Supplementary Materials

Supplementary Methods

Participant Inclusion and Exclusion Criteria

Allergic Asthmatics (AA)

A. Inclusion Criteria:

1. Baseline forced expiratory volume in 1 second (FEV_1) determined at the initial visit no less than 75% of the predicted value after bronchodilator administration.
2. Clinical history of allergic symptoms to cat or dust mite allergen and demonstrated skin reactivity (a positive allergen skin prick test).
3. Positive methacholine challenge, defined as a provocative concentration inducing a 20% reduction in FEV_1 (PC20) <16 mg/ml.
4. Life-long absence of cigarette smoking (defined as a lifetime total of less than 5 pack-years and none in 5 years).
5. Willing and able to give informed consent.
6. Expressed the desire to participate in an interview with the principal investigator.
7. Age between 18 and 50 years.

B. Exclusion Criteria:

1. Women of childbearing potential who are documented to be pregnant (based on urine beta-HCG testing), are sexually active and not using contraception, are seeking to become pregnant, or who are breast feeding.

2. Spontaneous asthmatic episode or clinical evidence of upper respiratory tract infection within the previous 6 weeks.
3. Participation in a research study involving a drug or biologic during the 30 days prior to the study.
4. Intolerance to albuterol, atropine, lidocaine, fentanyl, or midazolam.
5. Antihistamines within 7 days of the screening visit.
6. Presence of diabetes mellitus, congestive heart failure, ventricular arrhythmias, history of a cerebrovascular accident, renal failure, history of anaphylaxis, or cirrhosis.
7. Use of systemic steroids, increased use of inhaled steroids, beta blockers or monoamine oxidase inhibitors within 6 weeks of the initial visit.
8. Antibiotic use for respiratory disease within 1 month of the initial visit or a respiratory tract infection within 6 weeks of the bronchoscopy visits.
9. A history of asthma-related respiratory failure requiring intubation.
10. Quantitative skin-prick test positive reaction down to an allergen concentration of 0.056 bioequivalent allergy units (BAU) or allergy units (AU)/ml.
11. Participants with a high possibility of poor compliance with the study.
12. Cigarette smoking within the past 5 years or > 5 pack years total.
13. Having second-hand cigarette smoke exposure or indoor furry pets except in the case of dog, if the subject is not allergic to the dog and the subject has a negative skin test to dog.
14. Other lung diseases, such as sarcoidosis, bronchiectasis, or active lung infection.
15. Use of targeted biological therapy for asthma or allergic disorders including but not limited to benralizumab, dupilumab, mepolizumab, omalizumab, or reslizumab currently or within the last year.

16. Immunotherapy with cat or dust mite extract now or in the past.
17. Non-English speakers.
18. History of coagulopathy, thrombocytopenia, pulmonary hypertension, and/or use of anti-coagulants/anti-platelet drugs.

Allergic Non-asthmatic Controls (AC)

C. Inclusion Criteria:

1. History of either (a) allergic rhinitis (with one or more of the following symptoms: nasal congestion, sneezing, runny nose, postnasal drainage), (b) allergic conjunctivitis (ocular itching, tearing and/or swelling) or (c) contact allergy associated with cat dander or dust mite and a positive allergy test to the same allergen.
2. Baseline FEV_1 and forced vital capacity (FVC) determined at the initial visit no less than 80% of the predicted value.
3. Positive allergy skin prick test to cat dander or dust mite allergen.
4. Life-long absence of cigarette smoking (defined as a lifetime total of less than 5 pack-years and none in 5 years).
5. Willing and able to give informed consent.
6. Expressed the desire to participate in an interview with the principal investigator.
7. Age between 18 and 50 years.

D. Exclusion Criteria:

1. A history of asthma.
2. Exclusion criteria #1, 3-8, and 10-18 from section B (**see above**).

3. Positive methacholine challenge (PC20 <16 mg/ml).

Healthy Controls (HC)

E. Inclusion Criteria:

1. No history of allergy and negative allergen skin prick testing.
2. Inclusion criteria #2, 4-7 from section C (**see above**).

F. Exclusion Criteria:

1. Exclusion criteria #1-3 from section D (**see above**).

Medication Hold Parameters

The following medications were held for at least the time period listed below prior to study visits, for all subjects.

Medication	**Minimum time to withhold**
Montelukast	24 hours
Long-acting bronchodilators (LABA)	12 hours
Theophylline	12 hours
Short-acting bronchodilators (SABA)	6 hours
Antihistamines	7 days
Aspirin or Ibuprofen (prior to bronchoscopy visits only)	2 days

Inhaled corticosteroids	2 weeks
Inhaled corticosteroid/LABA	2 weeks

Single-cell RNA-sequencing and Computational Data Analysis

Read alignment and quantification

Raw sequencing data was pre-processed with CellRanger (v3.0.2, 10X Genomics) to demultiplex FASTQ reads, align reads to the human reference genome (GRCh38, v3.0.0 from 10X Genomics), and count unique molecular identifiers (UMI) to produce a *cell x gene* count matrix (*87*). For the co-culture data, matrices underwent an additional step of background correction with *remove-background* using CellBender (v0.1.0) with default parameters except for the learning rate, which was set to 5.10^{-5}, due to the high amount of ambient RNA molecules that result from the culture conditions (*88*). All count matrices were then aggregated with Pegasus (v0.17.2, Python) using the *aggregate_matrices* function (*89*). CellRanger parameters were adjusted to select the top 7,000 droplets, as we expected to have captured at least 6,000 cells from each 10X experiment. Since this 7,000-droplet cutoff likely also captured empty droplets and poor-quality cells, we next applied a more stringent cutoff: cells with >30% mitochondrial UMI or <500 unique genes detected were deemed low-quality cells and were filtered out of the matrix prior to proceeding with downstream analyses (**fig. S1B**). The percent of mitochondrial UMI was computed using 13 mitochondrial genes (*MT-ND6, MT-CO2, MT-CYB, MT-ND2, MT-ND5, MT-CO1, MT-ND3, MT-ND4, MT-ND1, MT-ATP6, MT-CO3, MT-ND4L, MT-ATP8*) using the *qc_metrics* function in Pegasus. The counts for each remaining cell in the matrix were then

log-normalized by computing the log1p(counts per 100,000), which we refer to in the text and figures as log(CPM). The detailed quality control statistics for these datasets are compiled in **data S2**.

Cell clustering and lineage-specific subclustering analysis strategy

A two-step clustering strategy was used to analyze this scRNAseq dataset. Briefly, our strategy consisted of first generating a low-resolution clustering solution of the data to identify global cell lineages (i.e., AEC, CD4 T cells, CD8 T cells, MNP, B cells, NK cells, and mast cells). Cell lineage identity was annotated based on the sets of marker genes uniquely expressed by each cluster, defined by unbiased differential expression analysis (see *Marker gene identification*). In addition to this set of unbiased markers, we confirmed lineage identity by assessing the expression of canonical marker genes. Where sufficient cells in a lineage were captured (i.e., AEC, CD4 T cells, CD8 T cells, MNP), we also performed subclustering analysis to identify stable subclusters within each cell lineage. For the subclustering analyses, cells that were likely to represent doublets were filtered using a biologically informed approach, where cells that co-expressed multiple canonical lineage markers were manually excluded due to the co-expression of *CD3D-EPCAM*, *CD3D-LYZ*, *CD3D-SLPI*, and *SLPI-LYZ*. This approach was validated using *scrublet* (*90*).

For the global clustering and lineage-specific subclustering analyses, 2,000 highly variable genes were selected using the *highly_variable_features* function in Pegasus and used as input for

principal component analysis (*89*). To account for technical variability between donors, the resulting principal component scores were aligned using the Harmony algorithm (*91*). The top 50 principal components were used as input for Leiden clustering (*92*) and Uniform Manifold Approximation and Projection (UMAP) algorithm (spread=1, min-dist=0.5) (*93*).

For lineage-specific subclustering analyses, we used a previously reported analytical strategy involving the quantification of cluster stability across multiple Leiden resolutions (range, 0.3-1.9) to determine the most stable clustering resolution solution for downstream analyses (*94*). We iteratively subsampled 90% of the data (20 iterations) and, at each iteration, made a new clustering solution with the subsampled data and calculated an Adjusted Rand Index (ARI) to compare it to the clustering solution from the full data. An ARI value close to 1 indicates that the clustering solutions for the subsampled data and the full data are similar, indicating a stable clustering solution. The highest resolution where the median ARI across all iterations was >0.9 was used in our initial clustering of the data. In two cases, clustering was refined manually to segregate small clusters of known rarer cell types with distinct biological functions (MNP: DC1 (*CLEC9A*) cells; AEC: serous cells) that could not be identified via unbiased clustering because too few cells were captured through our profiling effort.

Marker gene identification

The marker genes defining each distinct cell cluster from our global and lineage-specific subclustering analyses were determined by applying two complementary methods. First, we

captured genes with high expression in each cluster by calculating the area under the receiver operating characteristic (AUROC) curve for the log(CPM) values of each gene as a predictor of cluster membership using the *de_analysis* function in Pegasus. Genes with an AUC ≥ 0.75 were considered marker genes for a particular cluster. Second, we created a pseudobulk count matrix to identify genes with lower expression that were highly specific for a given cluster (*95*). Specifically, we summed the UMI counts across cells for each unique cluster/sample combination to create a matrix of *n genes x (n samples*n clusters)* and performed "one-versus-all" (OVA) differential expression (DE) analyses for each cluster using the *DESeq2* package (v1.32.0, R v4.1.0) (*96*). For each cluster, we used an input model *gene ~ in_clust*, where *in_clust* is a factor with two levels indicating if the sample was in or not in the cluster being tested. A Wald test was then used to calculate *P* values and compute a false discovery rate (FDR) using the Benjamini-Hochberg method. We identified marker genes that were significantly associated with a particular cluster as having an FDR<0.05. Non-overlapping marker genes (excluding ribosomal and mitochondrial genes) for each cluster were sequentially identified by first selecting genes with an AUROC ≥ 0.75, followed by those with an OVA pseudobulk FDR <0.05, and up to the top 50 genes were visualized with the *ComplexHeatmap* package (v2.8.0, R) (*97*). The full list of marker genes is compiled in **data S4**.

Differential gene expression analysis

Comparisons between disease groups and experimental conditions were performed on pseudobulk count matrices using the *DESeq2* package (v1.32.0, R v4.1.0). The input model was either *gene ~ group* (AA or AC) or *gene ~ condition* (baseline or allergen). Significant DEG

were identified using a Wald test (FDR<0.1). To test for an association between disease group and experimental condition, we concatenated the factors of interest (*group* and *condition*) to create a new factor level called *interaction* and then used an input model of *gene ~ interaction*. Significant DEG were identified using a likelihood ratio test (FDR<0.1). The full list of DEG is compiled in **data S6**.

Gene set scoring and gene set enrichment analyses (GSEA)

Gene set scoring was performed using the *calc_signature_score* function from Pegasus. The hillock cell gene set was curated based on the published signature from Montoro et al. (*23*). The MC4 (*CCR2*), MC2 (*SPP1*) and Mac2 (*A2M*) gene sets used to score the co-culture data were based on the top 50 markers genes for those clusters as determined by AUROC statistics and OVA pseudobulk statistics described above (see *Marker gene identification* and **data S4**).

GSEA was performed using the *fgsea* function from the *fgsea* package (v1.18.0, R) with 10,000 permutations to test for independence. For GSEA performed to validate cluster annotations, the input gene rankings for a given cluster were based on their OVA pseudobulk log fold-change values, where the gene with the highest log fold-change was ranked first and the lowest log fold-change ranked last. Only genes that were expressed in >5% of cells were included in the ranking lists. CD4 T cell gene sets for T_H2, T_H17, and T_HIFNR were derived from Seumois et al. (*33*). Additionally, CD4 T cell gene sets based on cytokine-induced cell states were generated by performing differential expression analysis on publicly available bulk RNAseq data published by

Cano-Gamez et al. (*39*). Samples derived from naïve T cells stimulated with either T_H1-, T_H2- or T_HIFNR-stimulating cytokines were used for an OVA differential expression analysis with *DESeq2* (v1.32.0, R). For each cell state, we used an input model *gene ~ cell_state* where *cell_state* is a factor with two levels indicating if the sample was stimulated by the cytokines for the given cell state, and a Wald test was used to identify genes that were associated with each state. Genes with an FDR <0.1 and a logFC>1 were considered cell-state specific genes and used as input for GSEA. Tissue resident and effector memory CD8 T cell gene sets were derived from Kumar et al. (*36*). All gene sets used for cellular annotation are compiled in **data S7**.

For GSEA performed for Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis, the input gene rankings for each cluster were based on the pseudobulk log fold-change values when comparing AA vs. AC at allergen, where the gene with the highest log fold-change (i.e., associated with AA) was ranked first and the lowest log fold-change (i.e., associated with AC) was ranked last. The input gene sets tested were derived from the KEGG pathways database and are compiled in **data S8**.

Ingenuity Pathway Analysis

Ingenuity Pathway Analysis (IPA; Qiagen) was performed on DEG identified using the gene ~ condition model for MC after allergen challenge and gene ~ interaction input model for AEC. IPA analysis identified canonical pathways in which DEG were overexpressed in each group (*pORA*<0.1). IPA of these same DEG was also used to identify upstream regulators of gene expression changes, with a z-score representing the predicted activation state of the regulator and

|z-score|>2 considered significant. The full list of predicted pathways and upstream regulators is compiled in **data S8**.

Disease association analysis

To identify the association between cluster abundance and disease group (AC, AA) at a given experimental condition (baseline, allergen), we used a mixed-effects association logistic regression model similar to that described by Fonseka et al. (*20*). We used the *glmer* function from the *lme4* package (v1.1-27.1, R) to fit a logistic regression model for each cell cluster. Each cluster was modelled independently as follows:

$$cluster \sim 1 + condition{:}group + condition + group + (1 \mid id)$$

where *cluster* is a binary indicator set to 1 when a cell belongs to the given cluster or 0 otherwise, *condition* is factor with 2 levels (baseline, allergen), *group* is a factor with 2 levels (AC, AA), and *id* is a factor with 8 levels indicating the participant. The notation *(1|id)* indicates that *id* is a random intercept. The least-squares means of the factors in the model were calculated and pairwise comparisons were performed using the means of the groups at each condition (e.g., AA vs. AC within baseline, AA vs. AC within allergen, etc.) using the *lsmeans* function from the *emmeans* package (v1.5.4, R). An adjusted $P<0.05$ using Tukey's HSD method indicated a significant association between cluster abundance and the corresponding group and condition. The detailed modeling outputs are compiled in **data S5**.

Least absolute shrinkage and selection operation (LASSO) regression modeling

Least absolute shrinkage and selection operation (LASSO) regression modeling similar to that described by Smillie et al. (*98*) was used to select genes from significant cell:cell interactions (identified using CellPhoneDB, see *CellPhoneDB analysis*) that predict cluster abundance after allergen challenge. For a given cluster *C*, we calculated the percent of cells from each sample that were represented by that cluster. All genes involved in a significant cell:cell interaction with *C* were collected, along with the percentage of cells in the interacting cluster that expressed the corresponding genes in each sample. We then created a matrix of predictor variables (*pred*) with dimensions n_samples x n_cell:cell interaction genes. We then used the *glmnet* function from the *glmnet* package (v4.1-2, R), where the percentages of *C* were the response variable, *pred* was the input predictor variables, and the penalty parameter *alpha* was set to 1. To determine the optimal lambda value, we performed 1,000 iterations of cross-validation via the *cv.glmnet* function in R and recorded the lambda value that resulted in the lowest mean-squared error at each iteration. The median of these values was used to predict the resulting coefficients with the *predict* function from the *glmnet* package in R. Genes with non-zero coefficients were included in our models. The fraction of variance explained by the resulting models were compared to 100 null models, where the percentages of the *C* were shuffled. Only models with an empirical $P<0.01$ were considered significant. All model coefficients are compiled in **data S9**.

Estimation of RNA velocity

Count matrices of spliced and unspliced transcript abundances were calculated using *velocyto* (v0.17.15, Python). These matrices underwent dimensionality-reduction via principal component

analysis (PCA) and the top 50 principal components were used to compute a k-nearest neighbor graph (k=30) that was used as input to estimate cellular velocity with *scVelo* (v0.2.4, Python) (*99*). *CellRank* (v1.5.1, Python) was used to estimate initial and terminal states and these estimations were used to recover latent time with the *recover_latent_time* function with *scVelo*. The coherence of the vector field was used as a measure of confidence for the RNA velocity results and was calculated using the *velocity_confidence* function with *scVelo*. Velocity streamlines were plotted on UMAP embeddings with the *velocity_embedding_stream* function and the lineage driver genes associated with the inferred trajectory were calculated with the *lineage_drivers* function from the *CellRank*. The full list of lineage driver genes is compiled in **data S10**.

CellPhoneDB analysis

To infer potential receptor: ligand interactions between cell-cell pairs, we used CellPhoneDB (v2.1.7, Python) (*67*) and ran the algorithm independently on AA and AC cells after allergen challenge. Each cell cluster was tested as both a sender (ligand) and receiver (receptor) population as defined by the algorithm, and all possible combinations of cell-cell pairs were tested. We restricted potential interactions to those where the receptor and ligand were each expressed in >10% of their respective cluster and at least 20 cells, with significance defined as an empirical $P<0.001$ cutoff. Mean was defined by the algorithm as the aggregate mean expression (log(CPM)) of the receptor and ligand genes in the cell-cell pair. Rank was defined by the algorithm as the number of times a receptor: ligand interaction was significant out of the total number of cell-cell pairs tested, reported as $-\log_{10}$(rank) with higher values indicating increasing

specificity of the interaction. A curated list of interactions between T_H2: AEC and T_H2: MNP subclusters as well as basal:MNP and goblet:MNP subclusters were visualized as dot plots using the *ggplot2* package (v3.3.3, R). The full list of predicted receptor-ligand interactions is compiled in **data S11**.

Linear modeling of sums of receptor:ligand pairs

Linear modeling was performed on the sums of selected basal cell-MNP gene pairs identified by CellPhone DB (see *CellPhoneDB analysis*, corresponding to **Fig. 7D**) to identify receptor-ligand pairs significantly associated with AC or AA. For each receptor:ligand pair, we summed the pseudobulk log(CPM) values for both genes. We then fit a linear model with the *lmFit* function from the package *Limma* (version 3.48.0, R) where the model was *gene_sum ~ 1 + group* (AA or AC), where *gene_sum* is the sum of the log(CPM) values of the gene pair. Significant differences between groups were determined using an FDR<0.1. The detailed model outputs are compiled in **data S12**.

NicheNet analysis

To predict ligand:downstream target pairs between MNP and AEC clusters with >50 DEG (FDR<0.1 and $\log_2$FC>0.5) between AA and AC after allergen challenge, we used the *Nichenetr* package (v1.0.0, R) (*75*). DEG lists (FDR<0.1 and $\log_2$FC>0.5) were generated in each of the input clusters (MNP: MC2, MC4; AEC: goblet, quiesGoblet, basal, suprabasal) by comparing AA versus AC after allergen challenge (see *Differential gene expression analysis*). Next, we

defined the AEC clusters as the “sender” cells (those expressing potential ligands) and limited the ligands to those that were in our AEC DEG list. Then, we defined the MNP clusters as the “receiver” cells and limited the potential downstream target genes to those that were in our MNP DEG list. We then prioritized ligands with the highest regulatory potentials to target genes as published by Browaeys et al. (*75*), which was downloaded from Zenodo (https://zenodo.org/record/3260758/files/ligand_target_matrix.rds). The analysis was repeated with AEC clusters defined as the “receivers” and the MNP clusters as the “senders”. The regulatory potentials of the prioritized ligands were visualized as heatmaps using the *ComplexHeatmap* package (v2.8.0, R) and Circos plots using the *circlize* package (v0.4.12, R) (*97*, *100*). All ligand:downstream target pairs and regulatory potentials are compiled in **data S13**.

Supplementary Figures and Figure Legends

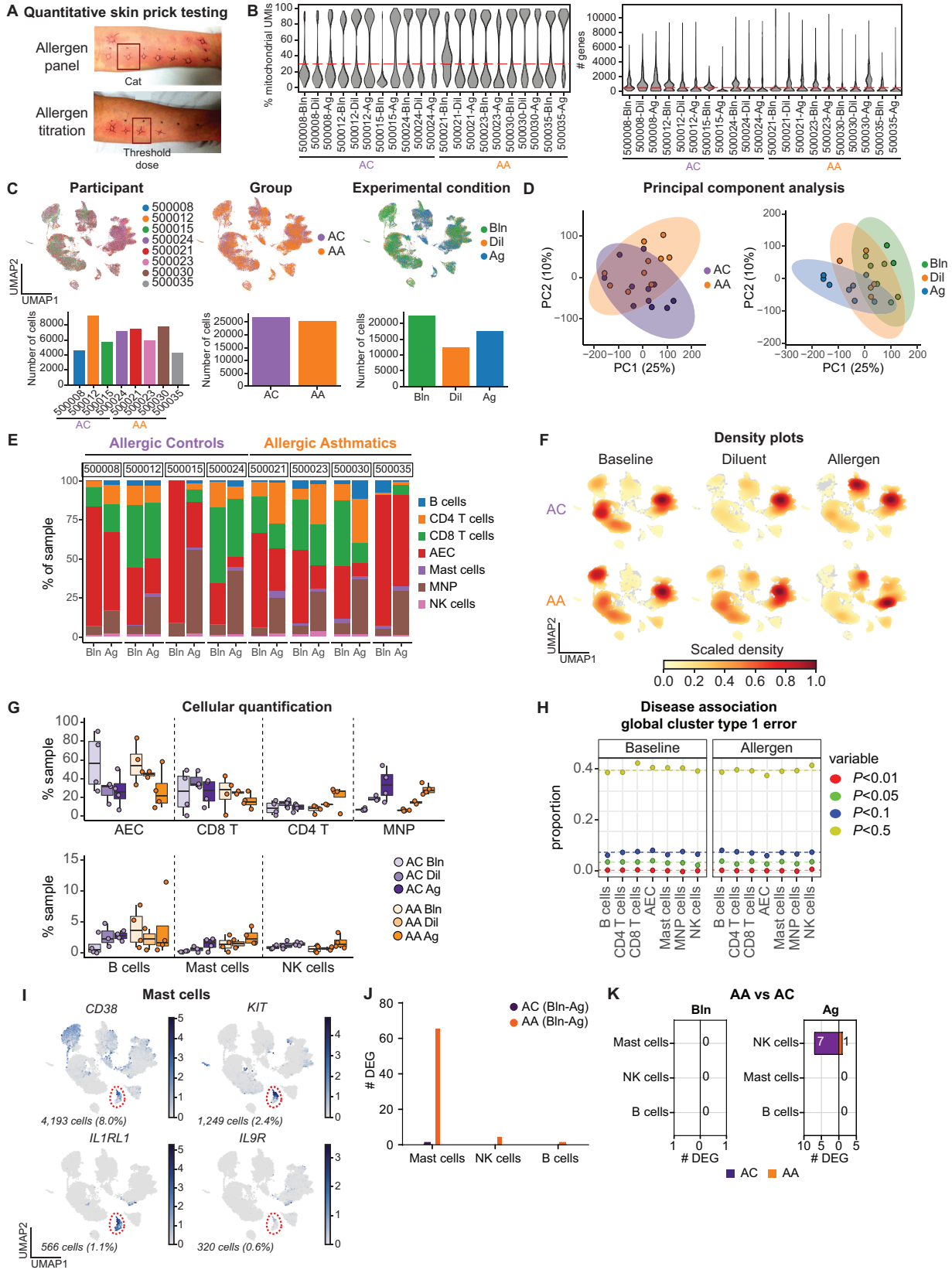

fig. S1. scRNAseq analysis quality metrics and overall cell composition. (**A**) Representative images of allergen skin prick testing (SPT) including quantitative skin SPT to determine threshold dose of allergen used during segmental allergen challenge (SAC). (**B**) Violin plots depicting the percentage of mitochondrial unique molecular identifiers (UMIs; left) and number of genes (right) by sample (n=21). Cells with >500 genes and <30% mitochondrial UMIs (dashed red lines) passed quality control (QC) filters and were used in downstream analysis. (**C**) UMAP feature plots (top row) showing cells post-QC filters and data integration using pseudo-coloring by participant identification number (left), disease group (middle), and experimental condition (right). Number of cells (bottom row) that passed QC filters from each participant (left), disease group (middle), and experimental condition (right). (**D**) Principal component analysis on the overall transcriptomes of each sample colored by disease group (left) and experimental condition (right). Ellipses indicate 95% confidence intervals. (**E**) Proportion distribution of the 7 cell lineages per sample. Bars represent the percentage of cells assigned to each color-coded cell lineage relative to the total cells for each sample. (**F**) UMAP embedding of cell density displaying the proportion of each cell lineage compared to every other cell lineage in all experimental conditions (Baseline: left, Diluent: middle, Allergen: right), faceted by disease group (AC: top, AA: bottom). (**G**) Contribution of each cell lineage defined in (E) shown as percentage (%) of total sample for each experimental condition. (**H**) Type-1 error for the disease association analysis in **Fig. 2E**. (**I**) Feature plots for genes enriched in mast cells using pseudo-coloring to indicate gene expression. Cell number and percentages (%) represent gene expression across all cell lineages. Scaled gene expression in log(CPM). (**J**) Number of DEG induced by SAC for mast cells, NK cells, and B cells in AC and AA. (**K**) Number of DEG between groups for mast cells, NK cells, and B cells. In (G), boxes represent the median (line) and interquartile

range (IQR) with whiskers extending to the remainder of the distribution no more than 1.5x IQR with dots representing individual samples.

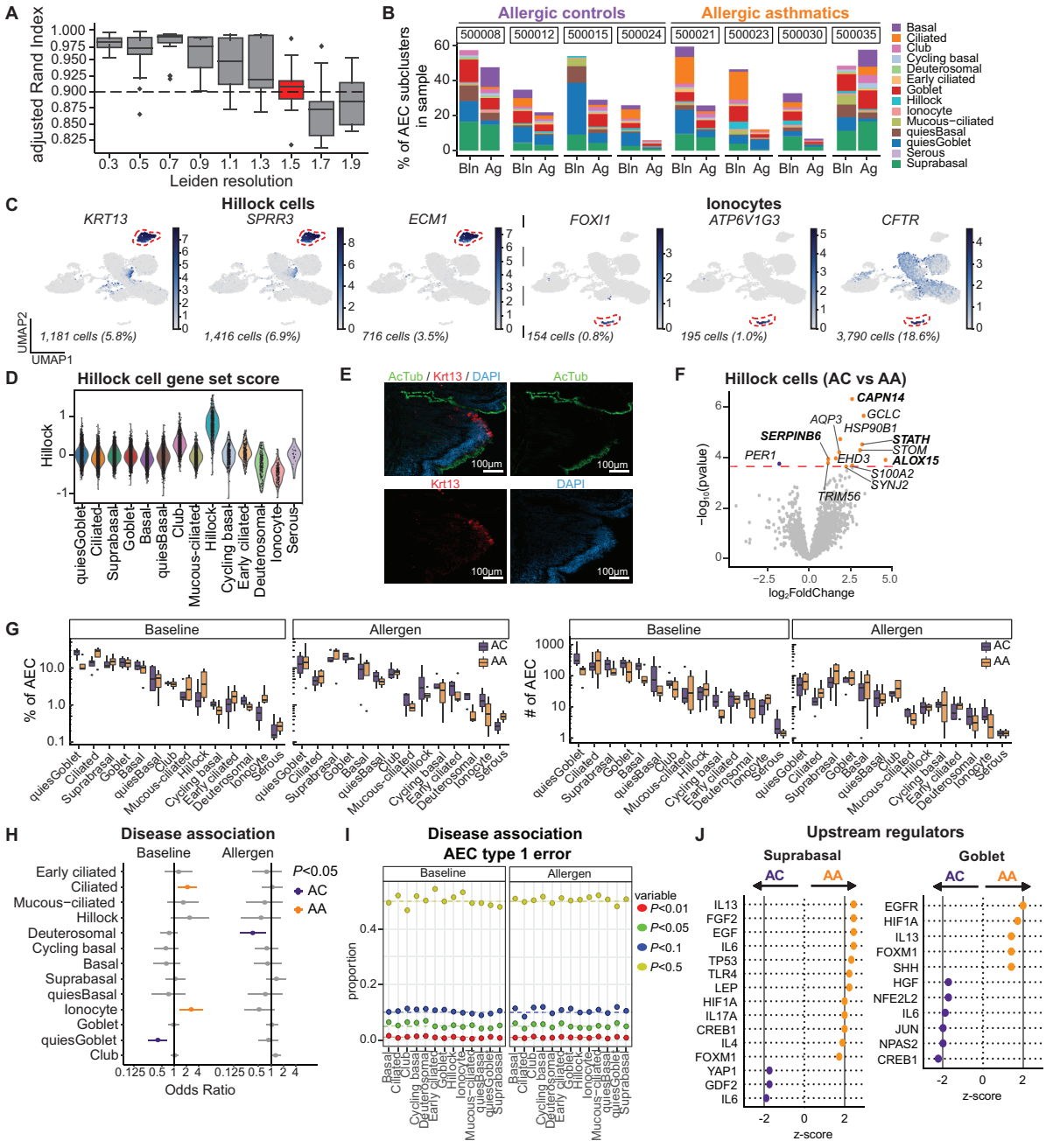

fig. S2. AEC subclustering and comparative analysis. (**A**) Subclustering stability of AEC reported as boxplots representing the distribution of the Adjusted Rand Indices (ARI). The red box indicates the Leiden resolution selected for downstream analyses. (**B**) Proportion distribution of the 14 AEC subsets per sample. (**C**) Feature plots using pseudo-coloring to indicate expression of top maker genes for hillock cells and ionocytes. Cell number and percentages (%)

represent gene expression across all AEC clusters. Scaled gene expression in log(CPM). (**D**) Hillock cell gene set score based on gene sets from Montoro et al. (*23*) applied across all AEC subclusters (**data S7**). (**E**) Immunofluorescence staining for KRT13 (red), acetylated tubulin (green), and DAPI (blue) in asthmatic airway tissue. (**F**) Volcano plot showing DEG between AC (left) and AA (right) in hillock cells. Horizontal dotted line represents FDR cutoff=0.1. Bolding indicates genes induced by IL-13. (**G**) Distributions of the percentage (left) and number (right) of AEC in each subset by group. Percentages represent the fraction of AEC that are categorized into each subset. (**H**) Odds ratio (OR) of disease association by AEC cluster at baseline and after allergen challenge. Color-coding denotes significant associations with AC (OR<1, purple) or AA (OR>1, gold). (**I**) Type-1 error for the disease association analysis in (H). (**J**) Predicted upstream regulators of DEG identified in suprabasal and goblet cells in **Fig. 3E** (AC: purple, AA: gold). Vertical solid lines represent z-score cutoff of |2|. In (A) and (G), boxes show the median (line) and IQR with whiskers extending to the remainder of the distribution no more than 1.5x IQR. In (D), DEG based on $\log_2$FC>0.5 and FDR<0.1 using the Wald test on pseudobulk count matrix.

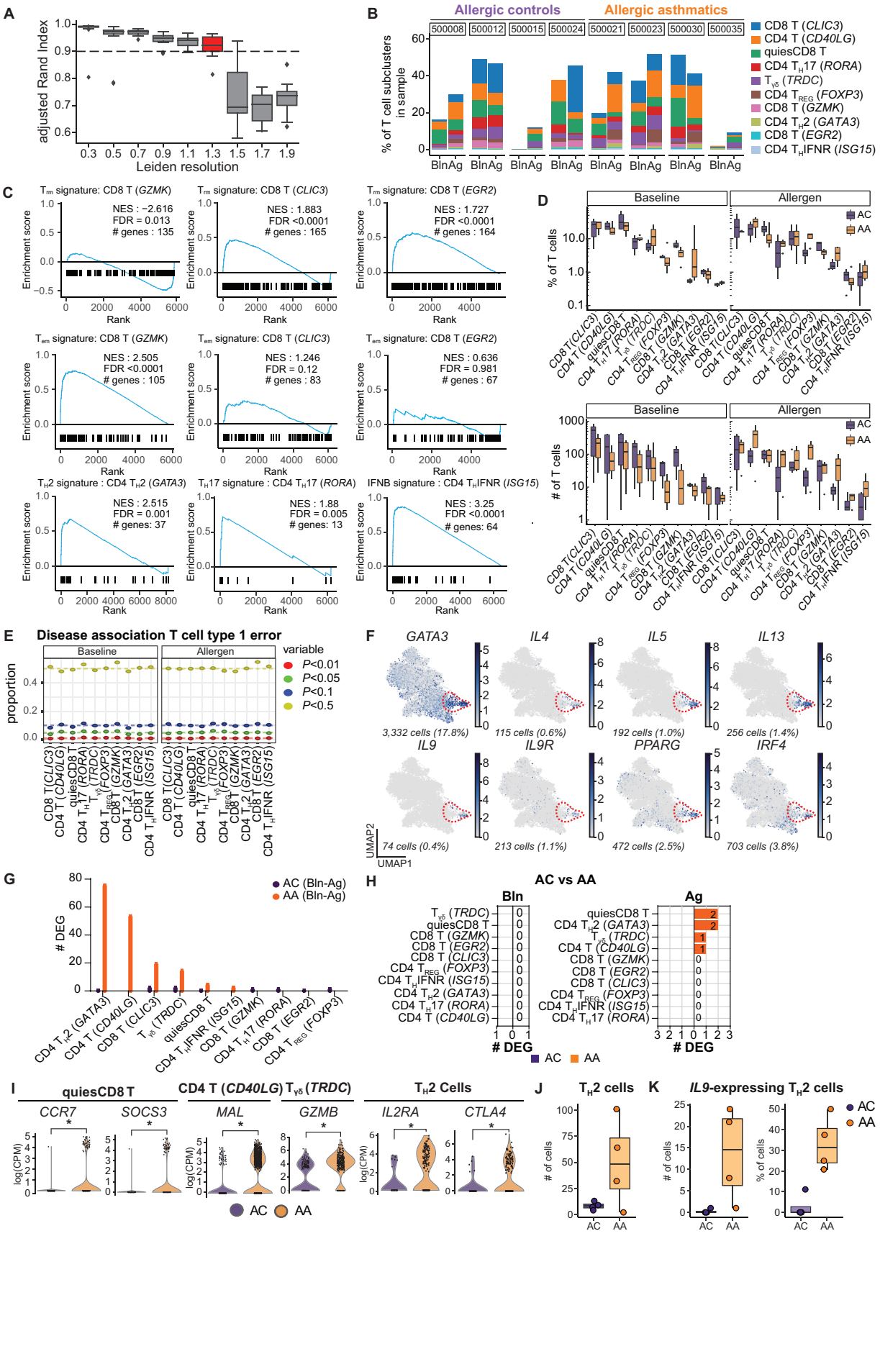
A
adjusted Rand Index
Leiden resolution
B
Allergic controls
Allergic asthmatics
% of T cell subclusters in sample
CD8 T (CLIC3)
CD4 T (CD40LG)
quiesCD8 T
CD4 TH17 (RORA)
Tγδ (TRDC)
CD4 TREG (FOXP3)
CD8 T (GZMK)
CD4 TH2 (GATA3)
CD8 T (EGR2)
CD4 THIFNR (ISG15)
C
Trm signature: CD8 T (GZMK)
NES : −2.616
FDR = 0.013
# genes : 135
Trm signature: CD8 T (CLIC3)
NES : 1.883
FDR <0.0001
genes : 165
Trm signature: CD8 T (EGR2)
NES : 1.727
FDR <0.0001
genes : 164
Tem signature: CD8 T (GZMK)
NES : 2.505
FDR <0.0001
genes : 105
Tem signature: CD8 T (CLIC3)
NES : 1.246
FDR = 0.12
# genes : 83
Tem signature: CD8 T (EGR2)
NES : 0.636
FDR = 0.981
# genes : 67
TH2 signature : CD4 TH2 (GATA3)
NES : 2.515
FDR = 0.001
# genes: 37
TH17 signature : CD4 TH17 (RORA)
NES : 1.88
FDR = 0.005
# genes: 13
IFNB signature : CD4 THIFNR (ISG15)
NES : 3.25
FDR <0.0001
genes: 64
Enrichment score
Rank
D
Baseline
Allergen
AC
AA
% of T cells
# of T cells
E
Disease association T cell type 1 error
proportion
variable
P<0.01
P<0.05
P<0.1
P<0.5
F
GATA3
IL4
IL5
IL13
IL9
IL9R
PPARG
IRF4
3,332 cells (17.8%)
115 cells (0.6%)
192 cells (1.0%)
256 cells (1.4%)
74 cells (0.4%)
213 cells (1.1%)
472 cells (2.5%)
703 cells (3.8%)
UMAP1
UMAP2
G
AC (Bln-Ag)
AA (Bln-Ag)
DEG
H
AC vs AA
Bln
Ag
DEG
AC
AA
I
quiesCD8 T
CD4 T (CD40LG) Tγδ (TRDC)
TH2 Cells
CCR7
SOCS3
MAL
GZMB
IL2RA
CTLA4
log(CPM)
J
TH2 cells
# of cells
K
IL9-expressing TH2 cells
# of cells
% of cells

fig. S3. T cell subclustering and comparative analysis. (**A**) Subclustering stability of T cells is reported as boxplots representing the distribution of the Adjusted Rand Indices (ARI). The red box indicates the Leiden resolution selected for downstream analyses. (**B**) Proportion distribution of the 10 T cell subsets per sample. (**C**) Gene set enrichment analysis (GSEA) of CD8 T cells based on tissue resident memory (T_{rm}; top row) and effector memory (T_{em}; middle row) gene sets from Kumar et al. (*33*) and GSEA of CD4 T_H2, T_H17, and T_HIFNR (bottom row) based on gene sets from Cano-Gamez et al. (*36*) (**data S7**). NES, normalized enrichment score. (**D**) Distributions of the percentage (top) and number (bottom) of T cells in each subset by group. Percentages represent the fraction of T cells that are categorized into each subset. (**E**) Type-1 error for the disease association analysis in **Fig. 4E**. (**F**) Feature plots showing genes enriched in T_H2 cells using pseudo-coloring to indicate gene expression. Cell number and percentages (%) represent gene expression across all T cell clusters. Scaled gene expression in log(CPM). (**G**) Number of DEG induced by SAC for all T cell clusters in AC and AA. (**H**) Number of DEG between groups at baseline and after allergen challenge for all T cell clusters. (**I**) Violin plots of DEG identified in (H) showing pairwise comparisons between groups. Each dot represents a single cell. *FDR<0.1. (**J**) Number of T_H2 cells in AC and AA after SAC. (**K**) Number of *IL9*-expressing T_H2 cells and percentage of T_H2 cells expressing *IL9* in AC and AA after SAC. In (G), (H), and (I), DEG based on $\log_2$FC>0.5 and FDR<0.1 using the Wald test on pseudobulk count matrix. Scaled gene expression in log(CPM). In (A), (D), (J), and (K), boxplots show median (line) and IQR with whiskers extending to the remainder of the distribution no more than 1.5x IQR with dots representing individual samples.

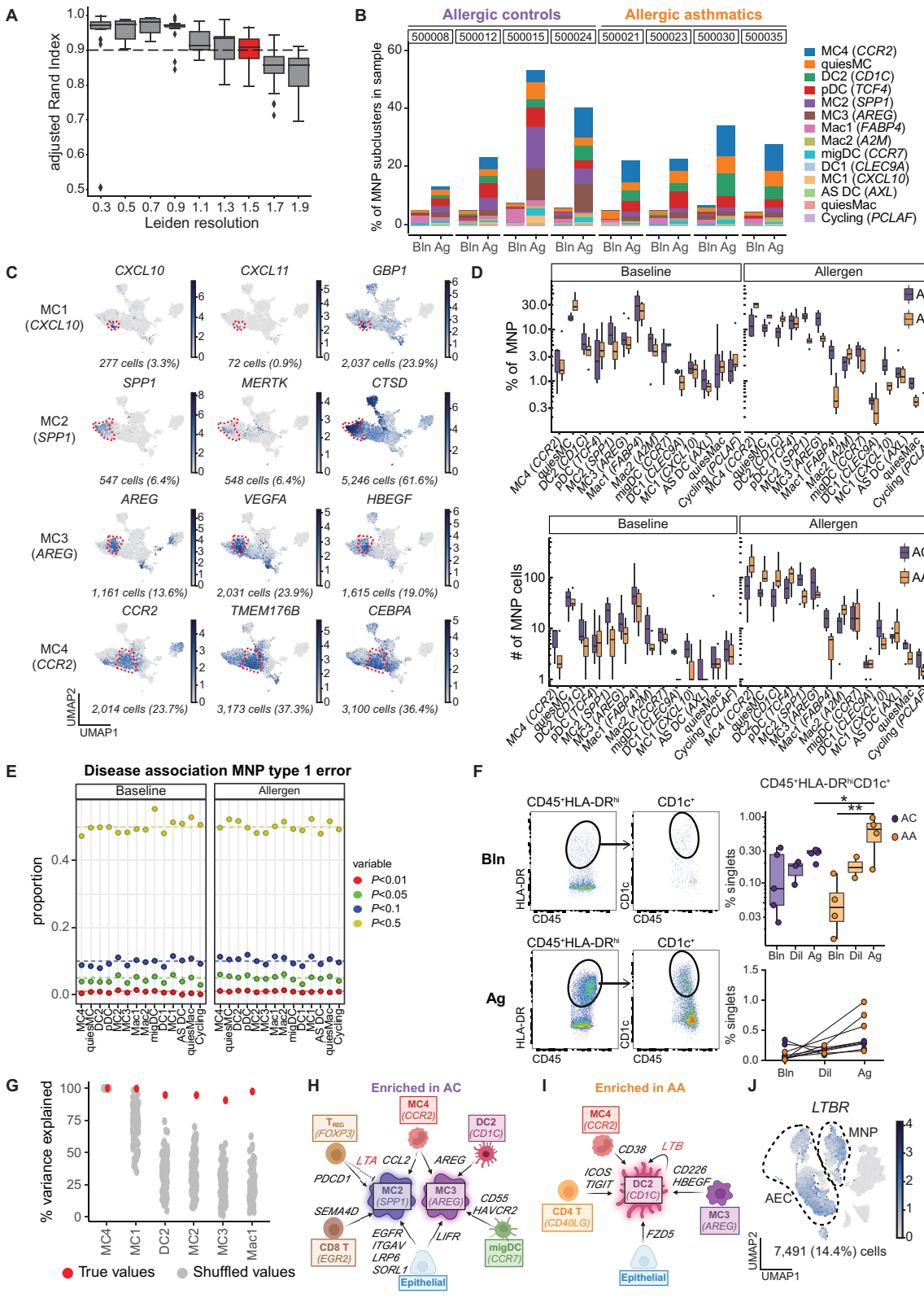
A
adjusted Rand Index
Leiden resolution
B
Allergic controls
Allergic asthmatics
% of MNP subclusters in sample
MC4 (CCR2)
quiesMC
DC2 (CD1C)
pDC (TCF4)
MC2 (SPP1)
MC3 (AREG)
Mac1 (FABP4)
Mac2 (A2M)
migDC (CCR7)
DC1 (CLEC9A)
MC1 (CXCL10)
AS DC (AXL)
quiesMac
Cycling (PCLAF)
C
CXCL10
CXCL11
GBP1
MC1 (CXCL10)
277 cells (3.3%)
72 cells (0.9%)
2,037 cells (23.9%)
SPP1
MERTK
CTSD
MC2 (SPP1)
547 cells (6.4%)
548 cells (6.4%)
5,246 cells (61.6%)
AREG
VEGFA
HBEGF
MC3 (AREG)
1,161 cells (13.6%)
2,031 cells (23.9%)
1,615 cells (19.0%)
CCR2
TMEM176B
CEBPA
MC4 (CCR2)
2,014 cells (23.7%)
3,173 cells (37.3%)
3,100 cells (36.4%)
UMAP1
UMAP2
D
Baseline
Allergen
AC
AA
% of MNP
# of MNP cells
E
Disease association MNP type 1 error
proportion
variable
P<0.01
P<0.05
P<0.1
P<0.5
F
CD45+HLA-DRhi
CD1c+
CD45+HLA-DRhiCD1c+
Bln
Ag
HLA-DR
CD45
CD1c
% singlets
Bln
Dil
Ag
G
% variance explained
True values
Shuffled values
H
Enriched in AC
TREG (FOXP3)
MC4 (CCR2)
DC2 (CD1C)
LTA
CCL2
AREG
PDCD1
MC2 (SPP1)
MC3 (AREG)
CD55
HAVCR2
SEMA4D
CD8 T (EGR2)
EGFR
ITGAV
LRP6
SORL1
LIFR
migDC (CCR7)
Epithelial
I
Enriched in AA
CD38
LTB
ICOS
TIGIT
CD226
HBEGF
CD4 T (CD40LG)
FZD5
J
LTBR
MNP
AEC
7,491 (14.4%) cells

fig. S4. MNP subclustering and comparative analysis. (**A**) Subclustering stability of MNP is reported as boxplots representing the distribution of the Adjusted Rand Indices (ARI). The red box indicates the Leiden resolution selected for downstream analyses. (**B**) Proportion distribution of the 14 MNP subsets per sample. (**C**) Feature plots of selected top marker genes in MC1-4 using pseudo-coloring to indicate gene expression. Cell number and percentages (%) represent gene expression across all MNP clusters. Scaled gene expression in log(CPM). (**D**) Distributions of the percentage (top) and number (bottom) of MNP in each subset by group. Percentages represent the fraction of MNP that are categorized into each subset. (**E**) Type-1 error for the disease association analysis in **Fig. 5E**. (**F**) Representative flow cytometry identifying DC2 in endobronchial brush samples at baseline (Bln; top row) and after allergen challenge (Ag; bottom row) from one AA participant. After excluding $CD326^+$ epithelial cells, $CD19^+$ B cells, and $CD4^+$ T cells, $CD45^+$HLA-DR^{hi} antigen presenting cells were identified. DC2 cells were defined as $CD45^+$HLA-$DR^{hi}CD1c^+$ and quantified in AC and AA at baseline (Bln) and after diluent- (Dil) and allergen-challenge (Ag). (**G**) Percent of the variance of the true proportion of MNP subclusters explained by the LASSO models (red points) vs. that of 100 iterations of shuffled proportions of MNP subclusters (grey points). Only subclusters that were enriched in either AC or AA after allergen challenge were tested. DC2 (*CD1C*), MC2 (*SPP1*), MC3 (*AREG*), and Mac (*FABP4*) had significant models (empirical $P<0.01$). (**H** and **I**) Schematics depicting the enrichment of MNP subclusters in each group as a function of positive (arrows) and negative (dashed line) associations with genes expressed by cell subsets. Only MNP subclusters with significant LASSO regression models (empirical $P<0.01$) are shown: MC2 and MC3 in AC (H) and DC2 in AA (I). (**J**) Feature plot using pseudo-coloring to indicate *LTBR* expression in the overall UMAP embedding. Cell number and percentages (%) represent gene expression across

all lineages. Scaled gene expression in log(CPM). In (A), (D), and (F), boxes show the median (line) and IQR with whiskers extending to the remainder of the distribution no more than 1.5x IQR with dots representing individual samples. In (F), *P* values were generated using a mixed effects model with Sidak correction to adjust for multiple comparisons. **P*<0.05, ***P*<0.01.

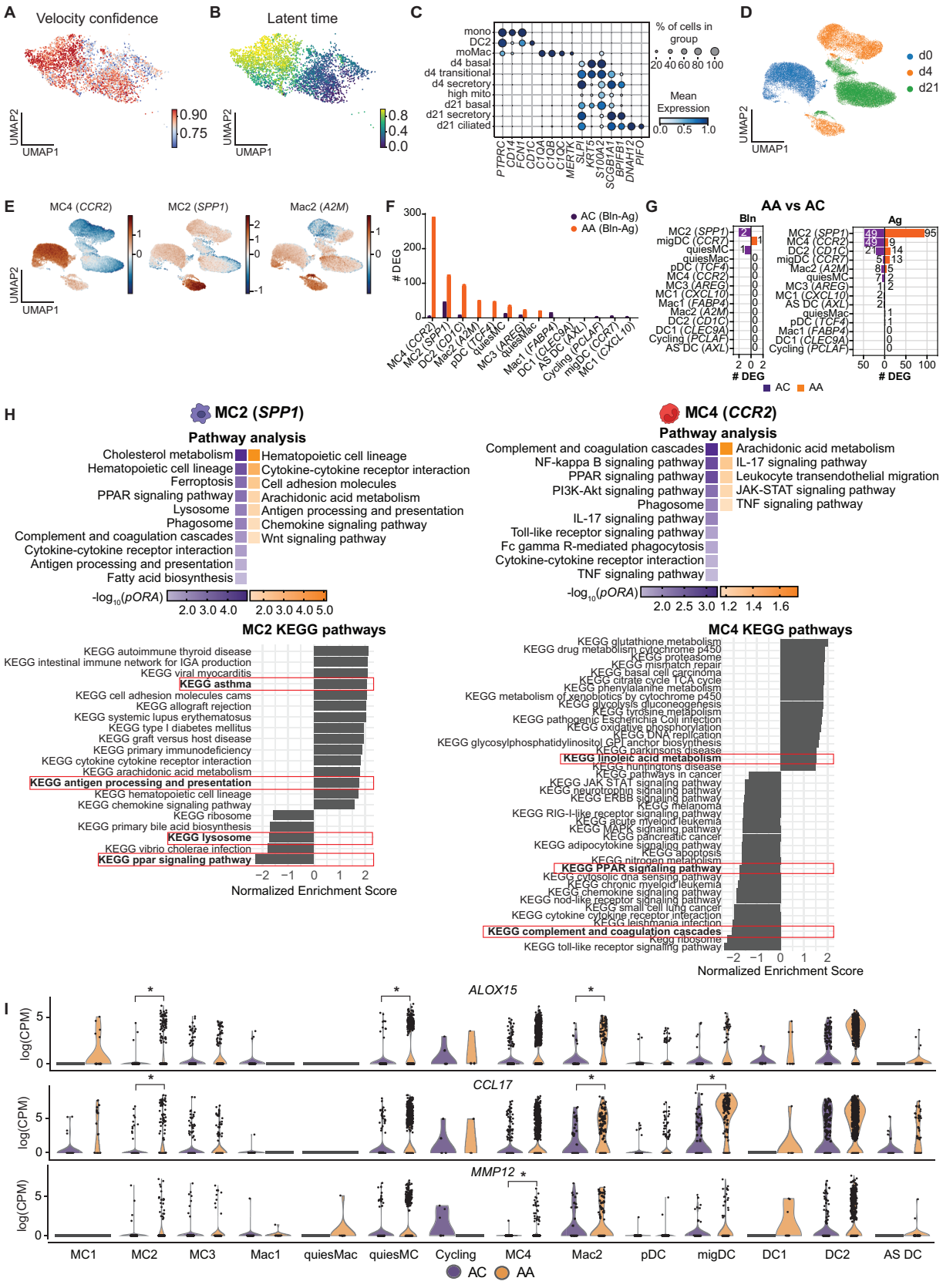
A
Velocity confidence
UMAP2
UMAP1
0.90
0.75
B
Latent time
0.8
0.4
0.0
C
mono
DC2
moMac
d4 basal
d4 transitional
d4 secretory
high mito
d21 basal
d21 secretory
d21 ciliated
PTPRC
CD14
FCN1
CD1C
C1QA
C1QB
C1QC
MERTK
SLPI
KRT5
S100A2
SCGB1A1
BPIFB1
DNAH12
PIFO
% of cells in group
20 40 60 80 100
Mean Expression
0.0 0.5 1.0
D
d0
d4
d21
E
MC4 (CCR2)
MC2 (SPP1)
Mac2 (A2M)
F
AC (Bln-Ag)
AA (Bln-Ag)
DEG
MC4 (CCR2)
MC2 (SPP1)
DC2 (CD1C)
Mac2 (A2M)
pDC (TCF4)
quiesMC
MC3 (AREG)
quiesMac
Mac1 (FABP4)
DC1 (CLEC9A)
AS DC (AXL)
Cycling (PCLAF)
migDC (CCR7)
MC1 (CXCL10)
G
AA vs AC
Bln
Ag
MC2 (SPP1)
migDC (CCR7)
quiesMC
quiesMac
pDC (TCF4)
MC4 (CCR2)
MC3 (AREG)
MC1 (CXCL10)
Mac1 (FABP4)
Mac2 (A2M)
DC2 (CD1C)
DC1 (CLEC9A)
Cycling (PCLAF)
AS DC (AXL)
MC4 (CCR2)
DC2 (CD1C)
migDC (CCR7)
Mac2 (A2M)
quiesMC
MC3 (AREG)
MC1 (CXCL10)
AS DC (AXL)
quiesMac
pDC (TCF4)
Mac1 (FABP4)
DC1 (CLEC9A)
Cycling (PCLAF)
DEG
AC
AA
H
MC2 (SPP1)
Pathway analysis
Cholesterol metabolism
Hematopoietic cell lineage
Ferroptosis
PPAR signaling pathway
Lysosome
Phagosome
Complement and coagulation cascades
Cytokine-cytokine receptor interaction
Antigen processing and presentation
Fatty acid biosynthesis
Hematopoietic cell lineage
Cytokine-cytokine receptor interaction
Cell adhesion molecules
Arachidonic acid metabolism
Antigen processing and presentation
Chemokine signaling pathway
Wnt signaling pathway
-log10(pORA)
2.0 3.0 4.0
2.0 3.0 4.0 5.0
MC4 (CCR2)
Pathway analysis
Complement and coagulation cascades
NF-kappa B signaling pathway
PPAR signaling pathway
PI3K-Akt signaling pathway
Phagosome
IL-17 signaling pathway
Toll-like receptor signaling pathway
Fc gamma R-mediated phagocytosis
Cytokine-cytokine receptor interaction
TNF signaling pathway
Arachidonic acid metabolism
IL-17 signaling pathway
Leukocyte transendothelial migration
JAK-STAT signaling pathway
TNF signaling pathway
-log10(pORA)
2.0 2.5 3.0 1.2 1.4 1.6
MC2 KEGG pathways
KEGG autoimmune thyroid disease
KEGG intestinal immune network for IGA production
KEGG viral myocarditis
KEGG asthma
KEGG cell adhesion molecules cams
KEGG allograft rejection
KEGG systemic lupus erythematosus
KEGG type I diabetes mellitus
KEGG graft versus host disease
KEGG primary immunodeficiency
KEGG cytokine cytokine receptor interaction
KEGG arachidonic acid metabolism
KEGG antigen processing and presentation
KEGG hematopoietic cell lineage
KEGG chemokine signaling pathway
KEGG ribosome
KEGG primary bile acid biosynthesis
KEGG lysosome
KEGG vibrio cholerae infection
KEGG ppar signaling pathway
−2 −1 0 1 2
Normalized Enrichment Score
MC4 KEGG pathways
KEGG glutathione metabolism
KEGG drug metabolism cytochrome p450
KEGG proteasome
KEGG mismatch repair
KEGG basal cell carcinoma
KEGG citrate cycle TCA cycle
KEGG phenylalanine metabolism
KEGG metabolism of xenobiotics by cytochrome p450
KEGG glycolysis gluconeogenesis
KEGG tyrosine metabolism
KEGG pathogenic Escherichia Coli infection
KEGG oxidative phosphorylation
KEGG DNA replication
KEGG glycosylphosphatidylinositol GPI anchor biosynthesis
KEGG parkinsons disease
KEGG linoleic acid metabolism
KEGG huntingtons disease
KEGG pathways in cancer
KEGG JAK STAT signaling pathway
KEGG neurotrophin signaling pathway
KEGG ERBB signaling pathway
KEGG melanoma
KEGG RIG-I-like receptor signaling pathway
KEGG acute myeloid leukemia
KEGG MAPK signaling pathway
KEGG pancreatic cancer
KEGG adipocytokine signaling pathway
KEGG apoptosis
KEGG nitrogen metabolism
KEGG PPAR signaling pathway
KEGG cytosolic dna sensing pathway
KEGG chronic myeloid leukemia
KEGG chemokine signaling pathway
KEGG nod-like receptor signaling pathway
KEGG small cell lung cancer
KEGG cytokine cytokine receptor interaction
KEGG leishmania infection
KEGG complement and coagulation cascades
Kegg ribosome
KEGG toll-like receptor signaling pathway
−2 −1 0 1 2
Normalized Enrichment Score
I
ALOX15
CCL17
MMP12
log(CPM)
MC1
MC2
MC3
Mac1
quiesMac
quiesMC
Cycling
MC4
Mac2
pDC
migDC
DC1
DC2
AS DC
AC
AA

fig. S5. Trajectory, co-culture, and selected MC comparative analysis. (**A**) Velocity confidence and (**B**) latent time analyses of selected MNP clusters corresponding to RNA velocity analysis performed in **Fig. 6A-B**. (**C**) Dot plot depicting gene expression levels and percentage of cells expressing genes across co-culture clusters identified in **Fig. 6C**. (**D**) Feature plot using pseudo-coloring to indicate day in the co-culture UMAP embedding. d0, day 0 (isolated blood $CD14^+$ monocytes). d4, day 4. d21, day 21. (**E**) Feature plot using pseudo-coloring to indicate airway MNP gene set score in the co-culture UMAP embedding, corresponding to gene set scores shown as violin plots in **Fig. 6D**. (**F**) Number of DEG induced by SAC for all MNP clusters in AC and AA. (**G**) Number of DEG between groups after allergen challenge for all MNP clusters. (**H**) Pathway analysis of DEG identified in (G) in MC2 (left) and MC4 (right). Ingenuity pathway analysis (IPA; top row) showing selected pathways out of the top 20 identified as enriched in each group based on gene overrepresentation (overexpression *P* value [*pORA*]<0.1). GSEA of KEGG pathways (bottom row) in AC (NES<0, FDR<0.1) and AA (NES>0, FDR<0.1). Red boxes indicate complementary pathways also identified by IPA. NES, normalized enrichment score. (**I**) Violin plots of key DEG in (G) depicting scaled gene expression distribution and pairwise comparisons between AC and AA, with each dot representing a single cell. *FDR<0.1. In (F), (G), and (I), DEG based on FDR<0.1 and $\log_2$FC>0.5 using the Wald test on pseudobulk count matrix.

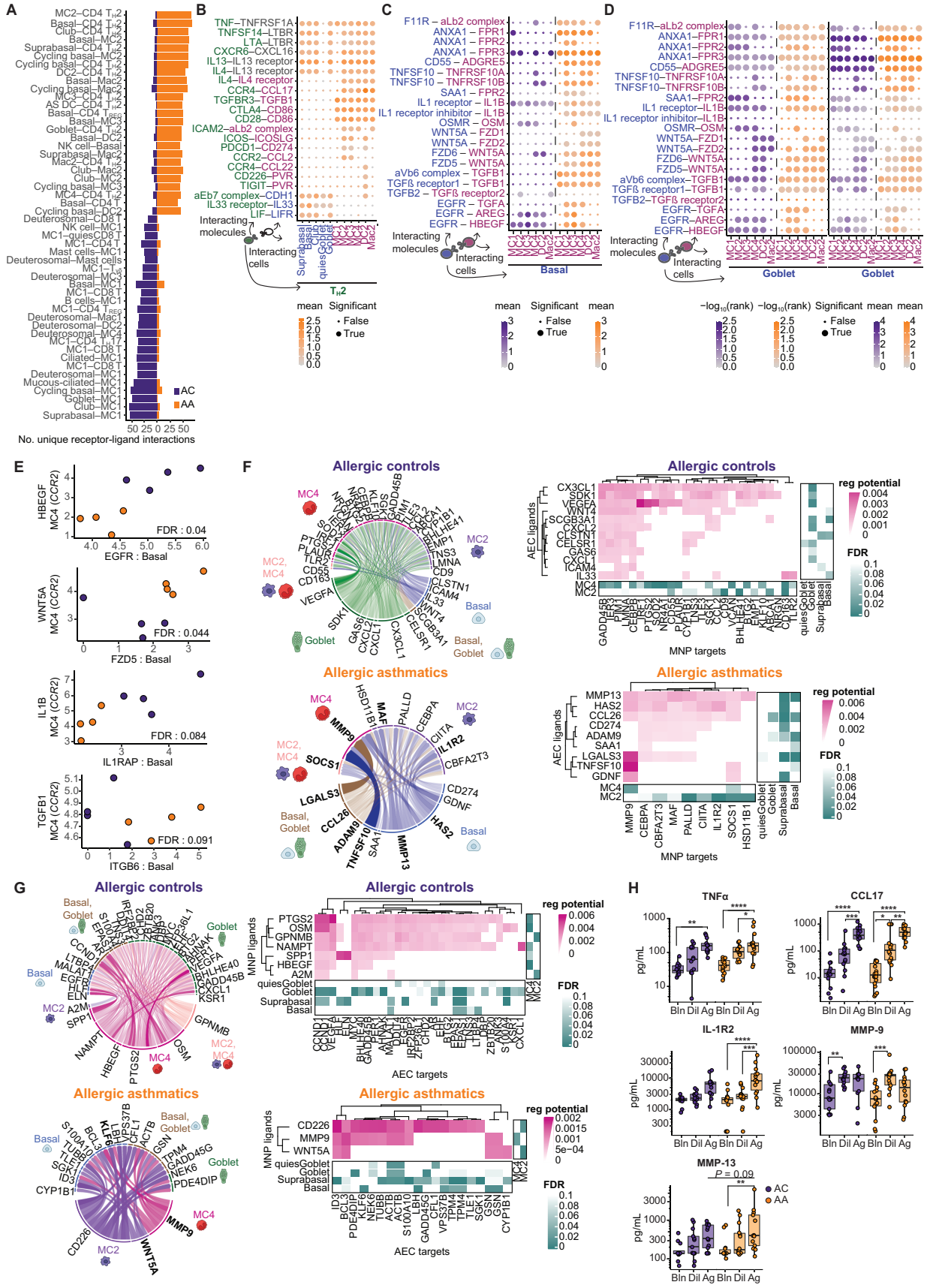
A
B
C
D
E
F
G
H
AC
AA
No. unique receptor-ligand interactions
Interacting molecules
Interacting cells
mean
Significant
False
True
Basal
Goblet
Allergic controls
Allergic asthmatics
reg potential
FDR
AEC ligands
MNP ligands
MNP targets
AEC targets
TNFα
CCL17
IL-1R2
MMP-9
MMP-13
P = 0.09
pg/mL
Bln Dil Ag Bln Dil Ag

fig. S6. Cell-cell interaction analysis. (**A**) Significant receptor-ligand interactions were predicted using CellPhoneDB v2.0 (**data S11**). Displayed are the top 50 cell-cell pairs with the greatest difference in the number (no.) of unique interactions between groups after allergen challenge, restricted to interactions between distinct cell lineages. Bar plots depict the number of unique receptor-ligand interactions in AC (purple) and AA (gold). (**B**) Dot plots of predicted interactions after SAC between T_H2-AEC and T_H2-MNP in AA, corresponding to **Fig. 7B**. Dot size indicates significance (true: empirical $P<0.001$) and color intensity indicates the aggregate mean expression of genes in each receptor-ligand pair. (**C**) Dot plots showing interactions after SAC between basal-MNP in AC (purple) and AA (gold), corresponding to **Fig. 7D**. Dot size indicates significance (true: empirical $P<0.001$) and color intensity indicates the aggregate mean expression of genes in each receptor-ligand pair. (**D**) Dot plots showing interactions after SAC between goblet-MNP in AC (purple) and AA (gold). Dot size indicates significance (true: empirical $P<0.001$) and color intensity indicates specificity of interaction to disease group [$-\log_{10}$(rank)] (left) or aggregate mean expression of genes in each receptor-ligand pair (right). (**E**) Linear modeling (**data S12**) of selected receptor-ligand pairs shown in **Fig. 7D**, depicting on a per-sample basis the log(CPM) expression of relevant genes in basal (x-axis) and MC4 (y-axis) cells. Each dot indicates an AC (purple) or AA (gold) sample collected after allergen challenge. (**F**-**G**) Circos plots and corresponding heatmap of NicheNet analysis (**data S13**) depicting ligands from AEC with their predicted downstream target genes in MC4 and MC2 (F) and ligands from MC4 and MC2 with their predicted downstream target genes in AEC (G). In the Circos plots, increasing line width and color intensity indicates higher regulatory potential score of the ligand on the downstream target gene. In the heatmap, color intensity indicates the regulatory potential score (pink) and FDR for each DEG (teal) associated with the interaction

occurring between each cell-cell pair. (**H**) BAL concentration of selected proteins (for TNFα, IL-1R2, MMP-9, and MMP-13, AC: n=13, AA: n=14; for CCL17, AC: n=14, AA: n=18). In (A) to (E), receptor-ligand interactions identified using CellPhoneDB v2.0 with an empirical $P<0.001$. In (E), linear modeling was performed on the sums of gene pairs, with FDR<0.1 considered significant. In (F) and (G), NicheNet analysis of DEG in AA compared to AC after SAC, based on FDR<0.1 and $\log_2$FC>0.5 using the Wald test on pseudobulk count matrix. In (H), boxes represent the median (line) and IQR with whiskers extending to the remainder of the distribution no more than 1.5x IQR with dots representing individual samples. *P* values were generated using a mixed effects model with Sidak correction to adjust for multiple comparisons. *$P<0.05$, **$P<0.01$, ***$P<0.001$, ****$P<0.0001$.

table S1. Antibodies used for experiments.

Antibody	Supplier	Cat. No.	RRID AB	Dilution
Flow cytometry				
Mouse anti-human CD326	Biolegend	324214	2098808	1:100
Mouse anti-human CD45	BD Biosciences	560779	1937332	1:100
Mouse anti-human CD3	Biolegend	300306	314042	1:50
Mouse anti-human CD19	Biolegend	363024	2564253	1:100
Mouse anti-human HLA-DR	Biolegend	307615	493589	1:100
Mouse anti-human CD1c	Biolegend	331519	10643413	1:100
Human Fc receptor blocking solution	Biolegend	422302	2869554	1:100
Immunofluorescence staining				
Mouse anti-human CD45	Biolegend	304056	2564155	1:25
Mouse anti-human MERTK	Biolegend	367608	2566401	1:25
Mouse anti-human CD45	Biolegend	304058	2564156	1:100
Rat anti-human C1q	Abcam	Ab11861	298643	1:400
Rabbit anti-human p63	Abcam	Ab124762	10971840	1:100
Rabbit anti-human cytokeratin 13	Abcam	ab92551	2134681	1:100
Mouse anti-human acetylated tubulin	Sigma	T6793	477585	1:400
Donkey anti-rat	Life Technologies	A21209	2535795	1:400
Goat anti-rabbit	Life Technologies	A11034	2576217	1:400
Human BD Fc Block	BD Biosciences	564219	2728082	1:100

Hoechst 33258	Biotium	40044		
VECTASHIELD Antifade Mounting Medium with DAPI	Vector Laboratories	H-1200	2336790	

Supplementary Datasets

data S1. Study participant characteristics. Demographics, allergen dosing, pulmonary function testing, and analyses performed for all study participants. HC, healthy control; AC, allergic non-asthmatic control; AA, allergic asthmatic; M, male; F, female; DP, *Dermatophagoides pteronyssinus* extract; Cat, cat hair extract; FEV_1, forced expiratory volume in 1 second; FVC, forced vital capacity; PC20, provocative concentration of methacholine inducing a 20% decline in FEV_1. BAL, bronchoalveolar lavage.

data S2. 10X sequencing statistics and metadata. Sample metadata and 10X Genomics experimental details. Rows represent individual samples with corresponding participant ID (id), disease group (group), and experimental condition in which the sample was collected (sample).

data S3. Cell numbers per cluster. Rows represent individual cell clusters with corresponding participant ID (id), disease group (group), and experimental condition in which the sample was collected (sample).

data S4. Marker genes for cluster identity. Each row represents a marker gene (gene) for each cell lineage and subcluster with corresponding area under the receiver operating curve values (AUROC), one-vs.-all (OVA) pseudobulk *P* value (OVA pseudobulk pval), OVA pseudobulk FDR (OVA pseudobulk FDR), and OVA pseudobulk $\log_2$(fold change) (OVA pseudobulk log-fc), marker gene significance as determined by AUROC (AUROC_marker) and pseudobulk (pseudobulk_marker) approaches, cell lineage (lineage), and cell cluster (cluster). Genes with an

AUC ≥0.75 or a pseudobulk FDR<0.05 are included for each cluster. This data file supports Figs. 2-5.

data S5. Disease association analysis. Cell cluster disease association analysis comparing AC vs. AA at baseline and after allergen challenge. The four sheets in this file include global cell lineages, T cells, MNP, and AEC. Each row represents a cluster (cluster) with corresponding experimental condition in which the sample was collected (sample), odds ratio (odds_ratio), lower limit of the 95% confidence interval (LCL), upper limit of the 95% confidence interval (UCL), and *P* value corrected for multiple comparisons using Tukey's HSD (p.value). This data file supports Figs. 2, 4, 5 and figs. S1-S4.

data S6. Differential gene expression analysis. Differential gene expression analysis comparing disease groups (AA vs. AC), experimental conditions (baseline vs. allergen), and modeling with a group:condition interaction term (only performed for AEC). Each sheet corresponds to a distinct differential expression comparison and includes the gene name (gene), normalized transcript counts averaged for all samples (baseMean), $\log_2$FC (log2FoldChange), standard error (lfcSE), Wald-statistic (stat), *P* value (pvalue), false discovery rate (FDR), significance (FDR<0.1), cell cluster (cluster), and comparison performed (contrast). The supplemental data file includes genes with a pvalue ≤ 0.2. The full list of genes tested can be accessed at GitHub (https://github.com/villani-lab/airway_allergic_asthma). This data file supports Figs. 3, 6 and figs. S1-S3 and S5.

data S7. Gene set signatures. Gene lists and corresponding references used to perform gene set scoring and gene set enrichment analysis. This data file supports Fig. 4 and figs. S2-S3.

data S8. Pathway and upstream regulator analyses. Pathway and upstream regulatory analyses performed on MC2 and MC4 using DEG between AA and AC after SAC (FDR<0.1). Upstream regulator analyses performed on suprabasal and goblet using DEG identified using the group:condition interaction term (FDR <0.1). Ingenuity pathway analysis (IPA) with each row representing a canonical pathway (pName) with corresponding number of DEG represented in the pathway (countDE), all genes in the pathway (countAll), *P* value (pv), perturbation accumulation *P* value (pAcc), combination *P* value (pComb), gene overrepresentation *P* value (pORA). KEGG pathway analysis with each row representing a canonical pathway with corresponding P value (pval), adjusted P value (padj), expression score (ES), normalized expression score (NES), and number of genes from the pathway found in our data set (size). Ingenuity upstream regulator analysis with each row representing a predicted upstream regulator (symbol) with corresponding Entrez Gene ID (entrez), log fold change (logFC), adjusted *P* value (adjpv), number of consistent DEG targets predicted to be significantly regulated (cDE_p), number of consistent DEG targets predicted to be regulated (cDE), all measured gene targets predicted to be regulated (cAll), combined FDR for predicted regulator (pv_comb_p_fdr), FDR for predicted regulator (pv_p_FDR), FDR for z-score (pv_zscore_fdr), activation z-score (zscore). This data file supports Fig. 6 and figs. S2 and S5.

data S9. LASSO analysis. Resulting variable coefficients from models produced by least absolute shrinkage and selection operation (LASSO) analysis. Each tab represents the LASSO

model from the 5 MNP clusters where a significant LASSO model was produced (P<0.01). Each tab has 3 columns: the variable coefficient in the model (coef), the cluster in which the gene was associated with MNP cluster abundance (cluster), and the gene that was associated with MNP cluster abundance (gene). This data file supports fig. S4.

data S10. Velocity analysis lineage drivers. Lineage driver genes for the inferred trajectory for selected MNP clusters. Each row represents a gene (feature) with corresponding correlation coefficient (corr), lower limit of the 95% confidence interval (ci_low), upper limit of the 95% confidence interval (ci_high), and *P* value (*pval*). This data file supports Fig. 6 and fig. S5.

data S11. CellPhoneDB analysis. All significant inferred interactions from CellPhoneDB v2.0. Each row represents a predicted receptor-ligand interaction (interacting_pair) with corresponding interacting clusters (cluster_pair), percentage of cells from AA and AC expressing the first gene of the interaction (aa_perc_a, ac_perc_a), percentage of cells from AA and AC expressing the second gene of the interaction (aa_perc_b, ac_perc_b), the number of cells from AA and AC expressing the first gene of the interaction (aa_ncells_a, ac_ncells_a), the number of cells from AA and AC expressing the second gene of the interaction (aa_ncells_b, ac_ncells_b), the rank of the interaction in AA and AC (aa_rank, ac_rank), the aggregate mean expression of genes in the putative interaction in AA and AC (aa_mean, ac_mean) and the group in which the interaction was found to be significant (significant_in; values can be “AA”, “AC” or “both”). This data file supports Fig. 7 and fig. S6.

data S12. Receptor ligand linear modeling. Linear modeling of predicted receptor-ligand interactions between basal-MC4 cells identified using CellPhoneDB v2.0 (**data S11**). Each row represents a predicted gene-gene interaction (gene_1, gene_2) with the corresponding cluster in which the genes are respectively expressed (cluster_1, cluster_2), log fold change (logFC), *P* value (p_val), and false discovery rate (FDR). This data file supports Fig. 7 and fig. S6.

data S13. NicheNet analysis. Each sheet represents a different NicheNet analysis for predicted AEC-MNP interactions in AA and AC. The first column represents the ligands expressed by the "sender" cell and the first row represents the downstream target expressed by the "receiver" cell. The values in the matrix represent the regulatory potential between the two genes as reported by the NicheNet algorithm. This data file supports Fig. 7 and fig. S6.