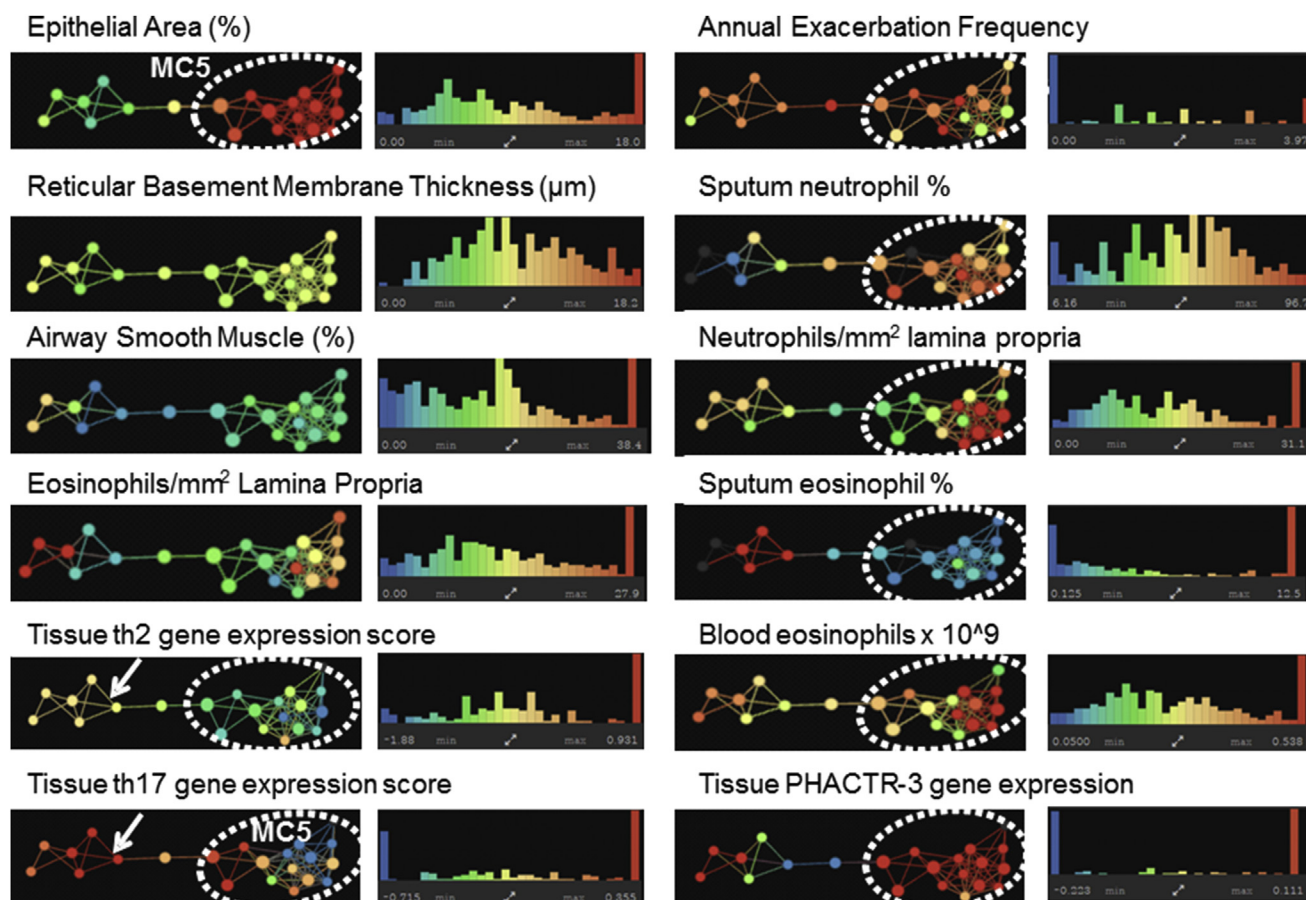


FIG 5. Remodeling^{Low} and nested microcluster: group 4 TDA network. Color annotations for pathological and clinical features are provided on the right-hand side of each corresponding TDA image. Overall, this group exhibited high treatment use and exacerbation frequency. However, microcluster 5 (MC5; white dotted outline) was identified within this network with a high epithelial percentage area and a non- T_H2 , primarily non- T_H17 tissue gene expression profile. Clinically, this subgroup demonstrated a high annual exacerbation frequency, high sputum percentage neutrophil count with a variable tissue neutrophil count, and concurrent blood eosinophilia in the absence of sputum eosinophilia. Gene expression of the DNA hypermethylation marker of the *PHACTR3* gene was enriched markedly in this subpopulation (in contrast to TDA network 1). White arrows represent regions within the network of reciprocal T_H2 and T_H17 gene expression in tissue.

exacerbations and lower lung function. Therefore it is possible that this group might be more amenable to existing therapeutic strategies targeting the ASM in asthmatic patients⁴⁴ and potentially novel therapies aimed at targeting tissue mast cells.⁴⁵ In support of this hypothesis, a recent report has identified significant attenuations in ASM mass in patients with a high baseline ASM mass undergoing thermoplasty.¹⁹

Finally, group 4 (Remodeling^{Low}) appeared to have frequent exacerbations in the absence of eosinophilic inflammation in tissue, sputum, and blood and group-level features of remodeling. However, we identified a microcluster enriched for increased epithelial area and sputum neutrophilia that appeared to be T_H2 low, with variable T_H17 gene expression in tissue. These observations highlight the fact that our understanding of the inflammatory mechanisms that drive airway pathology beyond the T_H2 and T_H17 axes in asthmatic patients remain poorly understood.

A common criticism of phenotyping studies is the lack of replication of clusters identified within external populations. We attempted to evaluate this concept using an independent replication

cohort of patients from the BOBCAT severe asthma study, comparing them with patients matched for asthma severity and treatment in the discovery cohort. It is well recognized that paraffin embedding promotes tissue shrinkage by approximately 30% in length and 50% in area when compared with GMA.²⁴ Therefore we normalized paraffin-embedded biopsy data from BOBCAT by using correction factors that assumed isotropic tissue shrinkage to allow for direct comparisons with GMA in the discovery severe asthma cohort. Our results are one of the first of their kind that demonstrate excellent replication of data topology in independent severe asthma cohorts and highlight the importance of using appropriate mathematic methods for topology replication.

Our study has a number of limitations. First, all of our observations with respect to pathology are cross-sectional. Second, we did not capture the full complement of remodeling features in biopsy tissue (eg, goblet cell hyperplasia). Third, substudies evaluating FEV₁ decrease and T_H2/T_H17 gene expression in tissue warrant further evaluation in studies with larger cohort sizes to study cluster-specific decrease and specific protein

TABLE II. Pathological features of the discovery population stratified by TDA grouping

	Mild-to-Moderate Asthma/Healthy ^{Continuum} , group 1 (n = 77)	Reticular Basement Membrane ^{High} , group 2 (n = 60)	Airway Smooth Muscle ^{High} , group 3 (n = 36)	Remodeling ^{Low} , group 4 (n = 19)	<i>P</i> value*
Total biopsy area (mm ²)	1.62 (1.07)	2.12 (1.46)§	1.49 (1.27)	1.36 (1.26)	.0256
Remodeling parameters					
Reticular basement membrane thickness (μm)	6.65 (4.14)‡§	12.83 (3.48)§	9.65 (3.83)	6.99 (1.64)‡§	<.0001
Epithelial area (%)	7.35 (5.98)	10.47 (10.31)§	5.83 (4.83)	14.87 (15.78)§	<.0001
ASM area (mm ²)	0.24 (0.24)§	0.16 (0.30)§	0.51 (0.43)	0.06 (0.08)‡§	<.0001
ASM area (%)	13.96 (16.79)§	9.77 (13.52)§	28.04 (17.77)	4.60 (6.73)§	<.0001
Glands (%)	0.00 (0.39)	0.00 (0.00)	0.00 (0.31)	0.00 (0.00)	.0740
Chalkley count	3.78 (2.00)	4.80 (2.30)	4.00 (2.50)	4.15 (0.70)	.0728
Tryptase-positive mast cells in ASM (cells/mm ²)	2.59 (6.12)	4.05 (7.26)	6.20 (8.19)	0.00 (6.25)	.0475
Lamina propria cells/mm ²					
Tryptase-positive mast cells	15.84 (15.96)	12.20 (17.66)	18.99 (28.00)	12.21 (14.93)	.1482
Major basic protein-positive eosinophils	6.81 (18.52)	9.75 (10.64)	9.87 (17.78)	7.04 (15.58)	.6008
Neutrophil elastase-positive neutrophils	9.55 (11.71)	9.04 (13.21)	14.21 (17.12)	8.05 (18.01)	.1977
Tissue gene expression					
T _H 2 gene expression score¶	−0.33 (1.03)	−0.29 (1.96)	−1.20 (0.46)	−1.17 (1.03)	.0635
T _H 17 gene expression score¶	−0.063 (0.95)	−0.316 (1.04)	0.587 (4.23)	0.097 (2.21)	.2184
<i>FKBP51</i> gene expression¶	0.20 (0.52)	0.28 (0.41)	0.52 (0.31)	0.38 (0.32)	.4473

All values not presented as means ± SEMs are presented as median (IQRs). Values in boldface indicate statistical significance.

*Statistical analysis across all groups. Adjusted *P* values (*P* < .05) are denoted as follows: †versus 1; ‡versus 2; §versus 3; and ||versus 4. Adjusted *P* values were obtained by using Bonferroni correction. All tests are Kruskal-Wallis tests.

¶Tissue gene expression analyses were only available in a subgroup of patients (see Fig E1).

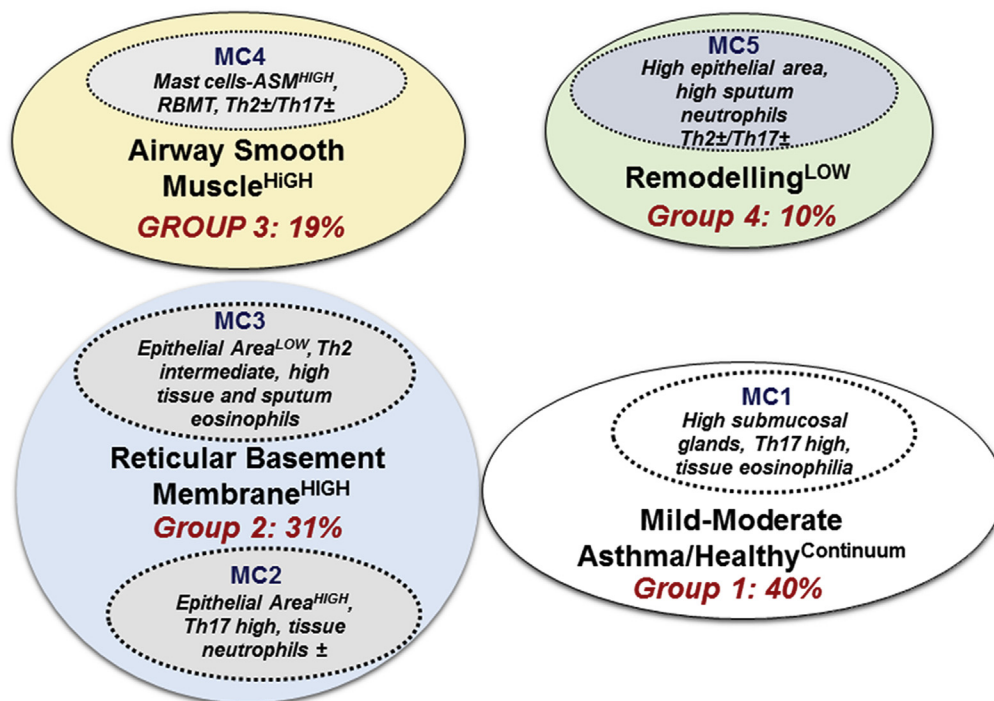


FIG 6. Discovery cohort, TDA cluster, and microcluster summary. A descriptive visual summary of the core pathological endotypes and microclusters identified in the discovery asthma and healthy volunteer population is shown. Group number refers to the TDA groups identified in Fig 1 with the corresponding percentage identifying the total proportion of the discovery population within a particular group. Five microclusters referred to in Figs 2-5 and denoted as MC1 to MC5 are summarized in the black dotted lines within their respective group/TDA network.

rather than gene expression markers of inflammatory disease, such as IL-17 and IL-22, in patients with T_H17 asthma.

Finally, treatment effects might have contributed to TDA group assignment, but in mitigation against this, TDA networks were

not generated by using GINA scores as an input function, and TDA groups showed clear differences in GINA treatment steps both within and between networks. Additionally, because of potential oversampling of patients with GINA steps 3 to 5 asthma

compared with steroid-naïve/low-dose inhaled steroid taking GINA step 1 to 2 asthma, future studies should compare pathological heterogeneity in both primary and secondary care asthma cohorts. Finally, although we were able to replicate broad topological structures in severe asthma, the replication of microclusters will require novel analytic methods and should form the basis of future research in this area.

In summary, we have used TDA to identify and visualize pathological disease clusters of asthma. Using this approach, we identified numerous microclusters (masked by group-level analysis) of disease that warrant further mechanistic evaluation and replicated pathology phenotypes by using appropriate mathematical methods in an independent population with severe asthma.

We thank Ayasdi (Menlo Park, Calif) and Dr Devi Ramanan for access to the Ayasdi software and SDK platforms for data analysis and Professors Alexander Gorban (Leicester, Mathematics) and Professor Gunnar Carlsson (Stanford, Mathematics) for their critique of the manuscript. We also thank Fiona Symon for pathology counts performed within the Leicester cohort and Ana Robles for pathology regions of interest identification within the BOBCAT cohort.

Key messages

- TDA identified 4 pathological phenotypes (Mild-Moderate Asthma/Healthy^{Continuum}, Reticular Basement Membrane^{High}, Airway Smooth Muscle^{High}, and Remodeling^{Low}) of patients with asthma/healthy volunteers, with numerous nested microclusters derived from bronchial biopsy specimens.
- Reciprocal T_H2 and T_H17 gene expression was demonstrated (1) across pathological phenotypes, (2) across the spectrum of asthma severity, and (3) in healthy volunteers.
- Our methodology provides new insight into pathological heterogeneity in asthmatic patients and is applicable to other complex chronic diseases.

REFERENCES

- Bousquet J, Mantzouranis E, Cruz AA, Ait-Khaled N, Baena-Cagnani CE, Bleecker ER, et al. Uniform definition of asthma severity, control, and exacerbations: document presented for the World Health Organization Consultation on Severe Asthma. *J Allergy Clin Immunol* 2010;126:926-38.
- Partridge MR, van der Molen T, Myrseth SE, Busse WW. Attitudes and actions of asthma patients on regular maintenance therapy: the INSPIRE study. *BMC Pulm Med* 2006;6:13.
- Robinson D, Humbert M, Buhl R, Cruz AA, Inoue H, Korom S, Hanania NA, et al. Revisiting type 2-high and type 2-low airway inflammation in asthma: current knowledge and therapeutic implications. *Clin Exp Allergy* 2017;47:161-75.
- Benayoun L, Druilhe A, Dombret MC, Aubier M, Pretolani M. Airway structural alterations selectively associated with severe asthma. *Am J Respir Crit Care Med* 2003;167:1360-8.
- Boucherat O, Boczkowski J, Jeannotte L, Delacourt C. Cellular and molecular mechanisms of goblet cell metaplasia in the respiratory airways. *Exp Lung Res* 2013;39:207-16.
- Siddiqui S, Sutcliffe A, Shikotra A, Woodman L, Doe C, McKenna S, et al. Vascular remodelling is a feature of asthma and non-asthmatic eosinophilic bronchitis. *J Allergy Clin Immunol* 2007;120:813-9.
- Siddiqui S, Hollins F, Saha S, Brightling CE. Inflammatory cell microlocalisation and airway dysfunction: cause and effect? *Eur Respir J* 2007;30:1043-56.
- Berair R, Brightling CE. Asthma therapy and its effect on airway remodelling. *Drugs* 2014;74:1345-69.
- Pretolani M, Dombret MC, Thabut G, Knap D, Hamidi F, Debray MP, et al. Reduction of airway smooth muscle mass by bronchial thermoplasty in patients with severe asthma. *Am J Respir Crit Care Med* 2014;190:1452-4.
- Choy DF, Hart KM, Borthwick LA, Shikotra A, Nagarkar DR, Siddiqui S, et al. TH2 and TH17 inflammatory pathways are reciprocally regulated in asthma. *Sci Transl Med* 2015;7:301ra129.
- Lum PY, Singh G, Lehman A, Ishkanov T, Vejdemo-Johansson M, Alagappan M, et al. Extracting insights from the shape of complex data using topology. *Sci Rep* 2013;3:1236.
- Li L, Cheng WY, Glicksberg BS, Gottesman O, Tamler R, Chen R, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci Transl Med* 2015;7:311ra174.
- Nielson JL, Paquette J, Liu AW, Guandique CF, Tovar CA, Inoue T, et al. Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nat Commun* 2015;6:8581.
- Sousa AR, Marshall RP, Warnock LC, Bolton S, Hastie A, Symon F, et al. Responsiveness to oral prednisolone in severe asthma is related to the degree of eosinophilic airway inflammation. *Clin Exp Allergy* 2017;47:890-9.
- Brightling CE, Bradding P, Symon FA, Holgate ST, Wardlaw AJ, Pavord ID. Mast-cell infiltration of airway smooth muscle in asthma. *N Engl J Med* 2002;346:1699-705.
- Siddiqui S, Mistry V, Doe C, Roach K, Morgan A, Wardlaw A, et al. Airway hyper-responsiveness is dissociated from airway wall structural remodeling. *J Allergy Clin Immunol* 2008;122:335-41.
- Shikotra A, Choy DF, Ohri CM, Doran E, Butler C, Hargadon B, et al. Increased expression of immunoreactive thymic stromal lymphopoietin in patients with severe asthma. *J Allergy Clin Immunol* 2012;129:104-11.
- Butler CA, McQuaid S, Taggart CC, Weldon S, Carter R, Skibinski G, et al. Glucocorticoid receptor β and histone deacetylase 1 and 2 expression in the airways of severe asthma. *Thorax* 2012;67:392-8.
- Butler CA, McQuaid S, Taggart CC, Weldon S, Carter R, Skibinski G, et al. Increased expression of bronchial epithelial transient receptor potential vanilloid 1 channels in patients with severe asthma. *J Allergy Clin Immunol* 2014;133:704-12.e4.
- Desai D, Newby C, Symon FA, Haldar P, Shah S, Gupta S, et al. Elevated sputum interleukin 5 and submucosal eosinophilia in obese individuals with severe asthma. *Am J Respir Crit Care Med* 2013;188:657-63.
- Jia G, Erickson RW, Choy DF, Mosesova S, Wu LC, Solberg OD, et al. Bronchoscopic exploratory research study of Biomarkers in Corticosteroid-refractory Asthma (BOBCAT) Study Group. Periostin is a systemic biomarker of eosinophilic airway inflammation in asthmatic patients. *J Allergy Clin Immunol* 2012;130:647-54.
- Bradding P, Feather IH, Howarth PH, Mueller R, Roberts JA, Britten K, et al. Interleukin 4 is localized to and released by human mast cells. *J Exp Med* 1992;176:1381-6.
- Sullivan P, Stephens S, Ansari T, Costello J, Jeffery P. Variation in the measurements of basement membrane thickness and inflammatory cell number in bronchial biopsies. *Eur Respir J* 1998;12:811-5.
- Miller PL, Meyer TW. Effects of tissue preparation on glomerular volume and capillary structure in the rat. *Lab Invest* 1990;63:862-6.
- Woodruff PG, Modrek B, Choy DF, Jia G, Abbas AR, Ellwanger A, et al. T-helper type 2-driven inflammation defines major subphenotypes of asthma [published erratum in *Am J Respir Crit Care Med* 2009;180:796]. *Am J Respir Crit Care Med* 2009;180:388-95.
- Choy DF, Modrek B, Abbas AR, Kummerfeld S, Clark HF, Wu LC, et al. Gene expression patterns of Th2 inflammation and intercellular communication in asthmatic airways. *J Immunol* 2011;186:1861-9.
- Edelsbrunner H, Letscher D, Zomorodian A. Topological persistence and simplification. *Discrete Comput Geom* 2002;28:511-33.
- Zomorodian A, Carlsson G. Computing persistent homology. *Discrete Comput Geom* 2005;33:249-74.
- Bubenik P. Statistical topological data analysis using persistence landscapes. *J Machine Learn Res* 2015;16:77-102.
- Terese Fasy B, Lecci F, Rinaldo R, Wasserman L, Balakrishnan S, Singh A. Confidence sets for persistence diagrams. *Ann Stat* 2014;42:2301-39.
- JavaPex: a research software package for persistent (co)homology, Proceedings of ICMS 2014, Lecture Notes in Computer Science 8592, 129–136, 2014. Available at: <http://appliedtopology.github.io/javapex/>. Accessed November 2016.
- Kelly MM, King EM, Rider CF, Gwozd C, Holden NS, Eddleston J, et al. Corticosteroid-induced gene expression in allergen-challenged asthmatic subjects taking inhaled budesonide. *Br J Pharmacol* 2012;165:1737-47.
- Huang YJ, Nariya S, Harris JM, Lynch SV, Choy DF, Arron JR, et al. The airway microbiome in patients with severe asthma: associations with disease features and severity. *J Allergy Clin Immunol* 2015;136:874-84.
- Wilson SJ, Ward JA, Sousa AR, Corfield J, Bansal AT, De Meulder B, et al. Severe asthma exists despite suppressed tissue inflammation: findings of the U-BIOPRED study. *Eur Respir J* 2016;48:1307-19.

35. Hubers AJ, Heideman DA, Burgers SA, Herder GJ, Sterk PJ, Rhodius RJ, et al. DNA hypermethylation analysis in sputum for the diagnosis of lung cancer: training validation set approach. *Br J Cancer* 2015;112:1105-13.
36. Hong Y, Ho KS, Eu KW, Cheah PY. A susceptibility gene set for early onset colorectal cancer that integrates diverse signalling pathways: implication for tumorigenesis. *Clin Cancer Res* 2007;13:1107-14.
37. Notterman DA, Alon R, Sierk AJ, Levine AJ. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Res* 2001;61:3124-30.
38. Knöbel M, O'Toole EA, Smith FJ. Keratins and skin disease. *Cell Tissue Res* 2015;360:583-9.
39. Burgess JK, Mauad T, Tjin G, Karlsson JC, Westergren-Thorsson G. The extracellular matrix—the under-recognized element in lung disease? *J Pathol* 2016;240:397-409.
40. Davies DE, Polosa R, Puddicombe SM, Richter A, Holgate ST. The epidermal growth factor receptor and its ligand family: their potential role in repair and remodelling in asthma. *Allergy* 1999;54:771-83.
41. Stinson SE, Amrani Y, Brightling CE. D prostanoid receptor 2 (chemoattractant receptor-homologous molecule expressed on TH2 cells) protein expression in asthmatic patients and its effects on bronchial epithelial cells. *J Allergy Clin Immunol* 2015;135:395-406.
42. Busse WW, Holgate S, Kerwin E, Chon Y, Feng J, Lin J, et al. Randomized, double blind, placebo-controlled study of brodalumab, a human anti-IL-17 receptor monoclonal antibody, in moderate to severe asthma. *Am J Respir Crit Care Med* 2013;188:1294-302.
43. Gandhi NA, Bennett BL, Graham NM, Pirozzi G, Stahl N, Yancopoulos GD. Targeting key proximal drivers of type 2 inflammation in disease. *Nat Rev Drug Discov* 2016;15:35-50.
44. Wright DB, Trian T, Siddiqui S, Pascoe CD, Johnson JR, Dekkers BG, et al. Phenotype modulation of airway smooth muscle in asthma. *Pulm Pharmacol Ther* 2013;26:42-9.
45. Beunk L, Verwoerd A, van Overveld FJ, Rijkers GT. Role of mast cells in mucosal diseases: current concepts and strategies for treatment. *Expert Rev Clin Immunol* 2013;9:53-63.

ONLINE SUPPLEMENTARY MATERIAL

**Airway Pathological Heterogeneity in Asthma: Visualisation of Disease Micro-Clusters
using Topological Data Analysis**

¹Salman Siddiqui, MRCP, PhD.

¹Aarti Shikotra, BSc, PhD.

¹Matthew Richardson BSc, PhD.

³Emma Doran PhD.

²David Choy BSc.

^{1,4}Alex Bell MMath.

²Cary D. Austin, MD, PhD.

²Jeffrey Eastham-Anderson

¹Beverley Hargadon RGN.

²Joseph R. Arron MD, PhD.

¹Andrew Wardlaw, FRCP, PhD.

¹Christopher E. Brightling, FRCP, PhD.

³Liam G. Heaney, FRCP, MD.

¹Peter Bradding, FRCP, DM.

¹ Department of Infection Immunity and Inflammation/ Institute for Lung Health University
of Leicester/Glenfield Hospital, Groby Road, Leicester, LE3 9QP.

² Genentech, Inc. South San Francisco, CA, USA.

³ Centre for Infection and Immunity, Health Sciences Building, Queens University Belfast,
Lisburn Road, Belfast, UK.

⁴ Department of Mathematics, University of Leicester, University Road, Leicester, United
Kingdom

METHODS

[1] PATIENT COHORTS

1.1 Patient characterisation within the discovery and replication cohorts is summarised in supplementary **Figure E1**. The discovery population consisted of 142 patients with adult asthma as annotated in the figure (orange ellipse), alongside the number of patients at each GINA treatment step. As the cohort were derived from patients attending both Leicester and Belfast severe asthma cohorts, central quality control and assurance measures were taken to ensure measurement standardisation across cohorts. Additionally 60 aged matched controls without asthma were included in the discovery cohort. The replication cohort was derived from the BOBCAT study (E2), with 59 patients with severe asthma derived from this cohort. Bronchial mucosal biopsies were embedded in GMA in the discovery cohort and paraffin in the replication cohort. The text box identifies additional phenotyping data acquired in the respective cohorts. This included whole biopsy gene expression analysis performed using two colour Whole Human Genome (WHG) 4x44k gene expression microarray analysis, in subsets of asthmatics in both cohorts and healthy volunteers in the discovery cohort, lung function decline data in a subset of asthma patients in the discovery cohort only, and clinical meta data in both cohorts (tables of % completeness of this data are provided in the **supplementary tables E1 a-d**).

[2] PATIENT CHARACTERISATION AND CLINICAL PHENOTYPING

2.1 Clinical definitions of asthma in the Discovery and BOBCAT-Replication cohorts

Discovery Cohort: All patients with asthma were current non-smokers, with a ≤ 10 pack year smoking history, and met the following diagnostic criteria (i) an expert physician diagnosis of asthma with (ii) objective evidence (defined as one or more of peak flow variation of $\geq 20\%$ over a two week period, bronchodilator reversibility of $\geq 12\%$ and 200 ml or AHR [methacholine $PC_{20} \leq 8\text{mg/ml}$]). Asthma treatment intensity was utilised as a marker of

disease severity and was defined according to the Global Initiative for Asthma (GINA) treatment steps with the following severity strata (GINA 1 n=30 [mild], GINA 2-3 n=31 [moderate], GINA 4-5 [severe] n=81). All patients with asthma had been free of an exacerbation and had been on stable asthma treatment for at least 6 weeks prior to study entry. Healthy volunteers (n=60) were identified from local advertising. They were selected on the basis of the following criteria (i) an absence of any historical diagnosis or symptoms of respiratory disease, (ii) current non-smokers with a smoking history of ≤ 10 pack years and (iii) normal airway physiology (methacholine PC₂₀ ≥ 16 mg/ml and normal pre-bronchodilator spirometry).

Replication Cohort: BOBCAT was a multicenter study conducted in the United States, Canada, and the United Kingdom to collect matched airway and blood samples in patients with moderate-to-severe asthma. Inclusion criteria required a diagnosis of moderate-to-severe asthma (confirmed by an FEV1 between 40% and 80% of predicted value, as well as evidence within the past 5 years of $>12\%$ reversibility of airway obstruction with a short-acting bronchodilator or methacholine sensitivity [PC₂₀ < 8 mg/mL) that was uncontrolled (as defined by at least 2 exacerbations in the prior year or a score of >1.50 on the ACQE3 while receiving a stable dose regimen [>6 weeks] of a high-dose ICS [>1000 μ g fluticasone or equivalent per day]) with or without a long-acting β -agonist. Permitted concomitant medications also included leukotriene receptor antagonists and oral corticosteroids [E1].

2.2 Allergy Testing

Discovery Cohort: Atopic status was assessed using skin prick testing to the common airborne allergens cat, dog, house dust mite, tree pollen, and *Aspergillus Fumigatus* (Alk-Abello, Denmark) and total IgE testing. Sensitisation was defined as a positive skin prick test

with a weal of ≥ 3 mm. Patients had withheld taking anti-histamines for a minimum of 72 hours prior to testing.

2.3 Sputum Induction

Discovery Cohort: Sputum induction was performed using previously described methods [E2]. Prior to testing, patients inhaled 200 μ g of salbutamol. Patients were asked to blow their nose and rinse their mouth with water. Sputum induction was undertaken using normal saline given via an ultrasonic nebulizer, followed by successive concentrations of saline from 3% up to 5% to aid expectoration. The expectorate was processed within 2 hours.

2.4 Pulmonary function testing

Discovery and Replication cohorts: All patients had standard spirometry (Vitalograph Ltd, Maids Moreton, UK) according to American Thoracic Society/European Respiratory Society guidelines [E3] with reversibility after 15 minutes using 200 μ g of inhaled salbutamol via a spacer.

Discovery Cohort: A sub-cohort of asthmatics [43 patients from Leicester and 8 patients from Belfast] had serial pre bronchodilator FEV₁ available. This population were utilised to evaluate the effect of remodelling cluster membership upon lung function decline. Pre bronchodilator FEV₁ was chosen *a priori* as bronchodilator responsive remodelling features such as ASM mass were input parameters in the cluster models.

2.5 Bronchoscopy

Discovery and Replication cohorts: Subjects underwent bronchoscopy conducted according to British Thoracic Society guidelines [E4]. Bronchial mucosal biopsy specimens were taken from the right middle lobe and lower lobe carinae, fixed in acetone, and embedded in GMA,

as described previously [E5]. A subset of biopsies [21 healthy controls and 51 asthmatic subjects] were also placed immediately in RNA preservative (RNAlater, Ambion, Austin, TX) and submitted for microarray analysis.

[3] ETHICAL APPROVAL STATEMENT

Discovery cohort: The studies through which patients were recruited were approved by the local Research Ethics Committee for both institutions (Leicester: Leicestershire, Northamptonshire, & Rutland Research Ethics Committee reference 04/Q2502/74; Belfast: Office of Research and Ethics Committee of Northern Ireland reference 06/NIR02/114; Northampton Research Ethics Committee 05/Q2502/8; Leicester Research Ethics Committee, reference Number 4977). Written informed consent was gained from all participants prior to their involvement. Replication cohort: For the BOBCAT study the research protocol was approved by relevant institutional review boards, and informed consent was obtained from all subjects before enrolment [E1].

[4] DATA COMPLETENESS SUMMARY

Full details of percentage missing data in the discovery cohort for (i) clinical meta data and (ii) pathology features according to GINA treatment intensity strata in **Tables E1 a-b** and according to topological data analysis (TDA) grouping **Tables E1 c-d**. Importantly input variables utilised for both factor and cluster analyses were derived from complete data ($\geq 90\%$ completeness) with imputation of median values utilised for missing data.

[5] CROSS COUNTING OF PATHOLOGY FEATURES

DISCOVERY COHORT: All pathological data were assessed by observers blinded to the identity of the patient. The population of patients identified for cross counts included both asthmatic and healthy subjects. A single blinded observer AS at Leicester performed cross

counts in a random selection of asthmatics and healthy cases to evaluate within and between centre repeatability.

Repeatability was assessed using two way analysis of variance and intra class correlation coefficient. **Table E2** reports the results of this analysis and demonstrates good to excellent repeatability (ICC of averaged measures > 0.75) for all of the pathology features.

[6] CORRECTION FOR TISSUE SHRINKAGE IN PARAFFIN EMBEDDED BIOPSIES: REPLICATION COHORT

Due to the known effect of paraffin embedding causing tissue shrinkage by approximately 30% in length and 50% in area when compared to GMA [E6], all biopsy measures in paraffin were normalised to account for this effect. **Table E3a** demonstrates the impact of not correcting for tissue shrinkage, in comparison to corrected data which is presented in **Table E3b**. We assumed that tissue shrinkage was isotropic and uniform within biopsies. The reticular basement membrane thickness was measured using the technique reported by Jeffrey *et al* [E7] and corrected for tissue shrinkage of 30% by multiplying values by 1.3. Lamina propria eosinophils counts were expressed per tissue unit area, as cells per square millimeter squared of the lamina propria. A correction factor of $\times 2.04$ ($1/[1-0.3]^2$) was applied to normalise for tissue shrinkage. The epithelial area was expressed as a percentage of the total biopsy area. No correction for shrinkage was applied as it was assumed that isotropic shrinkage of both epithelial and area and total biopsy area would cancel out. The airway smooth muscle area was measured in mm^2 and expressed per unit length of the basement membrane (mm^2/mm). Normalisation for shrinkage was therefore applied by using the correction factor of $\times 1.57$ ($[1-0.3]^2/1.3$). Tissue analysis was performed blind to the clinical data by JE and CA. The inter-observer variation was less than 5%.

[7] FEV₁ DECLINE MODELLING (DISCOVERY COHORT)

Pre bronchodilator FEV₁ decline modelling was performed using PROC GLIMMIX in SAS V9.4 using linear mixed effects models and the restricted maximum likelihood method as previously reported [E8]. Linear mixed model was fitted to evaluate patient specific lung function change trajectories. Fixed effects in the models included: Age, Height, Sex, Time, Random effects in the model were: intercept, and, time, for each patient (**Table E3a**).

Average change in FEV₁ over 1 year for each patient was determined as $= \beta_t + \gamma_{ti}$

Where β_t is the fixed effect for time and γ_{ti} is the random effect for time for patient I .

The algebraic expression for the decline model was:

$$Pre\ BD\ FEV = \beta_0 + \beta_1 Age + \beta_2 Height + \beta_3 Sex + \beta_4 Time + \gamma_i + \varepsilon$$

[8] GENE EXPRESSION ANALYSES DISCOVERY/REPLICATION COHORTS

Gene expression analysis was performed using R 3.3.1. A gene expression profile of Th2 and Th17 cytokine activity was defined as previously reported [E9-E11] in all subjects that had available data. Additionally a ‘discovery and replication approach’ was applied to the gene expression data as follows: (i) A subset of patients from the Discovery cohort that had undergone gene expression profiling were selected comprising healthy controls and patients having moderate to severe asthma GINA treatment step 3-5 (the asthma group). Where there were multiple probes associated with a single gene the probe with the greatest IQR was chosen, this was accomplished using the genefilter library in R. The Mann-Whitney test was applied for all genes in the subset to test for a difference in log ratio values between the asthma group and the healthy controls. The resulting p-values were then adjusted using FDR

(False Discovery Rate). The FDR adjusted p-values were then ranked and the top 20 differentially expressed genes in the severe asthma fraction of the discovery cohort were identified. (ii) The top 20 genes were carried forward for the same analysis within the BOBCAT Replication cohort (GINA 4-5), utilising the same population of healthy controls from the discovery cohort. The p-values were again adjusted using FDR. A p value of < 0.05 following adjustment at each stage for multiple discovery was utilised to determine significant association.

[6] ANALYTICAL METHODS FOR TOPOLOGICAL NETWORK DERIVATION

To generate data on pathological phenotypes in asthma a three-step statistical process was employed and incorporated (i) factor analysis to reduce the dimensionality of nine pathological features, (ii) topological data analysis with the input function derived from pathological factor analysis (The analytical process of TDA is depicted in **Figure E3**) and (iii) mapping of variables not utilised to generate TDA networks to the networks e.g. FEV₁ decline, biopsy gene expression.

STEP 1: Firstly, factor analysis was utilised for dimension reduction of the nine core pathological features. Four pathology features were selected on this basis [reticular basement membrane thickness, $\log_{10}(\text{epithelial area } \%+1)$, $\log_{10}(\text{ASM\% area}+1)$, and $\log_{10}(\text{eosinophils/mm}^2 \text{ lamina propria}+1)$].

STEP 2: The four features were then utilised as inputs into Topological Data Analysis (TDA) [E12-4], but were not $\log_{10}$ transformed in the TDA. TDA provides a powerful way of

visualising complex data by using a combination of unsupervised machine learning and data visualisation. TDA was performed with the Ayasdi CORE software (Ayasdi, Palo Alto, California). This approach represents a high dimensional data set as a structured three-dimensional network in which each “node” comprises subjects similar to each other in multiple dimensions. Lines or “edges” are drawn between nodes that contain shared data points. Statistical tests can then be performed on groups or features that emerge from the inherent structure of the data set. TDA was performed using the Ayasdi software and SDK (Ayasdi, Palo Alto, California).

Three inputs were used: a distance metric, one or more filter functions, and two resolution parameters (“resolution” and “percent overlap” or “gain”). A network of nodes with edges between them was created by using a force-directed algorithm. The nodes represent bins or “micro clusters” of data points, and 2 nodes are connected if their corresponding collections of data points have a point in common. Norm correlation was used as a distance metric, norm correlation is the Pearson correlation coefficient applied to the normalised variables, i.e., variables are transformed to follow a standard normal distribution. Two filter functions: neighbourhood lens 1 and neighbourhood lens 2 were applied, these lenses allow the embedding of high dimensional data into a two dimensional space by embedding a K nearest neighbours graph. Resolution and gain settings were selected where the network structure permits identification of subgroups.

STEP 3: Finally, in the third step, variables not utilised to generate the pathological clusters in TDA including clinical meta data were summarised in each TDA group by linkage to the group number.

Micro clusters were defined as regions within networks of patients (nodes) that were similar (connected by lines) and to avoid over interpretation of the data represented visible sub-structures within networks e.g. loops, tails and large regions within networks separated by significant distances. Microclusters were utilised to understand within network pathological heterogeneity within the discovery cohort.

A detailed descriptive mathematical narrative of figure E3 is provided below.

E3A: Geometric association and motivation from a data matrix of real numbers.

(i) Data analysis typically involves a table of numbers, with attributes (columns) a_i and object vectors (rows) v_i . The ‘data matrix’ is the corresponding subset of $\mathbb{R}^{m \times n}$, for n objects having m attributes. In this study, $m = 4$ and $n = 100$. (ii) The data items may be perceived as points in some geometric space. Up to 3 dimensions may be easily visualized at a time. Distance in this space may be associated with the notion of similarity (i.e. if points are close, then they are similar). (iii) Points in space may be considered to have some shape to them; it is our desire to extract an object resembling this shape, and examine its properties. The illustration suggests some continuous figure, composed of two possibly distinct sections.

E3B: Simple demonstration of ‘mapper’ algorithm to achieve the aforementioned goal.

(i) Consider a set of points S in the plane, bearing the shape inferred in (A). A metric function $d: S \times S \rightarrow \mathbb{R}$, may encode **similarity** for a set of objects S . In this demonstration, S is a subset of the plane ($S \subset \mathbb{R}^2$) and d is the Euclidean metric ($d(v_1, v_2) = \sqrt{|v_1 - v_2|^2}$). Let $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ assign each point to its x-coordinate. (ii) The image $f(S)$ is a subset of the real line, which may be partitioned into overlapping intervals for the purpose of evaluating these points with respect to their similarity (or closeness) in the x-direction. (iii) Cluster analysis is applied to the pre-images of these overlapping intervals (i.e. the sets of points in space that map to those intervals). Note that some clusters have points in common, e.g. C_1 and C_2 have

non-empty intersection ($C_1 \cap C_2 \neq \emptyset$), these clusters both contain the brown point. (iv) The product of this process is a simplicial complex, or simply a complex, that is a collection of objects called simplexes. The 0-dimesnional simplexes (termed 0-simplexes) are associated to the identified clusters. The 1-simplexes indicate clusters that have non-empty intersection (sharing at least one point in common), essentially indicating similarity of the x-coordinate. (v) A geometric visualization of this complex may be obtained in infinitely many ways (non-unique). Pseudo-physical simulation may be applied to achieve a desirable picture (easy to see the points and relationships). The 0-simplexes are initialized at time t_0 as a grid of points, and a force function is defined from the 1-simplexes. Time is then progressed to some final state (time t_f), concluding the visualization process (adding lines between the points for the 1-simplexes). This resulting visualization captures the idea that there are two entities present. Varying the number of intervals (resolution), and their degree of overlap (gain), will produce results of varying granularity (e.g. may capture the continuous relationship between the two entities).

E3C: Metric and lenses used in this study.

(i) The norm correlation metric determines the similarity of two vectors based on their component wise correlation, post normalization to zero mean with unit variance (z-score). Given vectors v_i, v_j , we calculate distance between them as $d(v_i, v_j) = 1 - r(\hat{v}_i, \hat{v}_j)$, where r is the Pearson product-moment correlation coefficient and $\hat{v}$ is v post normalization. The illustration demonstrates calculation of the norm correlation for two 4 dimensional vectors that have strong positive correlation (hence $d \approx 0$). (ii) The neighbourhood lenses project a vector onto its x and y values of an embedded k-nearest neighbour graph. The illustration demonstrates 3-nearest neighbours applied to a set of points, followed by a physical simulation (replacing connections with ‘springs’) leading to the embedded graph. This may then be covered by a grid of overlapping intervals, as discussed in B, to determine a complex

for visualization. N.B. here the image space is a subset of $\mathbb{R}^2$, with overlapping polygons,
rather than intervals on a line.

*In this study, we applied the mapper algorithm, using norm correlation metric and the
neighbourhood lenses, to a data set using 4 pathological features.*

[9] VIETORIS-RIPS FILTRATION AND TOPOLOGICAL SHAPE REPLICATION

To determine whether the pathological objects created in the discovery and replication severe
asthma cohorts were ‘similar’, we utilised a technique to identify data shape similarity in high
dimensional simplices called the Vietoris-Rips filtration technique. This method was utilised
as simple two dimensional visual comparisons of networks may be misleading when
comparing four dimensional pathological networks.

Figure E4 illustrates some of the essential ideas behind persistent homology. An example
data set is shown in the top left plot in **Figure E4b**, the data points appear to lie on a circle
with unit radius, the sample size $n=20$. The data set is a sample from an underlying
population, persistent homology enables us to make some statements about the shape of the
underlying population. Shapes can be described by a set of fundamental properties; connected
components, loops and voids. For example, a circle can be described as a set of connected
components along with a loop (or 1 dimensional hole) see **Figure E4a**, a sphere can be
described as a set of connected components along with a void (or 2 dimensional hole). We
can regard the data set shown in the top left of **Figure E4b**, as $n=20$ connected components,
if we centre a circle of radius ε on each point in the data set and let ε increase from 0 to 1
Figure OS E1 c-e the number of connected components decreases, at $\varepsilon = 0$ we have 20
connected components at $\varepsilon = 1$ we have one connected component, as we vary ε connected

components emerge and then disappear or are born and then die. As ε increases we begin to see the loop forming, the loop emerges at $\varepsilon = 0.9$ and dies at $\varepsilon = 1$, at $\varepsilon = 1$ all points are connected into a single point, so there is no loop or hole present. As ε is increased over a range of values ‘real’ features in the data set will persist. If ε_b is the value of ε when a feature is born and ε_d is the value of ε when a feature dies, then $\varepsilon_d - \varepsilon_b$ will be large for persistent features and small for transient features, transient features are likely to be noise. If we connect a subset of points that have overlapping or intersecting circles we form a network of points, these networks or groups of connected points can be described using simplices, a single point is a 0 dimensional simplex, two connected points (a straight line) is a 1 dimensional simplex, a triangle is a 2 dimensional simplex. In **Figure E4f** examples of simplices are shown for $\varepsilon = 0.06$ and $\varepsilon = 0.11$. If all subsets of points in the data set with intersecting circles and a common intersection are connected we obtain a set of simplices known as a simplicial complex, the particular simplicial complex that is formed in this case is known as the Čech complex. As we vary ε from 0 to 1 as in our example we get a sequence of Čech complexes, this sequence of complexes is called a filtration. The Čech complex can be quite computationally expensive to generate, the Vietoris-Rips complex is an approximation of the Čech complex, the Vietoris-Rips complex is constructed by joining simplices formed by considering only pairwise connections, this differs from the Čech complex where a simplex is included in the simplicial complex if the points in the simplex have a common intersection see Figure OS E1f. Informally we could say that with the Čech complex more computations are needed to achieve the same degree of connectedness as obtained in the Vietoris-Rips complex.

The sequence of Vietoris-Rips complexes obtained is known as the Vietoris-Rips filtration. The barcode plot for the example data set discussed above is shown in **Figure E4g**, barcode plots provide a topological summary, the endpoints of a bar represent the birth and death

times of the feature, hence long bars represent more persistent features, short bars represent noise. Earlier we said that shapes can be described by a set of fundamental properties; connected components, loops and voids, these properties can be mathematically described by homology groups, H_k is the k th homology group. H_0 describes connected components, H_1 describes loops (1 dimensional holes), H_2 describes voids (2 dimensional holes) and H_k describes k dimensional holes. Connected components, loops and voids can also be described by Betti numbers, the Betti number β_k counts the number of k dimensional holes, β_k is the rank of H_k . We see that β_0 counts the number of connected components, β_1 the number of loops and β_2 the number of voids. The barcode plots in **Figure E4g** shows the birth and death of features as a function of ε for the homology groups H_0 , H_1 and H_2 .

Persistent homology for the discovery severe asthma fraction and the replication (BOBCAT) data was computed using JavaPlex [E15] in MATLAB R2016b. Data sets comprising the four pathology features, reticular basement membrane thickness, epithelial area %, ASM% area, and eosinophils/mm² lamina propria in the discovery severe asthma fraction and replication cohort were examined. Vietoris-Rips filtration was used throughout.

336

337 **RESULTS**

338 **[1] ANALYSIS OF PATHOLOGICAL FEATURES IN THE DISCOVERY COHORT**
339 **ACCORDING TO GINA TREATMENT INTENSITY STRATA**

340 **Table E4 a-b** report clinical and pathological (**Tables E5 and Figure E5**) features in the
341 discovery population stratified by GINA treatment step.

342 **Table E6** reports the 90th percentile point of the pathology features within the health
343 population. The percentage of patients in each GINA treatment intensity stratum with values
344 above the 90th percentile point is presented as a comparator.

345 We identified that the pathological features were distributed differently according to asthma
346 GINA treatment intensity/severity strata. Specifically, all asthma severity strata had evidence
347 of a significant tissue eosinophilia. In contrast neutrophilic tissue inflammation was not a
348 feature at a group level in any of the asthma strata and was only observed above the 90th
349 percentile healthy population value in 13.3% with GINA 1 asthma, 21.4 % with GINA 2-3
350 asthma and 18.3% in GINA 4-5 asthma; $p = 0.43$.

351 Mast cell infiltration of the smooth muscle was significantly increased in GINA 1[4.17 (3.07-
352 11.47) cells /mm²] and GINA 2-3 [5.77 (1.87-17.43) cells /mm²] strata but not in the GINA
353 4-5 strata [4.01 (0-9.0 cells) /mm²]. In addition mast cells/mm² of the lamina propria were not
354 significantly increased in asthma when compared to healthy volunteers.

355 RBM thickness was significantly increased in all asthma strata compared to healthy
356 volunteers ($p < 0.0001$). 34.5 % of GINA 1, 32.2% GINA 2-3 and 59.2% of GINA 4-5
357 patients had a basement membrane thickness above the 90th percentile point in healthy
358 volunteers ($p < 0.0001$).

The Chalkley count (a marker of vascular remodelling) was significantly increased in mild GINA 1 (4.65 [3.16-5.67]; n= 12/30) and severe GINA 4-5 asthma (5.00 [4.00-6.00]; n= 55/81) compared to the healthy control group (3.40 [2.76-.3.95]; n= 32/60), p<0.0001.

In contrast we failed to observe significant differences in epithelial percentage area, glandular percentage area or percentage airway smooth muscle area between the GINA strata and healthy volunteers. However when expressed as a percentage above the 90th percentile healthy value, the absolute smooth muscle area in mm² was increased in asthma in 16% GINA 1, 32.1% GINA 2-3 and 36.4% GINA 4-5; p=0.003.

[2] GENE EXPRESSION

A total of eleven differentially expressed genes between severe asthma and healthy controls, (after accounting for false discovery), were replicated in the BOBCAT severe asthma cohort and are outlined in **Supplementary Table E7 and Figure E6**.

[3] FEV₁ DECLINE

Age and height were found to be statistically significant predictors of pre bronchodilator FEV₁. However it can be seen that age and height adjusted FEV₁ decline was not significant when considering time as fixed effect and random effect (decline model 1) **Supplementary tables E3a-b**. Additionally the decline slopes centred at 0 mls/year for all the TDA clusters had wide confidence intervals and we were not powered to assess cluster*decline interactions. **Supplementary Figure E7**.

[4] INVESTIGATION OF TOPOLOGICAL SIMILARITY: DISCOVERY AND REPLICATION SEVERE ASTHMA COHORTS

Clinical and pathological characteristics of the discovery and replication severe asthma (GINA 4-5) populations are presented in **tables E8a-b**.

Patients in the discovery and replication cohorts with severe asthma were matched for age and sex (**Supplementary Tables E8 a-b**). A significantly higher proportion of patients were receiving oral corticosteroids in the discovery cohort when compared to the replication cohort. The two cohorts were well matched for the degree of eosinophilic airway inflammation in blood and sputum, and for FeNO. In contrast the replication cohort had significantly poorer lung function and greater bronchodilator reversibility when compared to the discovery cohort. Tissue Th2 and Th17 gene expression cores were similar in both discovery and replication cohorts. However expression of the steroid exposure gene *FKBP5*, was significantly higher in the replication cohort when compared to the discovery cohort, indicative possibly indicative of significant batch effects rendering absolute numerical comparison of gene expression between the two cohorts difficult.

The results of computing persistent homology for the discovery and replication data sets are shown in Barcode plots **Figure E8**. The plots for discovery and replication populations are very similar in their dimension 0 homology but different in their dimension 1 homology. Two of the four input variable into TDA, eosinophils/mm² lamina propria and epithelial area % were transformed as they did not exhibit a normal distribution, using log₁₀ (epithelial area %+1) and log₁₀ (eosinophils/mm² lamina propria+1) transformation. Following both correction for 30% tissue shrinkage in paraffin and data transformation, both discovery and replication populations demonstrated highly similar level 0 and level 1 homologies indicative of homology replication.

However Visual inspection of networks (**Supplementary Figure E9a-b**) utilising TDA applied to the four pathological variables in the discovery cohort (GINA 4-5 only) generated a single network with four overlapping groups of patients with a high ASM% area, high epithelial % area, high lamina propria eosinophils and high reticular basement membrane thickness respectively.

Th2 and Th17 gene expression were mutually exclusive within both the discovery and replication networks

Thus in summary, there was a close similarity of the TDA networks between discovery and replication cohorts with respect to data shape and persistent homology after correcting for paraffin embedding related tissue shrinkage and data that were non-normally distributed (**Figure E8**). In contrast visual inspection of networks did not aid decision making with respect to phenotype replication (**Supplementary Figure E9 a-b**).

435

436 REFERENCES

437 [E1] Jia G, Erickson RW, Choy DF, Mosesova S, Wu LC, Solberg OD, Shikotra A, Carter R,
438 Audusseau S, Hamid Q, Bradding P, Fahy JV, Woodruff PG, Harris JM, Arron JR;
439 Bronchoscopic Exploratory Research Study of Biomarkers in Corticosteroid-refractory
440 Asthma (BOBCAT) Study Group. Periostin is a systemic biomarker of eosinophilic airway
441 inflammation in asthmatic patients. *J Allergy Clin Immunol*. 2012 Sep;130(3):647-654.

442 [E2] P. Chanez, O. Holz, P.W. Ind, R. Djukanović, P. Maestrelli, P.J. Sterk. Sputum
443 induction. *Eur Respir J*, 20 (2002), pp. 3s–8s

444 [E3] M.R. Miller, J. Hankinson, V. Brusasco, F. Burgos, R. Casaburi, A. Coates, *et al*.
445 Standardisation of spirometry. *Eur Respir J*, 26 (2005), pp. 319–338

446 [E4] British Thoracic Society Bronchoscopy Guidelines Committee, a Subcommittee of the
447 Standards of Care Committee of the British Thoracic Society. British thoracic society
448 guidelines on diagnostic flexible bronchoscopy. *Thorax*. 2001 February 01;56(suppl 1):i1-
449 i21.

450 [E5] Bradding P, Feather IH, Howarth PH, Mueller R, Roberts JA, Britten K, *et al*.
451 Interleukin 4 is localized to and released by human mast cells. *J Exp Med*. 1992 Nov
452 1;176(5):1381-6.

453 [E6] Miller PL, Meyer TW. Effects of tissue preparation on glomerular volume and capillary
454 structure in the rat. *Lab Invest*. 1990 Dec;63(6):862-6.

455 [E7] P. Sullivan, D. Stephens, T. Ansari, J. Costello, P. Jeffery. Variation in the
456 measurements of basement membrane thickness and inflammatory cell number in bronchial
457 biopsies. *Eur Respir J*, 12 (1998), pp. 811–815.

[E8] Newby C, Agbetile J, Hargadon B, Monteiro W, Green R, Pavord I, Brightling C, Siddiqui S. Lung function decline and variable airway inflammatory pattern: longitudinal analysis of severe asthma. *J Allergy Clin Immunol*. 2014 Aug;134(2):287-94.

[E9] Choy DF, Hart KM, Borthwick LA, Shikotra A, Nagarkar DR, Siddiqui S, Jia G, Ohri CM, Doran E, Vannella KM, Butler CA, Hargadon B, Sciurba JC, Gieseck RL, Thompson RW, White S, Abbas AR, Jackman J, Wu LC, Egen JG, Heaney LG, Ramalingam TR, Arron JR, Wynn TA, Bradding P. TH2 and TH17 inflammatory pathways are reciprocally regulated in asthma. *Sci Transl Med*. 2015 Aug 19;7(301):301ra129

[E10] 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 Sep 1;180(5):388-95. doi: 10.1164/rccm.200903-0392OC. Erratum in: *Am J Respir Crit Care Med*. 2009 Oct 15;180(8):796.

[E11] Choy DF, Modrek B, Abbas AR, Kummerfeld S, Clark HF, Wu LC, Fedorowicz G, Modrusan Z, Fahy JV, Woodruff PG, Arron JR Gene expression patterns of Th2 inflammation and intercellular communication in asthmatic airways. *J Immunol*. 2011 Feb 1;186(3):1861-9

[E12] 16. Lum PY, Singh G, Lehman A, Ishkanov T, Vejdemo-Johansson M, Alagappan M, Carlsson J, Carlsson G. Extracting insights from the shape of complex data using topology. *Sci Rep*. 2013;3:1236. doi: 10.1038/srep01236.

[E13] Li L, Cheng WY, Glicksberg BS, Gottesman O, Tamler R, Chen R, Bottinger EP, Dudley JT. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci Transl Med*. 2015 Oct 28;7(311):311ra174.

482 [E14] Nielson JL, Paquette J, Liu AW, Guandique CF, Tovar CA, Inoue T, Irvine KA,
483 Gense JC, Kloke J, Petrossian TC, Lum PY, Carlsson GE, Manley GT, Young W, Beattie
484 MS, Bresnahan JC, Ferguson AR. Topological data analysis for discovery in preclinical
485 Spinal cord injury and traumatic brain injury. Nat Commun. 2015 Oct 14;6:8581

486

487 [E15] JavaPex: A research software package for persistent (co)homology, Proceedings of
488 ICMS 2014, Lecture Notes in Computer Science 8592, 129-136, 2014, Software available at
489 <http://appliedtopology.github.io/javaplex/>

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505 **SUPPLEMENTARY TABLES**

506 **Table E1a: Percentage missing clinic data stratified by GINA treatment strata.**

	Healthy control subjects	Patients with mild asthma (GINA 1)	Patients with moderate asthma (GINA 2-3)	Patients with severe asthma (GINA 4-5)
Total numbers	60	30	32	80
Age (yrs)	0.00	0.00	0.00	0.00
Sex (M/F)	0.00	10.00	0.00	0.00
GINA treatment steps	N/A	0.00	0.00	0.00
Age onset mean	N/A	16.67	6.25	2.50
Disease duration mean	N/A	16.67	6.25	1.25
Atopic status (%)	30.00	16.67	12.5	17.5
Annual exacerbation frequency	N/A	63.33	21.87	18.75
Skin tests positivity (%)	100.00	63.33	28.12	22.22
Aspergillus skin test positivity (%)	48.33	50.00	28.12	21.25
IgE median (IQR)	46.67	20.00	21.88	13.75
Smoking (pys)	20.00	16.67	3.12	1.25
BMI (kg/m ²)	65.00	43.33	25.00	28.75
FEV ₁ Pre BD (L)	81.67	16.67	0.00	0.00
FEV ₁ Pre BD %	28.33	16.67	0.00	0.00
FEV ₁ Post BD (L)	43.33	10.00	9.37	16.25
FEV ₁ Post BD (%)	40.00	10.00	9.37	16.25
FEV ₁ /FVC % Post BD	40.00	10.00	9.37	16.25
BDP equivalent ICS dose/24 hours	N/A	20.00	6.25	3.75
Oral Steroids (%)	N/A	13.33	0.00	0.00
Sputum eosinophils (%)	60.00	16.67	21.87	27.50

Blood eosinophils $\times 10^9$	56.67	3.33	9.37	5.00
Sputum neutrophils (%)	56.67	16.67	25.00	27.50
PC ₂₀ Methacholine (mg/ml)	20.00	10.00	9.37	28.75
FEV ₁ BDR %	35.00	6.67	9.37	17.50
FE _{NO} 50 mls s ⁻¹ (ppb)	38.33	13.33	18.75	27.50

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

Table E1b: Percentage missing pathology data stratified by GINA treatment strata in the discovery cohort.

	Healthy control subjects	Patients with mild asthma (GINA 1)	Patients with moderate asthma (GINA 2-3)	Patients with severe asthma (GINA 4-5)
total numbers	60	30	32	80
Remodelling parameters				
Φ reticular basement membrane thickness	15.00	6.67	6.25	7.70
Φ epithelial percentage	23.33	0.00	6.25	3.75
airway smooth muscle area	21.67	16.67	6.25	5.00
Φ airway smooth muscle percentage	25.00	0.00	6.25	5.00
Glands percentage	25.00	0.00	6.25	5.00
Chalkey count	46.67	60.00	37.50	31.25
Smooth muscle cells/mm²				
Tryptase+ mast cells in the airway smooth muscle	45.00	40.00	37.50	35.00
Lamina propria cells /mm²				
Tryptase+ mast cells	15.00	3.33	6.25	12.50
Φ major basic protein+ eosinophils	11.67	3.33	12.50	7.50
neutrophil elastase+ neutrophils	11.67	0.00	6.25	12.50

Φ Highlighted in blue are the input variable derived from factor analyses to generate the reported TDA groups

531 **Table E1c: Percentage missing pathology data stratified by TDA grouping in the**
532 **discovery cohort.**

	Group 1 [n=77]	Group 2 [n=60]	Group 3 [n=36]	Group 4 [n=19]
Age (yrs.)	0.00	0.00	0.00	0.00
Sex (M:F)	6.49	0.00	8.33	0.00
BMI (kg/m ²)	42.86	31.67	38.89	47.37
GINA treatment steps ^^	0.00	0.00	0.00	0.00
Age of onset (years) \$	N/A	15.00	25.00	31.58
Disease duration (years) \$	N/A	15.00	25.00	31.58
Atopic status (%)	12.99	11.67	13.89	47.37
Aspergillus skin test positive (%)	36.36	13.33	22.22	31.58
Total IgE (KU)	32.47	16.67	25.00	26.32
Smoking (pys)	15.58	1.67	16.67	0.00
Annual exacerbation frequency \$	N/A	26.67	55.56	42.11
BDP equivalent ICS dose/24 hours	53.25	15.00	22.22	36.84
Oral Steroids (%)	49.35	13.33	22.22	31.58
Sputum eosinophils (%)	45.45	21.67	22.22	42.11
Blood eosinophils x10 ⁹	20.78	3.33	25.00	5.26

FE _{NO} 50 mls s ⁻¹ (ppb)	23.38	26.67	25.00	31.58
Sputum neutrophils (%)	42.86	23.33	22.22	42.11
FEV ₁ Pre BD(L)	14.29	1.67	22.22	0.00
FEV ₁ Pre BD %	15.58	1.67	22.22	0.00
FEV ₁ Post BD (L)	23.38	16.67	16.67	31.58
FEV ₁ Post BD (%)	23.38	13.33	13.89	31.58
FEV ₁ /FVC % Post BD	23.38	15.00	19.44	31.58
PC ₂₀ Methacholine (mg/ml)	14.29	18.33	19.44	31.58
FEV ₁ BDR %	15.58	15.00	13.89	42.11

§: Not applicable to TDA group 1 as this group primarily consisted of a large number of healthy volunteers, in whom these parameters would not be acquired. ^^: GINA treatment step was set to 0 in healthy volunteers.

546 **E1d: Percentage missing pathology data stratified by TDA Group in the discovery**
547 **cohort.**

	Group 1 [n=77]	Group 2 [n=60]	Group 3 [n=36]	Group 4 [n=19]
Remodelling parameters				
Φ Reticular basement membrane thickness (µm)	15.58	1.67	0.00	5.26
Φ Epithelial area (%)	15.58	0.00	0.00	5.26
Airway smooth muscle area (mm ²)	22.08	0.00	0.00	5.26
Φ Airway smooth muscle area percentage (%)	20.78	0.00	0.00	5.26
Glands percentage	18.18	0.00	0.00	5.26
Chalkey count	48.05	28.33	47.22	26.32
Smooth muscle cells/mm²				
Tryptase ⁺ mast cells in the airway smooth muscle	51.95	36.67	5.56	63.16
Lamina propria cells /mm²				
Tryptase ⁺ mast cells	12.99	8.33	2.78	5.26
Φ Major basic protein ⁺ eosinophils	14.29	1.67	0.00	5.26
Neutrophil elastase ⁺ neutrophils	10.39	6.67	0.00	10.53

548 Φ Highlighted in blue are the input variable derived from factor analyses to generate the
549 reported TDA groups

550

Table E2: Between observer intra-class correlation coefficients (ICC) for the pathology features in the discovery cohort

Pathology Feature	Intra class correlation coefficient (single measures)	Intra class correlation coefficient (average measures)
Reticular basement membrane thickness Number sampled (n=13)	0.75	0.86
Epithelial % Area Number sampled (n=13)	0.60	0.75
Mean Chalkley Count Number sampled (n=13)	0.98	0.99
% Airway Smooth Muscle Area Number sampled (n=13)	0.62	0.77
% Glands Number sampled (n=13)	1	1
Lamina propria eosinophils/mm² Number sampled (n=15)	0.97	0.99
Lamina propria neutrophils/mm² Number sampled (n=15)	0.97	0.98
Lamina propria mast cells/mm² Number sampled (n=16)	0.99	0.99
Mast cells/mm² airway smooth muscle Number sampled (n=16)	0.80	0.89

Table E3a: Linear mixed models of Pre Bronchodilator FEV₁ decline

Model	Fixed	Random	AIC	BIC
1	Age,Height,Sex,Time	Intercept,Time	830.80	838.12
2	Age,Height,Sex,Time	Intercept	865.35	870.24

AIC – Aikake information criteria, BIC – Bayesian information criteria

Table E3b: Linear mixed model output for model 1 in Table 3a

	Estimate	Standard Error	t Value	p-value	95% CI Lower	95% CI Upper
Intercept	-3.9048	1.5246	-2.56	0.0123	-6.9387	-0.8708
Time	-0.01346	0.008467	-1.59	0.1163	-0.03035	0.003425
Age	-0.02153	0.005004	-4.30	<.0001	-0.03135	-0.01171
Height	0.04447	0.008700	5.11	<.0001	0.02739	0.06155
Sex[Female]	0.04141	0.1730	0.24	0.8109	-0.2983	0.3811
Sex[Male]	0	.	.	.	.	.

567 **Table E4a: Clinical characteristics of the discovery population stratified by GINA**
568 **treatment intensity.**

	Healthy control subjects [n=60]	Patients with mild asthma (GINA 1) [n=30]	Patients with moderate asthma (GINA 2-3) [n=32]	Patients with severe asthma (GINA 4-5) [n=80]	p value†
Age (years), mean ± SEM	36.03 ± 1.91	40.50 ± 2.82	41.72 ± 2.97	45.62* ± 1.33	0.0016
Sex (M/F)	29/26	13/14	15/17	46/35	0.8054 [§]
BMI (kg/m ²)	24.10 (4.7)	28.1 (7.45)*	27.0 (4.8)	28.10 (8.5)*	0.0075
Age onset (years) mean ± SEM	N/A	26.48 ± 4.55	27.57 ± 9.92	25.04 ± 2.14	0.8327
Disease duration (years) ± SEM	N/A	14.89 ± 2.69	13.62 ± 2.40	20.52 ± 1.62	0.0386
Atopic status (%)	38.1	76.1	84.6	69.7	0.0002[§]
Annual exacerbation frequency	N/A	0 (0)	0 (0.5)	2 (4)	<0.0001
Aspergillus Fumigatus skin test positive (%)	0	20	14.3	20.6	0.7290 [§]
Total IgE (KU)	32.30 (95.6)	95.70 (183.17)	201.0 (204.5)*	194.0 (400.1)*	0.0007
Smoking (pack years)	0 (0)	0 (0)	0 (2.5)	0 (2)	0.0905
FEV ₁ Pre BD (L) s ± SEM	5.31 ± 1.7	3.03 ± 0.21	2.97 ± 0.17	2.64 ± 0.08	0.0877
FEV ₁ Pre BD % ± SEM	97.09 ± 4.4	93.09 ± 3.09	91.92 ± 4.46	81.69 ± 2.13*	0.0021
FEV ₁ Post BD (L) ± SEM	6.07 ± 2.78	3.2 ± 0.20	3.20 ± 0.15	2.82 ± 0.09	0.0887
FEV ₁ Post BD % ± SEM	98.46 ± 4.32	97.95 ± 2.91	98.40 ± 2.45	86.36* ± 2.19	0.0022
FEV ₁ /FVC Post BD ± SEM	82.98 ± 1.5	76.78 ± 2.08*	77.13 ± 1.60	72.22 ± 1.30*	<0.0001
FEV ₁ BDR %	0 (1)	5.89 (11.5)*	4.88 (6.06)*	2.65 (10.29)*	0.0003
PC ₂₀ Methacholine (mg/ml)	32(0)	0.58 (2.82)*	1.50 (3.19)*	0.80 (2.19)*	0.3661

569

570 All values not presented as mean ± SEM are presented as median (IQR). BMI = Body Mass
571 Index. BD =Bronchodilator, BDR = Bronchodilator Reversibility. † Statistical analysis across
572 all groups unless otherwise stated. *p < 0.05 compared with healthy control subjects. §

Moderate compared with severe asthma. Adjusted p-values obtained using Bonferonni correction All tests are Kruskal-Wallis unless indicated otherwise. \$=Chi-Squared test.

575

Table E4b: Treatment and Inflammatory profile of the discovery population stratified by GINA treatment intensity

	Healthy control subjects	Patients with mild asthma (GINA 1) [n=30]	Patients with moderate asthma (GINA 2-3) [n=32]	Patients with severe asthma (GINA 4-5) [n=80]	p value†
BDP equivalent ICS dose/24 hours ¥	N/A	0(0)	775 (450) [§]	1600 (1000)	<0.0001
Oral steroids (%)	N/A	0	0	37	<0.0001[§]
Sputum eosinophils (%)	0.25 (0.75)	2.4 (5.25) *	3.8 (14)*	4.0 (15.81)*	<0.0001
Blood eosinophils x10 ⁹	0.11 (0.12)	0.28 (0.28)*	0.24 (0.24)*	0.29 (0.42)*	<0.0001
Sputum neutrophils (%)	54.28 (55.09)	54.00 (58)	47.10 (53.67)	59.00 (46.29)	0.6409
FE _{NO} 50 mls s ⁻¹ (ppb)	14 (16.5)	19.15 (27.25)	27.95 (23.55)*	29.25 (30.93)*	0.0001

578

† Statistical analysis across all groups unless otherwise stated. *p < 0.05 compared with healthy control subjects. ¥ Ratio for Budesonide Turbuhaler calculated as 1.5. § Moderate compared with severe asthma. Adjusted p-values obtained using Bonferonni correction. All tests are Kruskal-Wallis unless indicated otherwise. \$=Chi-Squared test.

583

584

585

586

587

588 **Table E5: Pathological features of the discovery population stratified by GINA**
589 **treatment intensity**

	Healthy control subjects	Patients with mild asthma (GINA 1)	Patients with moderate asthma (GINA 2-3)	Patients with severe asthma (GINA 4-5)	p value†
Total numbers	60	30	32	80	
Total biopsy area (mm ²)	1.52 (1.02)	1.52 (1.64)	1.68 (1.4)	1.83 (1.54)	0.22
Reticular basement membrane thickness (µm)	6.86 (3.58)	9.3 (5.53)*	9.32 (4.8)*	11.41 (4.45)*	< 0.0001
Epithelial percentage area	7.53 (5.57)	6.64 (9.15)	5.87 (8.79)	10.00 (11.46)	0.1344
Airway smooth muscle area (mm ²)	0.14 (0.24)	0.16 (0.29)	0.18 (0.34)	0.24 (0.44)	0.0677
Airway smooth muscle % area	11.50 (15.61)	12.82 (22.47)	14.38 (20.51)	14.82 (19.7)	0.7068
Glands %	0 (0)	0 (0.28)	0 (0)	0 (0)	0.6577
Mean Chalkley Count	3.4 (1.19)	4.65 (2.57) *	4.15 (1.43)	5 (2)*	< 0.0001
<i>Smooth muscle cells/mm²</i>					
Tryptase ⁺ mast cells	1.64 (6.19)	4.17 (8.4)*	5.77 (13.65)*	4.01 (8.09)	0.011
<i>Lamina propria cells /mm²</i>					
Tryptase ⁺ mast cells	14.93 (15.31)	19.02 (18.67)	19.26 (24.34)	11.64 (15.04)	0.117
Major Basic Protein ⁺	2.78 (10.90)	14.22	11.69 (16.66)*	10.98	< 0.0001

eosinophils		(21.9)*		(20.23)*	
Neutrophil Elastase ⁺ neutrophils	8.82 (17.35)	9.54 (18.52)	10.03 (14.78)	10.65 (14.76)	0.6380
Th2 gene expression score	0.21 (0.38)	0.91 (2.42)	-0.77 (2.07)*	-0.88 (1.19)	0.0208
Th17 gene expression score	-0.066(0.57)	-0.73 (0.49)	-0.001 (0.99)	-0.27 (2.00)	0.2639
FKBP51 gene expression score	0.0286(0.19)	0.0133(0.27)	0.440(0.18)*	0.395(0.44)*	< 0.0001

590

591 All values presented as median (IQR). † Statistical analysis across all groups unless otherwise
592 stated. *p < 0.05 compared with healthy control subjects. Adjusted p-values obtained using
593 Bonferroni correction All tests are Kruskal-Wallis unless indicated otherwise.

594

595

596

597

598

599

600

601

602

603

604

605

606

607

608 **Table E6: 90th percentile point analysis of pathology features stratified by GINA**
609 **treatment intensity**

Remodelling parameters	90 th Percentile Value in Healthy Volunteers	GINA 1 Asthma % > 90 th percentile point	GINA 2-3 Asthma % > 90 th percentile point	GINA 4-5 Asthma % > 90 th percentile point	Chi-square test [^]
Reticular basement membrane thickness (µm)	10.81	34.48	30.00	59.21	<0.0001
Epithelial area (%)	18.21	13.33	14.28	19.23	0.4246
Airway smooth muscle area (mm ²)	0.3857	16	29.03	36.36	0.0037
Airway smooth muscle percentage (%)	35.34	12	12.90	9.09	0.8912
Submucosal glands (%)	1.762	10	3.22	11.69	0.5924
Mean Chalkey count (units)	4.440	66.67	37.50	40.74	<0.0001
Airway smooth muscle cells/mm ²					
Tryptase+ mast cells in the airway smooth muscle	7.236	33.33	44.44	33.96	0.0282
Lamina propria cells /mm ²					
Tryptase+ mast cells	34.68	6.90	25.00	9.86	0.1128
Major basic protein+ eosinophils	18.54	44.83	25.00	34.67	0.0018
Neutrophil elastase+ neutrophils	27.81	13.33	24.14	18.31	0.3087

[^] Comparison of proportion across all four groups (healthy controls set at 10%)

612 **Table E7: Replicated Biopsy Gene Expression Markers in Severe Asthma**

613

Gene	Genomic Location		¹ Entrez Gene/ ² Gene Card summary/ ³ PubMed asthma search results
RPL23AP32	Chr:2p	Ribosomal Protein L23a Pseudogene 32	² RPL23AP32 (Ribosomal Protein L23a Pseudogene 32) is a Pseudogene, and is affiliated with the lncRNA class. ³ No relevant published papers related to asthma.
YME1L1	Chr:10p	YME1 Like 1 ATPase	¹ The protein encoded by this gene is the human ortholog of yeast mitochondrial AAA metalloprotease, Yme1p. It is localized in the mitochondria and can functionally complement a yme1 disruptant yeast strain. It is proposed that this gene plays a role in mitochondrial protein metabolism and could be involved in mitochondrial pathologies. ³ No relevant published papers related to asthma.
PHACTR3	Chr: 20q	Phosphatase And Actin Regulator 3	¹ This gene encodes a member of the phosphatase and actin regulator protein family. The encoded protein is associated with the nuclear scaffold in proliferating cells, and binds to actin and the catalytic subunit of protein phosphatase-1, suggesting that it functions as a regulatory subunit of protein phosphatase-1. ³ No relevant published papers related to asthma.
ROBO3	Chr:11q	Roundabout Guidance Receptor 3	¹ This gene is a member of the Roundabout (ROBO) gene family that controls neurite outgrowth, growth cone guidance, and axon fasciculation. ROBO proteins are a subfamily of the immunoglobulin transmembrane receptor superfamily. SLIT proteins 1-3, a family of secreted chemorepellants, are ligands for ROBO proteins and SLIT/ROBO interactions regulate myogenesis, leukocyte migration, kidney morphogenesis, angiogenesis, and vasculogenesis in addition to neurogenesis. ³ No relevant published papers related to asthma.
SLC25A16	Chr:10q	Solute Carrier Family 25 Member 16	¹ This gene encodes a protein that contains three tandemly repeated mitochondrial carrier protein domains. The encoded protein is localized in the inner membrane and facilitates the rapid transport and exchange of molecules between

			the cytosol and the mitochondrial matrix space. ² This gene has a possible role in Graves' disease. SLC25A16 (Solute Carrier Family 25 Member 16) is a Protein Coding gene. Diseases associated with SLC25A16 include Graves' Disease and Subclavian Artery Aneurysm. Among its related pathways are Metabolism of water-soluble vitamins and cofactors. ³ No relevant published papers related to asthma.
KRT24	Chr:17q	Cytokeratin-24	¹ This gene encodes a member of the type I (acidic) keratin family, which belongs to the superfamily of intermediate filament (IF) proteins. Keratins are heteropolymeric structural proteins which form the intermediate filament. These filaments, along with actin microfilaments and microtubules, compose the cytoskeleton of epithelial cells. ³ No relevant published papers related to asthma.
TMC6	Chr:17q	Transmembrane Channel Like 6	² TMC6 (Transmembrane Channel Like 6) is a Protein Coding gene. Diseases associated with TMC6 include Epidermodysplasia Verruciformis and Superficial Mycosis. ³ No relevant published papers related to asthma.
PKD1P6	Chr:16p	Polycystin 1, Transient Receptor Potential Channel Interacting Pseudogene 6	² PKD1P6 (Polycystin 1, Transient Receptor Potential Channel Interacting Pseudogene 6) is a Pseudogene. Diseases associated with PKD1P6 include Polycystic Kidney Disease. ³ No relevant published papers related to asthma.
FKBP5	Chr:6p	FK506 Binding Protein 5	² FKBP5 (FK506 Binding Protein 5) is a Protein Coding gene. Diseases associated with FKBP5 include Major Depressive Disorder And Accelerated Response To Antidepressant Drug Treatment and Post-Traumatic Stress Disorder. Among its related pathways are Glucocorticoid receptor regulatory network and TNF-alpha/NF-kB Signalling Pathway. FKBP51, is up-regulated in the airways of allergen-challenged asthmatic subjects taking inhaled budesonide when compared to placebo [37]. Increased bronchial epithelial gene expression of FK506 binding protein (FKBP5), correlated with Actinobacteria (16sRNA) in the airway epithelium [38]. Increased bronchial biopsy gene expression in the UBIOPRED severe asthma cohort when compared to healthy controls [39].

IGLJ3	Chr: 22q	Immunoglobulin Lambda Joining 3	² IGLJ3 (Immunoglobulin Lambda Joining 3) is a Protein Coding gene. ³ No relevant published papers related to asthma.
PHACTR1	Chr:20q	Phosphatase And Actin Regulator 1	¹ The protein encoded by this gene is a member of the phosphatase and actin regulator family of proteins. This family member can bind actin and regulate the reorganization of the actin cytoskeleton. It plays a role in tubule formation and in endothelial cell survival. Polymorphisms in this gene are associated with susceptibility to myocardial infarction, coronary artery disease and cervical artery dissection. ³ No relevant published papers related to asthma.

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628 **Table E8a: Pathological Characteristics of the Severe Asthma Fraction: Discovery and**
629 **Replication (uncorrected for tissue shrinkage in paraffin)**

630

	Replication	Discovery	
--	--------------------	------------------	--

	N=64	N=80	
Remodelling parameters			
Reticular basement membrane thickness (µm)	8.87[3.11]	11.405[4.37]	<0.0001
Epithelial area (%)	8.5[10.3]	10[11.13]	0.979
Airway smooth muscle area (Replication: mm²/mm, Discovery: % total biopsy area)	0.02935[0.0192]	0.24233[0.433]	N/A [^]
Lamina propria cells /mm²			
Major basic protein ⁺ eosinophils	22.47[34.3]	10.885[19.23]	0.005817
Th2 gene expression score	-0.07 [3.00]	-0.876 [1.31]	N/A\$\$
Th17 gene expression score	0.08[2.09]	-0.27 [2.06]	N/A\$\$
FKBP51 gene expression	1.658[2.20]	0.40[0.47]	N/A\$\$

All values are presented as median (IQR). † Statistical analysis between Severe Asthma Fraction Discovery and Replication. ^ N/A as the two measurements of airway smooth muscle mass cannot be directly compared as they are expressed differently in both cohorts. .
\$\$ Direct comparison is not possible between cohorts due to the possibility of batch effects which were not corrected for in the statistical analysis.

Table E8b: Pathological Characteristics of the Severe Asthma Fraction: Discovery and Replication (Corrected for paraffin embedding)

	Replication N=64	Discovery N=80	p value
Remodelling parameters			

Reticular basement membrane thickness (μm)	11.45(3.8)	11.52 (4.37)	0.5528
Log ₁₀ Epithelial area (%)	1.05(0.45)	0.95(0.42)	0.2986
Airway smooth muscle area Replication (mm ² /mm), Discovery: % total biopsy area	0.02935(0.0192)	0.24233(0.433)	N/A \$
Lamina propria cells /mm²			
Log ₁₀ Major basic protein ⁺ eosinophils	1.08(0.68)	1.07(0.67)	0.6217

All values are presented as median (IQR). † Statistical analysis between Severe Asthma Fraction Discovery and Replication. All tests are Kruskal-Wallis. ¶ Due to paraffin embedding apply test using length and area corrected variables. \$ Direct comparisons are not possible due to the airway smooth muscle area being expressed as a function of basement membrane length in the replication cohort and total biopsy area in the discovery cohort.

Table E8c: Clinical Characteristics of the Severe Asthma Fraction: Discovery and Replication

	Replication	Discovery	
--	-------------	-----------	--

	N=64	N=80	
Age (yrs.), mean $\pm$ SEM	47.21 $\pm$ 1.4	45.53 $\pm$ 1.35	0.479
Sex (% Male)	41.8%	60%	0.073 [‡]
GINA treatment steps	4[n=53],5[n=11]	4[n=47],5[n=33]	0.003^{\$}
Number on Oral Steroids	11/64	33/80	0.0019[‡]
Blood eosinophils $\times 10^9$	0.235[0.28]	0.29[0.39]	0.091
Sputum eosinophils (%)	5.7[14.875]	4[15.063]	0.821
FE _{NO} 50 mls s ⁻¹ (ppb)	25.5[18.38]	29.25[29.83]	0.416
Sputum neutrophils (%)	51[31.1]	59[44.11]	0.304
FEV ₁ Pre BD % $\pm$ SEM	60.16 $\pm$ 1.35	81.69 $\pm$ 2.13	<0.0001
FEV ₁ BDR %	14.93[14.93]	2.65[9.59]	<0.0001

All values not presented as mean $\pm$ SEM are presented as median (IQR). † Statistical analysis between Severe Asthma Fraction Discovery and Replication. All tests are Kruskal-Wallis, except ‡=Two proportion Z-test and \$=Chi Squared Test.

SUPPLEMENTARY FIGURE LEGENDS

Figure E1

Schematic outlining the discovery and replication population cohorts. The discovery population consisted of 142 patients with adult asthma as annotated in the figure (orange ellipse), alongside the number of patients at each GINA treatment step. As the cohort were derived from patients attending both Leicester and Belfast severe asthma cohorts, central

quality control and assurance measures were taken to ensure measurement standardisation across cohorts (**see supplementary material**). Additionally 60 aged matched controls without asthma were included in the discovery cohort. The replication cohort was derived from the BOBCAT study (20), with 59 patients with severe asthma derived from this cohort. Bronchial mucosal biopsies were embedded in GMA in the discovery cohort and paraffin in the replication cohort. The text box identifies additional phenotyping data acquired in the respective cohorts. This included whole biopsy gene expression analysis performed using two colour Whole Human Genome (WHG) 4x44k gene expression microarray analysis, in subsets of asthmatics in both cohorts and healthy volunteers in the discovery cohort, lung function decline data in a subset of asthma patients in the discovery cohort only, and clinical meta data in both cohorts (tables of % completeness of this data are provided in the **supplementary tables E1 a-d**).

Figure E2

Photomicrographs of the nine pathological features evaluated in the discovery cohort across GINA treatment steps. Percentage (a) epithelium, (b) smooth muscle and (c) glandular areas of the total biopsy area, (d) RBM thickness (black arrow) (e) EN4+ immunostained vessels. Numbers of positively stained nucleated (f) MBP+ eosinophils, (g) NE+ neutrophils and (h) tryptase+ mast cells in the lamina propria and (i) tryptase+ mast cells within the ASM bundle areas were counted and expressed per mm². Pathology features (i) % airway epithelial area, (ii) % airway smooth muscle area, (iii) eosinophils /mm² of the lamina propria and (iv) RBM thickness were used as input parameters into subsequent topological data analysis after data reduction using factor analysis.

** Represents variables that were selected post factor analysis as input functions for topological data analysis.*

Figure E3

Applying mapper algorithm (Ayasdi) to a set of continuous data. See narrative description in the online supplemental text for further details.

E3A: Geometric association and motivation from a data matrix of real numbers.

E3B: Simple demonstration of ‘mapper’ algorithm to achieve the aforementioned goal.

E3C: Metric and lenses used in this study.

In this study, we applied the mapper algorithm, using norm correlation metric and the neighbourhood lenses, to a data set using 4 pathological features.

Figure E4

Fundamental ideas behind persistent homology, shapes can be characterised using certain properties, connected components, loops and voids. **Figure E4a:** a circle can be described by a set of connected components and a loop. **Figure E4b-e:** For the example data set by varying the radius ϵ , we observe the birth and death of topological features such as connected

components, loops and voids. **Figure E4f:** Illustrates the construction of simplices by connecting points that are a distance of $d \leq \epsilon$ apart. The Vietoris-Rips complex is constructed by joining simplices formed by considering only pairwise connections. **Figure E4g:** barcode plot for the example data set E1b-e above. The barcode plots provide a topological summary of the data, the endpoints of a bar represent the birth and death times of the feature, hence long bars represent more persistent features, short bars represent noise. For the exemplar data from a circle with unit radius, long bars represent persistent features short bars transient features that can be regarded as topological noise. The barcode shows that the sample data is characterised by a single connected component and a loop.

Figure E5

Analysis of pathology features in asthmatic bronchial biopsies according to GINA treatment intensity strata [Mild (GINA 1), Moderate (GINA 2-3), Severe (GINA 4-5)]. It can be seen that patients with high levels of corticosteroid exposure (GINA 4/5) continue to have basement membrane and vascular remodelling and demonstrate a refractory eosinophilic tissue phenotype. * Indicate intergroup comparisons with a p value of < 0.05 .

Figure E6

Gene expression data in the discovery and replication cohorts using two colour Whole Human Genome (WHG) 4x44k gene expression microarray analysis. For consistency, patients with GINA3-5 treatment steps asthma were evaluated in the discovery cohort (as the replication cohort comprised entirely of GINA 4 (primarily) and GINA 5 patients). All probes were ranked for association in order of p value, with corresponding log ratio of expression. The top 20 genes (based upon p value of association) were 'pulled through' into the replication cohort (BOBCAT) where a more limited analysis of differential expression against discovery healthy volunteers was under taken for these selected probes only – correcting for 20 comparisons. The figure demonstrates a heat map of expression (log ratio values) of the 11/20 replicated genes that reached significance after correction for multiple comparisons in the replication cohort. Individual columns represent individual patients, rows represent the 11 genes which are annotated.

**** Identifies three genes selected based upon a priori knowledge (FKBP5 a potential marker of adherence and corticosteroid exposure), potential mechanistic relevance to phenotype (KRT24) and potential role in epigenetic modification following treatment with steroids in asthma (PHACTR-3) for further annotation in subsequent topological data analysis networks.**

Figure E7

A subgroup of asthmatic patients (87/142) in the discovery population had longitudinal lung function (pre bronchodilator FEV₁) data acquired following bronchial biopsy sampling. Median (Q1-Q3) follow up duration was 2.73[5.34-0.77] years. A linear mixed model was fitted to evaluate patient specific lung function change trajectories. Fixed effects in the model were: Age, Height, Sex and Time. Random effects in the model were: Intercept and Time for each patient. Average change in FEV₁ over 1 year for each patient was determined as $\beta_t + \gamma_{ti}$ where β_t is the fixed effect for time and γ_{ti} is the random effect for time for patient i . It can be seen that overall the median decline slopes in mls/year centre at zero but with wide quartile 1: quartile 3 ranges across TDA groups.

Figure E8

Barcode graphs for discovery and replication, the endpoints of a bar represent the birth and death times of the feature, hence long bars represent more persistent features, short bars represent noise. The black bars represent connected components, the red bars loops, the x-axis is labelled time, this reflects the birth and death of features as the spatial resolution is increased over time. Increasing similarity of the discovery (**a, c**) and replication cohorts (**b, d**) can be seen after correction for tissue shrinkage in paraffin in the replication cohort and log₁₀ normalisation of non-normally distributed pathology features (**c vs d**). After this series of corrections the two populations appear to have highly similar level 0 which can be seen visually when inspecting the filtration profile of level 0 homologies and for level 1 homologies is annotated with the red bounded shapes. Further details on the theory of barcode graphs can be found in **supplementary Figure E4** and in the supplementary material.

Figure E9

TDA networks generated using the neighbourhood 1 and 2 lens and norm correlation metric in the discovery population of severe GINA treatment step 4/5 asthmatics only (**E9a**) and replication population (**E9b**). 4 pathological input parameters [ASM% area, reticular basement membrane thickness, epithelial area, ASM% area, and eosinophils/mm² lamina propria- left to right] derived from factor analysis of the 9 pathological features were utilised to generate the TDA networks. The networks were created after correcting the replication data for tissue shrinkage and log₁₀+1 transformation of two of the pathology parameters due to a non-normal distribution. In both network regions of high ASM mass, lamina propria eosinophils, basement membrane thickening and epithelial area can be seen. However these regions appear to visually overlap differently in discovery and replication networks. This highlights the challenges of ‘visual two dimensional replication’ of multi-dimensional pathological networks. In both the discovery and replication networks, Th2 and Th12 gene expression appear to show a reciprocal relationship, indicating that this relationship is conserved in severe asthma and a repeatable observation.

DISCOVERY

126 Asthma
50 controls
Leicester

16 Asthma
10 controls
Belfast

N=142 Asthma
N= 60 healthy volunteers

GINA 1= 30
GINA 2-3 = 32
GINA 4-5 = 80 (GINA 5 =32)
Median Q3:Q1 GINA 4[4-3]

GMA

REPLICATION

N=59 Asthma

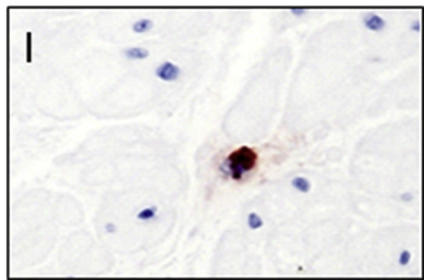
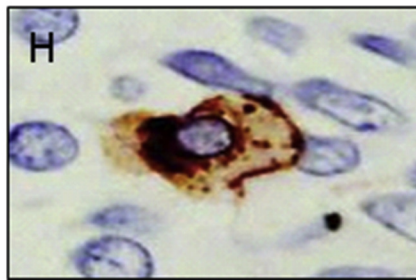
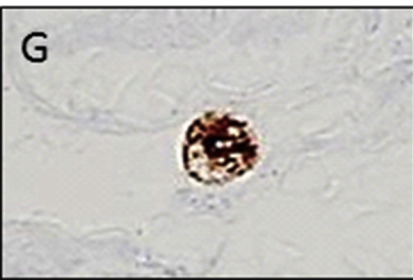
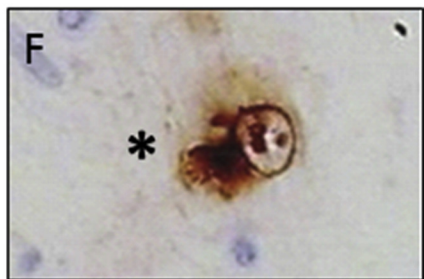
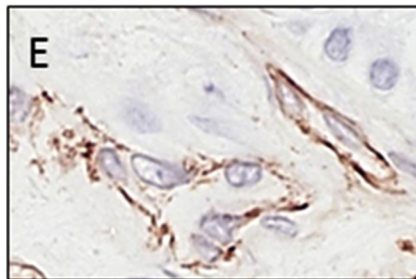
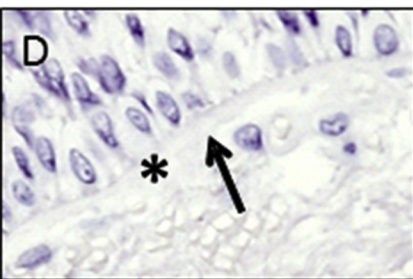
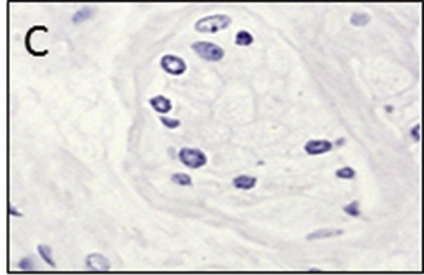
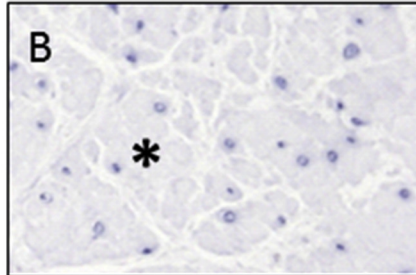
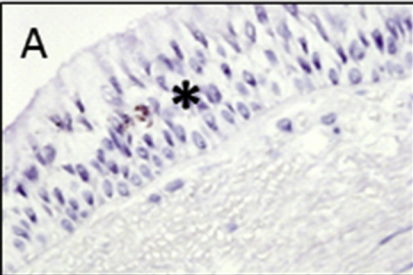
BOBCAT

Median Q1:Q3
GINA 4[4-4]

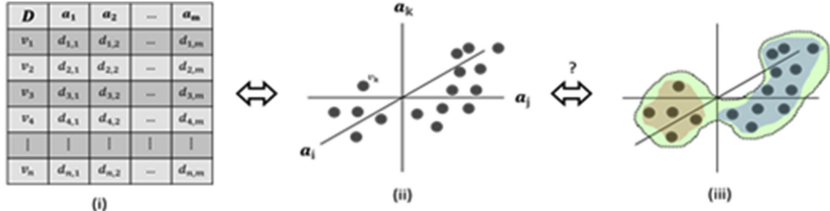
Paraffin



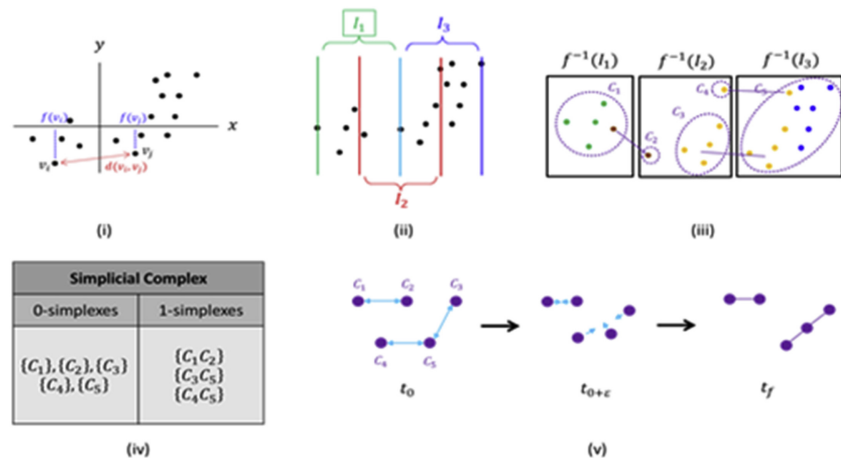
- Immunopathology
- Clinical meta data
- Tissue gene expression subset [n=48/142 Asthma, 19/60 healthy volunteers: discovery, n=33/59 BOBCAT: replication]
- FEV₁ declinesubset: discovery only [n=67/142]



A



B



C

$$v_i = (v_{i,1}, \dots, v_{i,n}) \quad v_j = (v_{j,1}, \dots, v_{j,n}) \quad d(v_i, v_j) = 1 - r(\hat{v}_i, \hat{v}_j)$$

$$r(x, y) = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \sqrt{n \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2}}$$

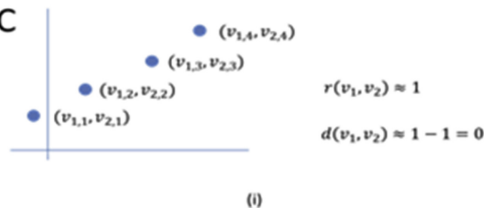
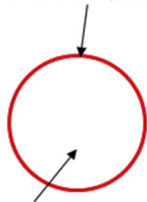


Figure E4a

Connected Components



1 Dimensional Hole

Figure E4b

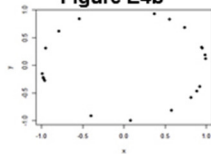


Figure E4c

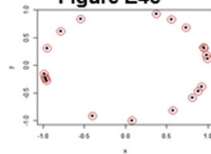


Figure E4d

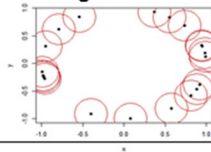


Figure E4e

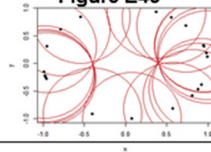
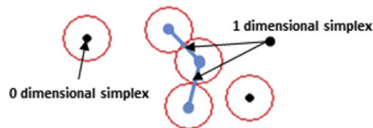
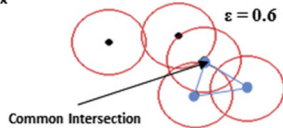


Figure E4f

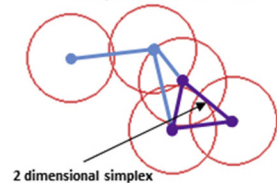
Simplices, radius $\varepsilon = 0.11$



Simplices included in Čech complex



Simplices, radius $\varepsilon = 0.6$



Simplices included in Vietoris-Rips complex

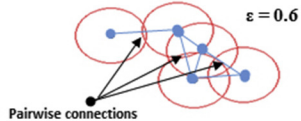
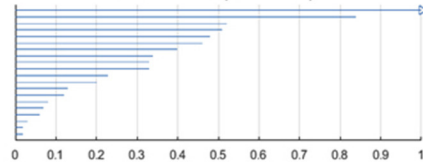
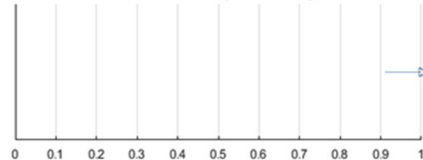


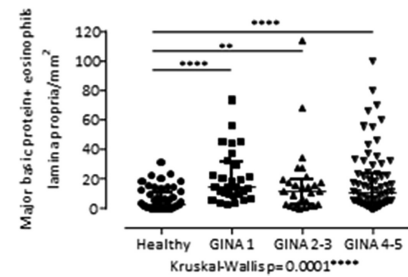
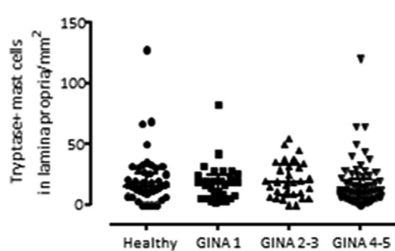
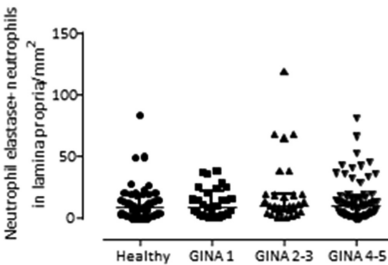
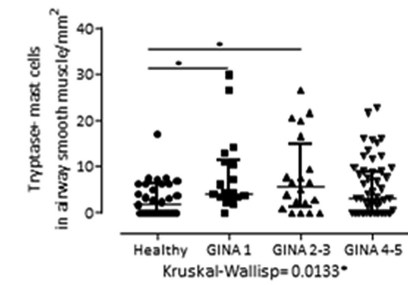
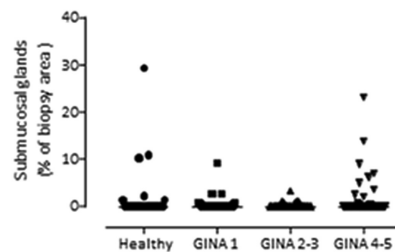
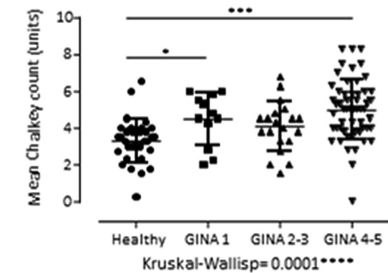
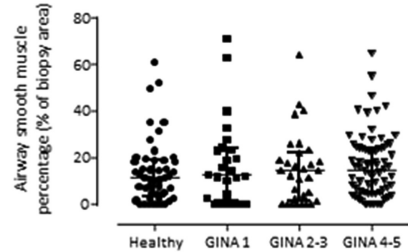
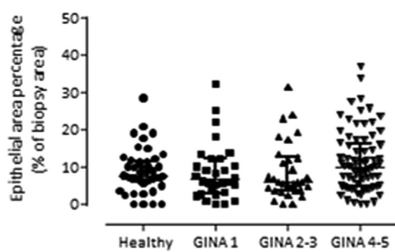
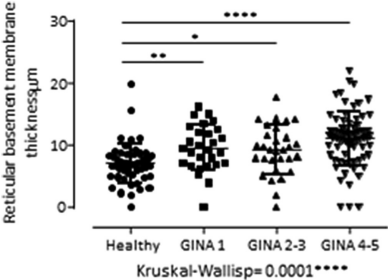
Figure E4g

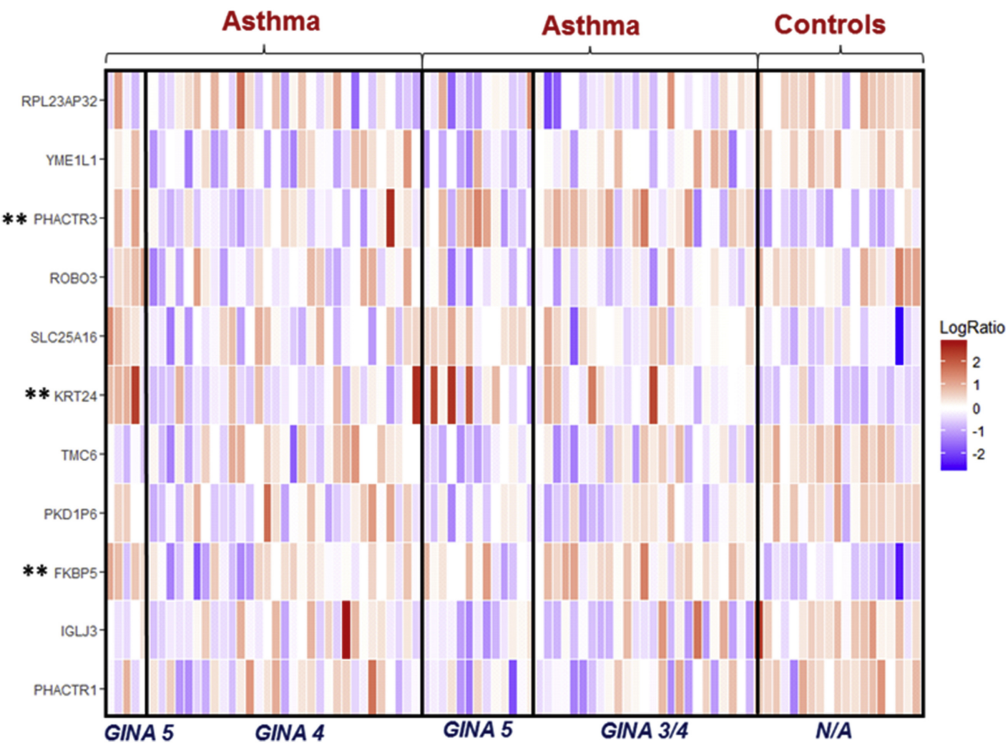
Uniform Circle (dimension 0)

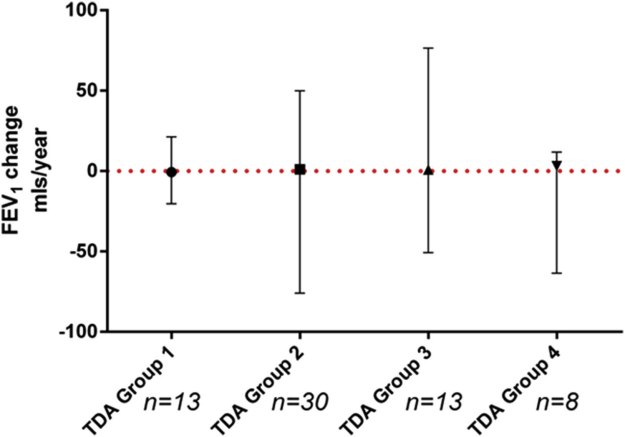


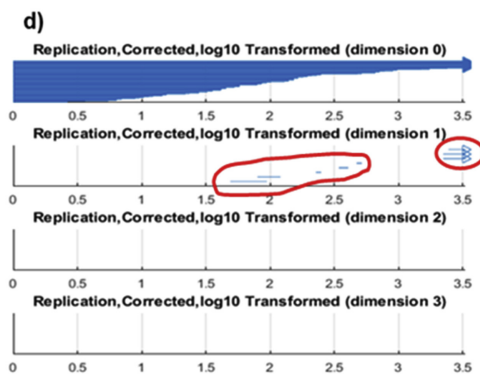
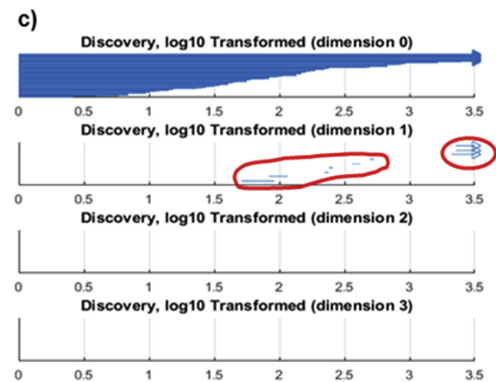
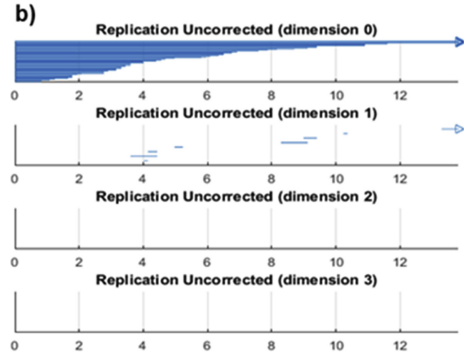
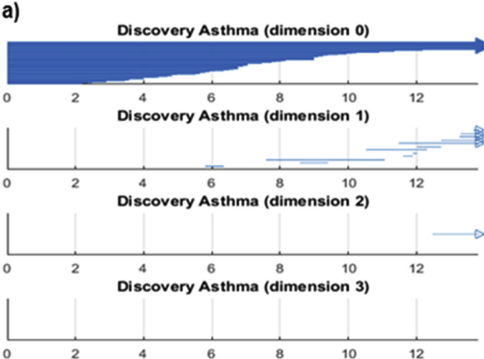
Uniform Circle (dimension 1)

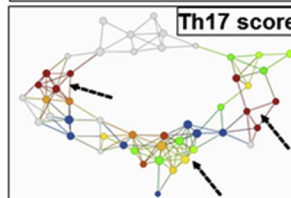
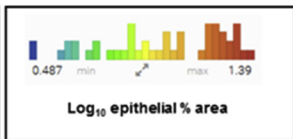
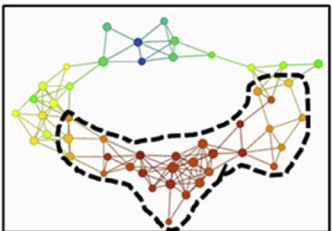
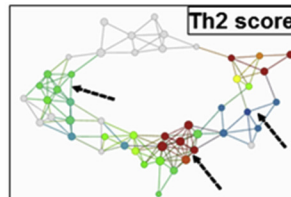
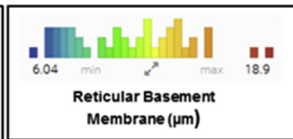
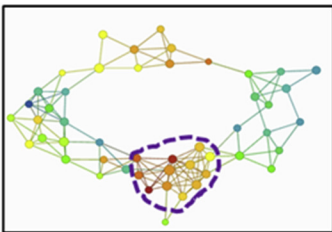
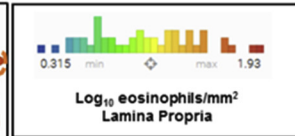
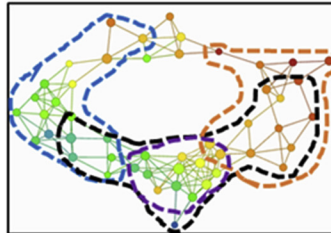
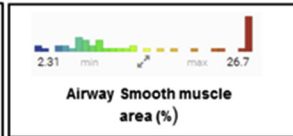
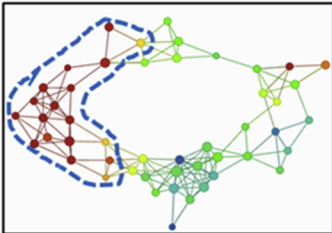




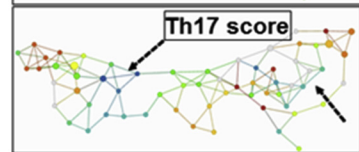
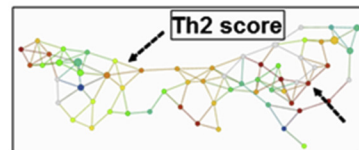
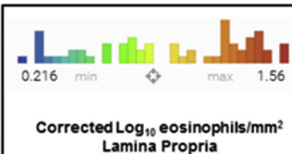
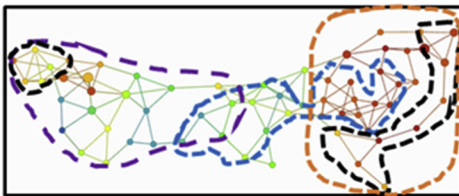
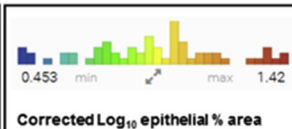
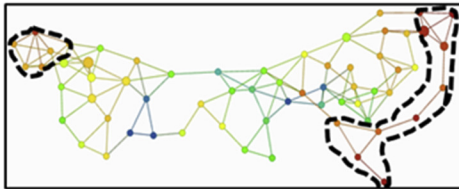
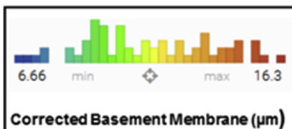
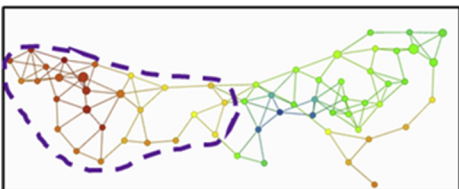
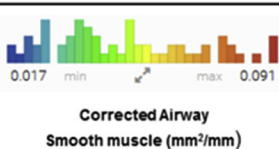
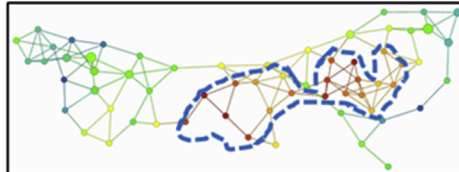
BOBCAT Cohort**Discovery Cohort**







DISCOVERY SEVERE ASTHMA



REPLICATION SEVERE ASTHMA