Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Beyond the T2/Non-T2 Dichotomy: Cross-Cohort Airway Transcriptomics Reveals a Continuous Type 2 Axis and Reproducible Heterogeneity in T2-Low Asthma

Kairui Meng
Chengdu University of Traditional Chinese Medicine
Yuanming Liu
Department of Respiratory and Critical Care Medicine, People's Hospital of Pengzhou City
Lu Wang
Chengdu University of Traditional Chinese Medicine
Xueqing Gong
Chengdu University of Traditional Chinese Medicine
Wengjun Tang
Hospital of Chengdu University of Traditional Chinese Medicine
Guobing Jia
Chengdu University of Traditional Chinese Medicine
Yanmei Wang
Institute of Traditional Chinese Medicine, Sichuan Academy of Chinese Medicine Sciences
Chengshi He
191336930440163.com

Hospital of Chengdu University of Traditional Chinese Medicine

Research Article

Keywords: Asthma, Type 2 inflammation, Airway transcriptomics, T2-low asthma, Endotypes, Biomarkers

Posted Date: August 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9900269/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Asthma airway inflammation is not fully captured by the conventional binary T2/non-T2 framework. We aimed to determine whether cross-cohort airway transcriptomics could define a more reproducible inflammatory structure, including both a continuous T2-related axis and heterogeneity within T2-low disease.

Methods

Six GEO-derived human airway epithelial or sputum transcriptomic cohorts (n = 454) were analyzed, including three discovery cohorts and three independent validation cohorts. T2 inflammation was quantified using a core three-gene signature (POSTN, CLCA1, and SERPINB2). A T2-associated epithelial axis was derived by integrating cross-cohort differential expression, pathway activity, and co-expression network analysis. Within T2-low samples, non-T2 endotypes were defined using axis-based clustering and evaluated by predictive modeling, differential expression, pathway analysis, and external validation.

Results

Core T2 signals showed stable directional concordance across cohorts. The core3-based T2 score robustly stratified samples across discovery and validation cohorts. A leakage-controlled elastic-net model discriminated T2-high from T2-low samples with strong performance and retained good discrimination in independent validation cohorts. The T2 epithelial axis score correlated strongly with the T2 score and remained positively associated with eosinophilic phenotype in external cohorts (pooled OR = 1.77, 95% CI 1.18–2.65; P = 0.0059). Within T2-low asthma, two reproducible non-T2 endotypes, E1 and E2, were identified. E1 was characterized by higher IFN_Antiviral, IL1_Inflammasome, and Th17/Neutrophil program scores, and a minimal E1/E2 classifier retained good discrimination (pooled AUROC = 0.899).

Conclusions

These findings support a layered model of asthma inflammation comprising a continuous T2 epithelial axis together with reproducible internal heterogeneity within T2-low disease.

Background

Asthma is a heterogeneous chronic airway disease in which inflammatory and structural processes vary across patients and over time[1]. Although the type 2 (T2) inflammatory paradigm has substantially improved molecular stratification and therapeutic targeting—particularly in eosinophilic asthma—the

conventional binary framework of T2 versus non-T2 remains incomplete. In clinical practice, many patients labelled as T2-low show marked variability in disease behaviour, treatment response and exacerbation risk, suggesting that T2-low asthma is unlikely to represent a single biologically uniform category[2]. Recent work has also emphasised the need to move from broad phenotypes towards molecularly grounded endotypes using integrative computational approaches.

Airway transcriptomic profiling has provided key insight into asthma endotypes by identifying epithelial and immune programs associated with T2 inflammation, eosinophilia and airway remodeling[3]. Core epithelial T2-associated genes, including SERPINB2, POSTN and CLCA1, have shown consistent value for molecular stratification[4]. However, much of the transcriptomic literature remains cohort-specific, with substantial differences in sampling source, cohort composition and clinical annotation. This heterogeneity complicates cross-study reproducibility and makes it difficult to distinguish stable biological structure from cohort-specific signal. Recent multicohort analyses underscore both the promise of cross-cohort molecular stratification and the need for rigorously reproducible analytical frameworks[5].

A further challenge is conceptual: asthma inflammation is often represented using discrete labels, despite evidence that key molecular processes—especially T2 inflammation—behave as continuous gradients rather than sharply separated states[6]. Dichotomisation can aid interpretation, but it may also reduce statistical power and obscure biologically meaningful variation[7], particularly when linking transcriptomic features to clinical phenotypes or when building predictive models intended to generalize across datasets.

Uncertainty is even greater in T2-low asthma. Compared with T2-high disease, T2-low asthma lacks widely adopted molecular biomarkers and has fewer mechanism-guided therapeutic options[8, 9]. Contemporary work in non-T2 severe asthma has implicated interferon (IFN)-related responses, inflammasome pathways, IL-17/neutrophilic inflammation and oxidative stress, but how these processes are organised within a reproducible cross-cohort molecular structure remains insufficiently resolved[10]. A central unresolved question is whether T2-low asthma is best viewed as a diffuse residual category or as a set of identifiable molecular subtypes with distinct program-level biology.

To address these gaps, we established a cross-cohort airway transcriptomic framework using six GEO datasets in a prespecified discovery–validation design, with cohort-aware validation to support generalizability across structured datasets[11]. The analytical strategy prioritized reproducibility and minimized information leakage through fold-restricted preprocessing, discovery-derived frozen thresholds and independent external validation, in line with emerging guidance emphasising rigorous design and risk-of-bias control in prediction modeling workflows[12]. We aimed to characterize (i) the stability of T2-associated airway signals across heterogeneous cohorts, (ii) a continuous epithelial axis capturing T2-related molecular variation, and (iii) structured molecular heterogeneity within T2-low asthma, including reproducible non-T2 subtype programs.

By integrating cross-cohort consensus differential expression, pathway- and network-level evidence, leakage-controlled predictive modeling and random-effects meta-analysis, this study tests whether asthma airway inflammation is better represented as a hierarchical immune architecture—comprising a dominant T2 gradient and internal program-level divergence within T2-low disease—rather than as a simple binary split. This conceptual and analytical direction is aligned with current efforts to use integrative molecular profiling to refine asthma stratification beyond conventional clinical labels[13], and with emerging spatially resolved atlases showing structured regional and cellular variation in diseased human lung tissue[40]. Such a framework may provide a more biologically aligned basis for biomarker development, patient stratification and prospective evaluation of mechanism-guided therapies[15].

Methods

Study design and datasets

Six GEO airway epithelial or sputum transcriptomic cohorts were included in this study[16]. Three cohorts (GSE67472, GSE41861, and GSE43696) were assigned to the discovery set for feature derivation and model development, whereas three independent cohorts (GSE45111, GSE143303, and GSE148004) were reserved for external validation (Supplementary File 1: Table S1). Eligible cohorts comprised human airway samples with phenotype annotations sufficient to support asthma classification and T2-related stratification. In total, 454 samples were analyzed. Discovery cohorts were used for model development under leave-one-dataset-out (LODO) evaluation, whereas validation cohorts were withheld from feature selection and model tuning.

General preprocessing and information leakage control

To minimize information leakage, all supervised preprocessing and modeling procedures were performed using training data only and then applied unchanged to held-out data. This principle was maintained throughout the analytical workflow, including imputation, feature filtering, scaling, standardization, and model fitting. For cross-cohort analyses, expression values were standardized within each GSE by gene-wise Z-scoring. For derived quantitative traits, an additional within-cohort standardization step was applied where appropriate to improve comparability across datasets. Genes absent from a given cohort were handled in a manner consistent with this within-cohort standardized framework.

Data preprocessing, cross-cohort integration, and differential expression analysis

For each dataset, expression profiles were matched to phenotype annotations by sample ID. Data were log2-transformed where necessary and quantile-normalized within each cohort[17]. Probes were

mapped to gene symbols using the corresponding GPL annotation files, and multiple probes mapping to the same gene were collapsed by their mean expression value.

For visualization-only quality control, discovery datasets were merged on shared genes and batch-corrected with ComBat while retaining biological group information in the design matrix[18]. These batch-corrected data were used exclusively for principal component analysis and quality assessment. All downstream analyses were conducted using within-cohort preprocessed matrices in order to preserve a leakage-free design.

T2 inflammation scoring and stratification

T2 inflammation was initially quantified using a core three-gene signature consisting of POSTN, CLCA1, and SERPINB2, hereafter referred to as the core3 score[4]. Within each cohort, the T2 score was calculated as the mean of gene-wise Z-scores for these genes. When T2 status was available from the original study annotation, it was harmonised as T2-high, T2-mid, or T2-low. Otherwise, samples were stratified according to within-cohort T2 score quantiles, with the upper and lower portions used to define T2-high and T2-low groups, respectively, and intermediate samples excluded from binary analyses. The robustness of stratification was assessed using score distributions, correlations between component genes and the composite score, and between-group comparisons.

In validation cohorts with eosinophilic phenotype information, T2-like groups were defined according to study-specific criteria, such as sputum eosinophils ≥ 3% versus < 3%, consistent with commonly used definitions of sputum eosinophilia in asthma[19]. As a sensitivity analysis, an extended T2 panel was derived from genes showing consistent positive association with the core3 score across discovery cohorts. Extended-panel scores were calculated in the same manner and compared with core3-based scores in both discovery and validation datasets.

Leakage-free baseline model for T2 classification

A baseline classifier for T2-high versus T2-low asthma was developed in the discovery set using elastic-net logistic regression under LODO evaluation[11, 14]. Predictors were restricted to genes shared across discovery cohorts and represented by within-cohort standardized expression values. Within each training fold, preprocessing procedures were fitted using training data only, and model hyperparameters were selected by nested cross-validation. Model performance was summarized using the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (AUPRC) [21]. After internal evaluation, the final model was trained on the full discovery set and tested in the independent validation cohorts using the prespecified preprocessing pipeline and locked feature set[22].

Discovery of a T2-associated epithelial axis

A T2-associated epithelial axis was defined in the discovery cohorts by integrating three complementary sources of evidence: differential expression, pathway activity, and co-expression network structure. Differentially expressed genes (DEGs) were identified by comparing T2-high and T2-low samples within each cohort using limma[20, 23], and consensus DEGs were defined on the basis of cross-cohort directional concordance. Pathway activity was assessed using single-sample gene set enrichment analysis, and network structure was examined using weighted gene co-expression network analysis (WGCNA)[24, 25]. Candidate axis genes were prioritized from overlapping evidence across these analyses, and the direction of each gene was assigned according to the consensus pattern observed across cohorts.

The T2 epithelial axis score was subsequently calculated from the aggregate expression of positively and negatively weighted axis genes. Detailed parameter settings and sensitivity analyses for DEG consensus rules, single-sample gene set enrichment analysis, and WGCNA are provided in the Supplementary File 1: Supplementary Methods.

Clinical association, meta-analysis, and predictive evaluation of the T2 epithelial axis

The T2 epithelial axis score was integrated with available clinical phenotype data after harmonisation of variable names and class labels. Thresholds used to derive categorical axis-defined groups were determined in the discovery set and then applied unchanged to the validation set. Associations between the axis score and clinical outcomes were assessed within each cohort using linear or logistic regression, depending on the endpoint, and cohort-specific estimates were then pooled by random-effects meta-analysis[26].

To evaluate predictive value, models incorporating the T2 epithelial axis were compared with baseline approaches using measures of discrimination and calibration, including AUROC, AUPRC, Brier score, calibration intercept, and calibration slope[27]. Internal evaluation was performed using dataset-level cross-validation, and external performance was assessed in the independent validation cohorts.

Explainable machine learning and derivation of a minimal T2 gene panel

To derive a parsimonious and interpretable transcriptomic signature of T2 status, an XGBoost[28] classifier was trained using axis-related genes under strict LODO evaluation across discovery cohorts. Feature importance was quantified using SHAP values[29], and genes that were repeatedly prioritized across folds were used to define a compact gene panel. The final model was retrained on the full discovery set and externally evaluated in the validation cohorts. Additional implementation details, including tuning procedures and panel selection criteria, are provided in the Supplementary File 1: Supplementary Methods.

Definition of non-T2 endotypes (E1 and E2) and predictive modeling

Within the non-T2 space, endotypes were defined using a rule-based strategy based on multiple standardized axis scores. Within each cohort, candidate axes were evaluated by unsupervised clustering, and the axis that showed the most stable within-cohort separation across datasets was selected as the primary non-T2 stratification axis[30, 31]. Samples in the higher-scoring cluster were labelled E1, whereas those in the lower-scoring cluster were labelled E2.

To assess reproducibility, these rule-based labels were then predicted using elastic-net logistic regression based on the set of axis scores. Model performance and calibration were evaluated under LODO analysis in the discovery cohorts and subsequently examined in the external validation datasets[11, 22]. Detailed clustering and calibration procedures are described in the Supplementary File 1: Supplementary Methods.

Minimal gene panel for E1 versus E2 within T2-low asthma

Within T2-low samples, a minimal transcriptomic panel for distinguishing E1 from E2 was derived using LODO LASSO logistic regression[14]. Candidate genes were restricted to the intersection of non-T2 axis genes and genes shared across the discovery cohorts. Class imbalance was addressed during model training, and gene selection was based on reproducibility across folds together with coefficient stability. The resulting locked panel was refitted in the full discovery set and externally validated in independent cohorts with E1/E2 labels[22]. Detailed feature selection rules, hyperparameter settings, and calibration procedures are provided in the Supplementary File 1: Supplementary Methods.

Differential expression and functional enrichment analysis for E1 versus E2

Within T2-low samples from the discovery cohorts, differential expression between E1 and E2 was evaluated using limma on within-cohort standardized gene expression[23]. Consensus signals were defined using prespecified directional concordance and significance criteria across cohorts, and robustness was examined by sensitivity analyses using alternative consensus rules and statistical thresholds. Because differences within T2-low asthma were expected to be subtler than those between T2-high and T2-low disease, emphasis was placed on cross-cohort consistency and pooled evidence rather than on any single cohort-specific DEG list.

Functional interpretation of E1/E2-associated genes was performed using over-representation analysis for MSigDB Hallmark and Gene Ontology Biological Process terms with clusterProfiler[32, 33]. Additional pooled-evidence analyses, including directional-consistency and meta-analytic approaches, are described in the Supplementary File 1: Supplementary Methods.

Mechanistic validation of E1/E2 endotypes

Mechanistic validation was performed in T2-low samples with definitive E1/E2 labels. In eligible cohorts, E1–E2 differential expression was re-evaluated using an empirical Bayes framework[23], and consensus genes were defined according to prespecified cross-cohort criteria. Transcription factor activity was inferred using the CollecTRI regulatory network[34]. Programme-level differences and associations with available clinical phenotypes were analyzed using covariate-adjusted regression models, and cohort-specific estimates were pooled by random-effects meta-analysis[26]. Transcription factor target-set enrichment and Gene Ontology enrichment analyses were further used to interpret subtype-associated molecular programs. Missing data were handled using complete-case analyses with adaptive covariate selection to minimize sample loss.

Results

Cross-cohort integration and preprocessing define the analytical framework and reveal stable asthma-related transcriptional differences

Six GEO-derived human airway transcriptomic cohorts (n = 454) were integrated into a discovery set and an independent validation set. Marked differences in sample composition were observed across cohorts, particularly in the proportions of bronchial epithelial and sputum samples (Fig. 1A). ComBat correction reduced cohort-related structure in principal component analysis and increased overlap among samples from different cohorts (Fig. 1B,C).

To characterize global transcriptional differences at the disease level, differential expression analysis comparing asthma and control samples identified significant upregulation of POSTN, CST1, CPA3, and KIT, together with downregulation of SCGB3A1 and C3 (adjusted $P < 0.05$; logFC > 0.585; Fig. 1D; Supplementary File 1).

Core3-based T2 scoring robustly stratifies samples across discovery and validation cohorts

A T2 inflammation score derived from the core genes SERPINB2, POSTN, and CLCA1 showed consistent positive correlations with expression of its component genes across all cohorts (Spearman's $\rho = 0.494$–0.954; discovery set, 0.801–0.954; validation set, 0.494–0.753; Fig. 2A). In the discovery cohorts, the T2 score increased stepwise across T2-low, T2-mid, and T2-high strata. The same gradient was reproduced in the independent validation cohorts (Fig. 2C,D).

In sensitivity analyses, most genes in the extended20 panel also showed positive associations with the core T2 score across cohorts (105 of 115 available gene–cohort correlations, 91.3%), with directions concordant with the core T2 signal (Fig. 2B). An extended20-based score likewise separated T2 strata in

both the discovery and validation sets and showed a more continuous distribution than the core3-based score (Fig. 2E,F; Supplementary File 2).

A leakage-controlled elastic-net baseline model generalizes across cohorts

Under leave-one-dataset-out (LODO) evaluation, an elastic-net logistic regression model stably discriminated T2-high from T2-low samples across the discovery cohorts (AUROC, 0.945–0.987; AUPRC, 0.961–0.987; median AUROC, 0.972; median AUPRC, 0.974; Fig. 3A). Good discrimination was retained in the independent validation cohorts, although cohort-level heterogeneity remained evident (Fig. 3B,C).

Inspection of model coefficients showed positive weights for canonical T2-associated genes, including SERPINB2, POSTN, and CLCA1, whereas IL18RAP carried a negative weight (Fig. 3D). Median expression distributions of epithelial, stromal, endothelial, and immune cell-type marker genes showed no global cross-cohort shift, arguing against a trivial cell-composition explanation for model performance (Fig. 3E; Supplementary File 3).

A T2-associated epithelial axis captures T2 biology and retains external predictive value

In the discovery cohorts, the T2 epithelial axis score correlated positively with the T2 score (Spearman's $\rho \approx 0.76–0.78$ across cohorts) and clearly discriminated T2-high from T2-low samples (Fig. 4A,B). A classifier trained on axis-related features retained good discriminative performance under LODO evaluation in the discovery set and in the independent validation cohorts (Fig. 4C–E). SHAP-based feature attribution identified SERPINB2, CLCA1, and POSTN as stable major contributors. Consistent with the elastic-net baseline model, the T2 epithelial axis score also separated T2-high from T2-low samples across both discovery and validation cohorts, with concordant effect directions (Fig. 4F,G; Supplementary Files 4–8).

Clinical generalizability was further evaluated in eosinophilic phenotype-labelled cohorts (GSE143303 and GSE45111). In these cohorts, the T2 epithelial axis score remained positively associated with eosinophilic phenotype after adjustment for age and sex (random-effects pooled OR = 1.77, 95% CI 1.18–2.65, P = 0.0059; I^2 = 0%; Supplementary File 9). The prespecified axis-high group showed a directionally consistent but weaker association (OR = 2.91, P = 0.053; Supplementary File 9).

Non-T2 endotypes (E1 and E2) are reproducible across cohorts and can be predicted from **axis-level features**

Within T2-low samples, two non-T2 endotypes, E1 and E2, were identified and aligned across cohorts. Relative to E2, E1 showed higher mean scores on inflammation- and stress-related axes, including IFN_Antiviral, IL1_Inflammasome, Th17/Neutrophil, and OxStress_NRF2, together with opposite trends

for the BarrierRepair and EMT/Fibrosis axes (Fig. 5A). The proportional distribution of E1 and E2 was similar between the discovery and validation cohorts (Fig. 5B).

A model trained on subtype labels stably discriminated E1 from E2 under discovery LODO evaluation, and this performance was retained in the independent validation cohorts (Fig. 5C–E). Feature-importance analysis assigned the largest weight to the IFN_Antiviral axis (|β| = 1.11), followed by the Th17/Neutrophil and IL1_Inflammasome axes (Fig. 5F; Supplementary File 10).

A minimal E1/E2 panel identifies a reproducible inflammatory-like program within T2-low asthma

Among T2-low samples in the discovery cohorts, a locked minimal-feature classifier robustly discriminated E1 from E2 (pooled AUROC = 0.899; pooled AUPRC = 0.766) and showed good calibration. CXCL10, CASP1, IFIT2, and IFI27 were consistently the dominant contributors, with concordant coefficient directions and stable selection frequencies under LODO evaluation.

Cross-cohort differential expression did not yield stringent consensus genes at FDR < 0.05. However, under prespecified sensitivity criteria (FDR < 0.10, concordant direction, and cross-cohort support), reproducible E1/E2-associated signals emerged across alternative consensus rules, yielding 266 consensus genes. Integration of REML random-effects meta-analysis with combined P-value methods further expanded the candidate set to 330 genes (303 E1-like and 27 E2-like), supporting cross-cohort reproducibility. E1-like candidates were significantly enriched for immune and inflammatory pathways, including interferon-γ response, interferon-α response, TNFα signalling via NFκB, and inflammatory response (Supplementary Files 11–12).

Consistent with these findings, marker-program analysis showed a broad shift of IFN- and inflammation-related programs towards E1, including IFN_Antiviral, IL1_Inflammasome, and Th17/Neutrophil. Several of these associations remained significant after random-effects meta-analysis and multiple-testing correction. By contrast, epithelial homeostasis/repair and oxidative stress-related programs showed no stable E2 bias and were generally weak or non-significant. Cohort-level heatmaps recapitulated these patterns (Fig. 6E), supporting a reproducible and biologically interpretable molecular divergence between E1 and E2 (Supplementary File 13).

Discussion

Clinically, our findings suggest that asthma airway inflammation may be stratified more reproducibly, and in a way that may better align with biologically meaningful treatment-related heterogeneity, by combining a continuous T2 burden axis with structured variation within T2-low disease rather than relying solely on the conventional binary T2/non-T2 framework. Based on six GEO airway transcriptomic cohorts from patients with asthma (n = 454), this study systematically characterized T2-related airway signals and the molecular organisation of T2-low asthma within a cross-cohort integration and independent validation

framework. Three mutually reinforcing observations support this model. First, core T2-associated molecular signals showed stable directional concordance across cohorts despite differences in sample source and cohort composition. Second, a continuous epithelial inflammatory axis derived from core T2 genes robustly captured T2 status in asthma and was consistent with eosinophilic phenotypes. Third, T2-low asthma did not represent a single homogeneous “non-T2” category, but instead comprised reproducible molecular endotypes across cohorts, one of which was dominated by IFN/inflammatory programs. Taken together, these findings support a more structured view of asthma inflammation as an immune architecture comprising both a principal gradient and internal modular differentiation, rather than a simple two-class system[2, 5].

Despite substantial differences in sample type and cohort composition, core T2 genes such as SERPINB2, POSTN and CLCA1 retained concordant expression trends across asthma cohorts, consistent in direction with previous transcriptomic studies of T2/eosinophilic asthma[2]. This consistency indicates that T2-associated epithelial signals are not merely dataset-specific observations, but represent a stable and transferable molecular dimension across heterogeneous asthma populations. Notwithstanding differences in disease severity, inhaled corticosteroid use and airway sampling approaches, these signals remained directionally robust, supporting their utility as a reproducible basis for cross-cohort stratification.

Notably, both the core three-gene score and the extended-panel score showed gradient-like distributions rather than two sharply separated groups. This pattern supports the existence of a continuous T2 inflammatory spectrum in asthma and suggests that T2 inflammation is better understood as a quantifiable dimension of immune activity than as a fixed categorical label. From a clinical perspective, this view may be more informative than binary classification, because therapeutic response—particularly to biologics targeting IL-5 or IL-4/IL-13 pathways—is also likely to vary along a gradient of inflammatory intensity rather than being determined entirely by arbitrary thresholds[6]. Thus, a continuous T2 axis may capture clinically relevant biological variation that is obscured by dichotomisation.

On the basis of these observations, we constructed a T2-associated epithelial axis score. This continuous T2 axis score stably discriminated T2-high from T2-low individuals in both discovery and validation cohorts and was significantly associated with eosinophilic phenotypes. Importantly, it was not simply statistically convenient; it also appeared biologically more informative than dichotomised grouping, showing stronger associations and greater cross-cohort stability. These findings suggest that discretisation of continuous biological processes can mask meaningful variation in asthma. Mechanistically, the epithelial T2 axis is likely to reflect not only the presence or absence of IL-4/IL-13-driven signalling, but also the integrated output of epithelial barrier remodeling, mucus-secretory program activation and amplification of eosinophilic inflammation[6, 35]. The T2 epithelial axis should therefore be viewed not only as a stratification tool, but also as a composite readout of T2-related pathological intensity in asthma.

Beyond the T2 principal axis, one of the most mechanistically informative findings was the identification of an E1 endotype within T2-low asthma characterized by IFN/inflammatory programs. Across multiple cohorts, E1 showed higher scores for IFN_Antiviral, IL1_Inflammasome, Th17/Neutrophil and OxStress_NRF2 programs. Enhanced IFN antiviral programs may reflect a molecular background related to viral exposure or antiviral responses, raising the possibility that this subtype is linked to susceptibility to virus-associated exacerbations[36]. Increased IL-1 inflammasome and Th17/neutrophil-related signals are also consistent with pathway features previously associated with neutrophilic phenotypes and reduced corticosteroid responsiveness[8]. In addition, the involvement of oxidative stress-related programs supports a broader framework of innate immune activation in non-T2 asthma. Together, these findings suggest that E1 represents a biologically distinct T2-low molecular subtype enriched for innate immune, antiviral and inflammasome-related activity, rather than a generic “non-T2” state.

By contrast, E2 did not display a dominant program that mirrored E1 with comparable magnitude and opposite direction, but instead resembled a background state with globally lower or more diffuse inflammatory signalling. This asymmetry suggests that variation within the T2-low space is driven primarily by the degree and composition of inflammatory activation rather than by a simple opposition between two symmetrical mechanistic modules. Accordingly, T2-low asthma is better understood as a biologically structured domain than as a residual negative class. Given the current lack of reliable biomarkers and targetable therapeutic strategies for T2-low asthma, this perspective provides a testable framework for future mechanistic validation and treatment-oriented subgrouping[9, 37].

An important point is that the E1/E2 conclusion did not depend solely on a lenient single-gene differential-expression threshold. Because differences within T2-low asthma are expected to be subtler than the broader contrast between T2-high and T2-low disease, a sensitivity-oriented framework was necessary to reduce false negatives arising from modest effect sizes, platform differences and between-cohort heterogeneity, while avoiding overinterpretation of cohort-specific noise. For this reason, the E1/E2 comparison was interpreted primarily through convergent evidence—directional consistency across cohorts, pathway-level reproducibility and external model performance—rather than through the expectation that every informative gene would meet the same stringent FDR threshold in every dataset. Consistent with this interpretation, a minimal-feature model built from E1/E2 labels retained stable discriminative performance in external cohorts, and its key genes were all involved in IFN- or inflammasome-related pathways[10]. Although single-gene analyses did not yield a highly consistent cross-cohort list of significant genes, pathway-level and model-level signals were substantially more robust under a directional-consistency framework[38, 39]. Taken together, these findings indicate that the E1/E2 distinction is better understood as a program-level stratification signal than as a fixed DEG list.

Overall, these results support a layered and mechanistically informed model of asthma inflammation in which a continuous epithelial T2 principal axis is superimposed on subtype-level divergence within T2-low disease. More importantly, from a translational perspective, this framework suggests that asthma stratification may benefit from moving beyond label-based classification towards the joint consideration

of axis intensity (for T2 burden) and program composition (for T2-low heterogeneity), which may better reflect biological variation and gradients of treatment response.

Several limitations should be acknowledged. The analyses were based on public bulk transcriptomic data, making it difficult to fully exclude the influence of changes in cellular composition. Stratification was established primarily at the molecular level, and its relationships with exacerbation risk, corticosteroid responsiveness and biologic treatment efficacy require prospective validation. The cross-sectional design also precludes assessment of the temporal stability of inflammatory axes. In addition, because the E1/E2 distinction was derived from integrated public datasets rather than a prospectively recruited cohort, its clinical relevance should be regarded as hypothesis-generating until confirmed in independent longitudinal studies with harmonised phenotyping.

Future work should prioritise prospective and longitudinal validation of this two-layer framework. In particular, it will be important to determine whether the continuous T2 epithelial axis predicts gradients of biologic response, whether the E1 endotype is enriched in patients with virus-associated exacerbations or poor corticosteroid responsiveness, and whether these transcriptomic structures remain stable over time or shift with treatment and disease activity. Integration with single-cell or spatial transcriptomics will also be required to resolve the cellular origins of the observed signals and to define how epithelial, myeloid and lymphoid compartments contribute to the principal-axis and subtype architecture[13, 40]. In summary, these findings support a stratification framework for asthma airway inflammation that combines a continuous T2 axis with reproducible program-level heterogeneity within T2-low disease, providing a more reproducible and potentially more treatment-relevant representation of airway inflammatory biology than the conventional T2/non-T2 dichotomy[15].

Conclusions

In conclusion, this study supports a stratification framework for asthma airway inflammation that combines a continuous T2 epithelial axis with reproducible program-level heterogeneity within T2-low disease. Rather than treating T2-low asthma as a uniform residual category, this framework provides a more reproducible and biologically informative representation of airway inflammatory variation. These findings offer a practical basis for biomarker development and for future validation of mechanism-guided stratification, particularly in clinically difficult T2-low populations.

Declarations

Ethics approval and consent to participate

The original studies included in the GEO datasets were conducted with appropriate ethical approval and informed consent as reported in the source publications. The present study analyzed publicly available, de-identified transcriptomic datasets from the NCBI Gene Expression Omnibus (GEO), including GSE67472, GSE41861, GSE43696, GSE45111, GSE143303, and GSE148004. No new participants were recruited and no new samples were collected for this study. The authors did not have access to

information that could identify individual participants during or after data collection. Therefore, additional institutional ethics approval and informed consent were not required for this secondary analysis.

Consent for publication

All authors read and approved the final manuscript and agreed to be accountable for all aspects of the work.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary files

Supplementary File1-14: Supplementary tables and analytical outputs, including dataset metadata, differential expression results, model outputs, sensitivity analyses, and supporting summary tables.

Supplementary Figure: S1 Fig-S10 Fig supporting the main analyses.

Funding

This work was supported by the Sichuan Provincial Administration of Traditional Chinese Medicine (25ZDIZX017); the Science and Technology Department of Sichuan Province Provincial Scientific Research Institute Basic Scientific Research Project (2024JDKY0027, 2024-4-804, 2024-4-808); and the Sichuan Provincial Administration of Traditional Chinese Medicine Special Research Project on TCM Science and Technology (2020JC0089, 2024MS152, 2024MS518). The funders had no role in study design, data collection and analysis, interpretation of data, writing of the manuscript, or the decision to submit the article for publication.

Author Contribution

Kairui Meng: Conceptualization; Methodology; Software; Formal analysis; Data curation; Visualization; Writing – Original Draft. Yuanming Liu: Resources; Clinical interpretation of results; Funding acquisition. Lu Wang: Software; Formal analysis; Visualization; Writing – Review & Editing. Xueqing Gong: Methodology; Software; Formal analysis; Visualization. Wenjun Tang: Resources; Interpretation of results. Guobing Jia: Investigation; Data curation. Chengshi He: Supervision; Project administration; Funding acquisition. Yanmei Wang: Supervision; Project administration; Funding acquisition.

Acknowledgements

The authors gratefully acknowledge the Gene Expression Omnibus (GEO) database for providing valuable data resources essential for this research.

Data Availability

All the data are contained in the attachment.

References

1. Wenzel SE. Asthma phenotypes: the evolution from clinical to molecular approaches. Nat Med. 2012;18(5):716–25.
2. Ray A, Das J, Wenzel SE. Determining asthma endotypes and outcomes: complementing existing clinical practice with modern machine learning. Cell Rep Med. 2022;3(12):100857.
3. Jackson ND, et al. Single-cell and population transcriptomics reveal pan-epithelial remodeling in type 2-high asthma. Cell Rep. 2020;32(1):107872.
4. Peters MC, et al. Measures of gene expression in sputum cells can identify TH2-high and TH2-low subtypes of asthma. J Allergy Clin Immunol. 2014;133(2):388–94.
5. Lee I, et al. Multicohort analysis of bronchial epithelial cell expression in healthy subjects and patients with asthma reveals four clinically distinct clusters. Am J Respir Cell Mol Biol. 2025;73(1):73–87.
6. Jakiela B, et al. Type 2 gene expression signature in severe asthma associates with more advanced airway remodeling. Clin Transl Allergy. 2025;15(6):e70060.
7. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ. 2006;332(7549):1080.
8. Liu T, Woodruff PG, Zhou X. Advances in non-type 2 severe asthma: from molecular insights to novel treatment strategies. Eur Respir J. 2024;64(2).
9. Ricciardolo FLM, Carriero V, Bertolini F. Which therapy for non-Type(T)2/T2-low asthma. J Pers Med. 2021;12(1).
10. Khalil BA, Sharif-Askari NS, Halwani R. Role of inflammasome in severe, steroid-resistant asthma. Curr Res Immunol. 2023;4:100061.
11. Roberts D et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2016;40.
12. Collins GS, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open. 2021;11(7):e048008.
13. Macowan M, et al. Deep multiomic profiling reveals molecular signatures that underpin preschool wheeze and asthma. J Allergy Clin Immunol. 2025;155(1):94–106.

14. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22.
15. Quon S, et al. Technological advances in the search for a CURE for asthma. Med J Aust. 2025;223(Suppl 10):S36–41.
16. Clough E, et al. NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. Nucleic Acids Res. 2024;52(D1):D138–44.
17. Bolstad BM, et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19(2):185–93.
18. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–27.
19. Westerhof GA, et al. Biomarkers to identify sputum eosinophilia in different adult asthma phenotypes. Eur Respir J. 2015;46(3):688–96.
20. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol. 1995;57(1):289–300.
21. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015;10(3):e0118432.
22. Collins GS, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
23. Ritchie ME, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
24. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.
25. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
26. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Softw. 2010;36(3):1–48.
27. Debray TPA, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data (TRIPOD-Cluster): explanation and elaboration. BMJ. 2023;380:e071058.
28. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: Association for Computing Machinery; 2016. pp. 785–794.
29. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Neural Inform Process Syst. 2017.
30. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press; 1967. pp. 281–297.

31. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.
32. Wu T, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation. 2021;2(3):100141.
33. Liberzon A, et al. The Molecular Signatures Database hallmark gene set collection. Cell Syst. 2015;1(6):417–25.
34. Müller-Dott S, et al. Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. Nucleic Acids Res. 2023;51(20):10934–49.
35. Siddiqui S et al. Epithelial miR-141 regulates IL-13-induced airway mucus production. JCI Insight. 2021;6(5).
36. Doni Jayavelu N et al. Bronchial epithelial transcriptome reveals dysregulated interferon and inflammatory responses to rhinovirus in exacerbation-prone pediatric asthma. JCI Insight. 2025;10(24).
37. Kim SR. Next-generation therapeutic approaches for uncontrolled asthma: insights into the heterogeneity of non-type 2 inflammation. Allergy Asthma Immunol Res. 2024;16(1):1–5.
38. Tseng GC, Ghosh D, Feingold E. Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res. 2012;40(9):3785–99.
39. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
40. Firsova AB, et al. Spatial single-cell atlas reveals regional variations in healthy and diseased human lung. Nat Commun. 2025;16(1):9745.

Figures

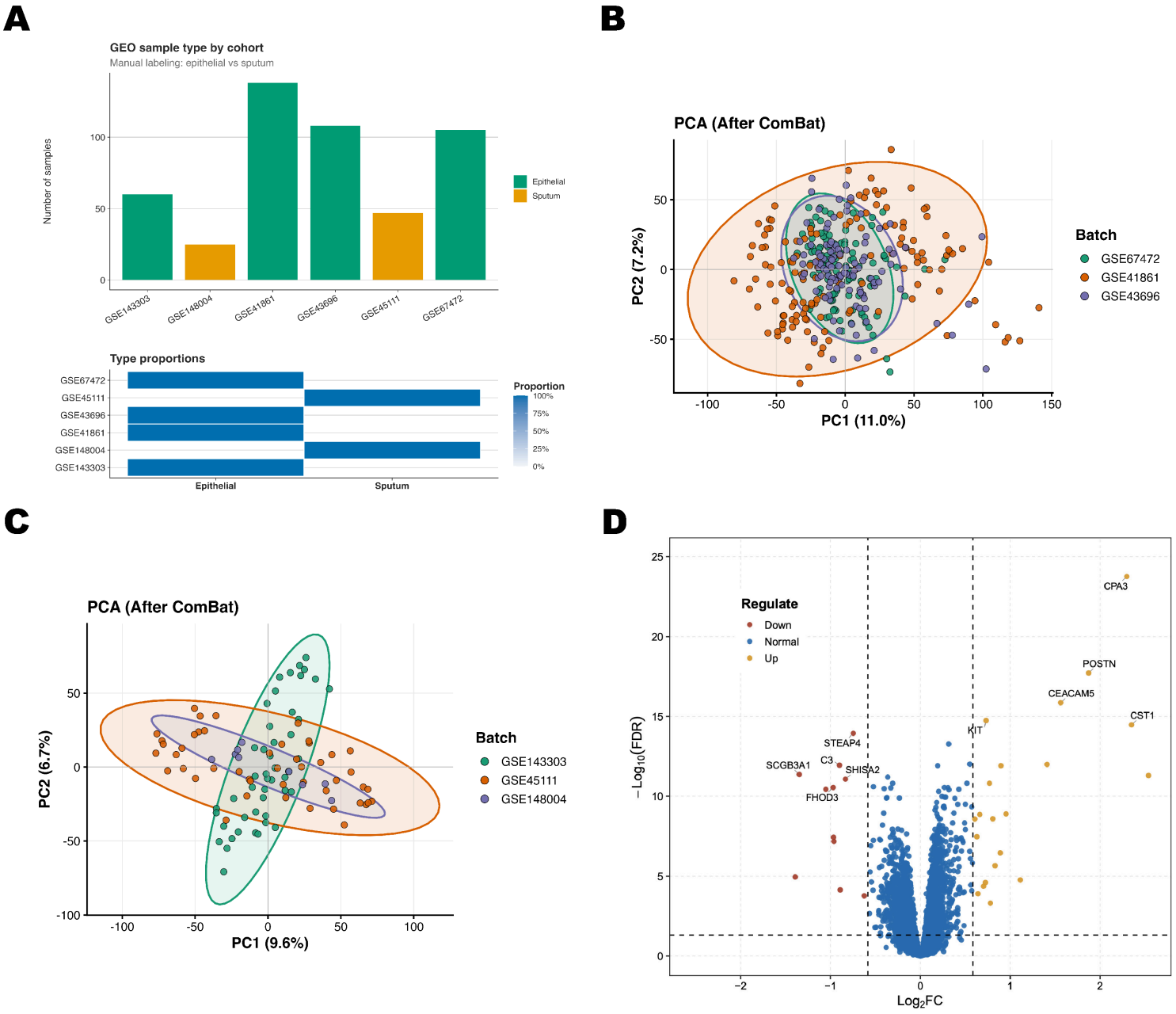


Figure 1

Cross-cohort integration and preprocessing define the analytical framework and reveal stable asthma-related transcriptional differences. Six GEO-derived cohorts were integrated into discovery and validation sets. (A) Sample-type composition by cohort (bronchial epithelium, green; sputum, yellow). (B–C) PCA after ComBat batch correction for discovery (B) and validation (C), with points colored by cohort. (D) Volcano plot of asthma vs control differential expression; colours denote upregulated (yellow), downregulated (red) and non-significant (blue) genes; dashed lines indicate significance thresholds (FDR < 0.05; −log10(FDR) = 1.30; |log2FC| > 0.585). Labelled genes highlight canonical T2-associated markers. Volcano plots for the validation cohorts are shown in S1 Fig.

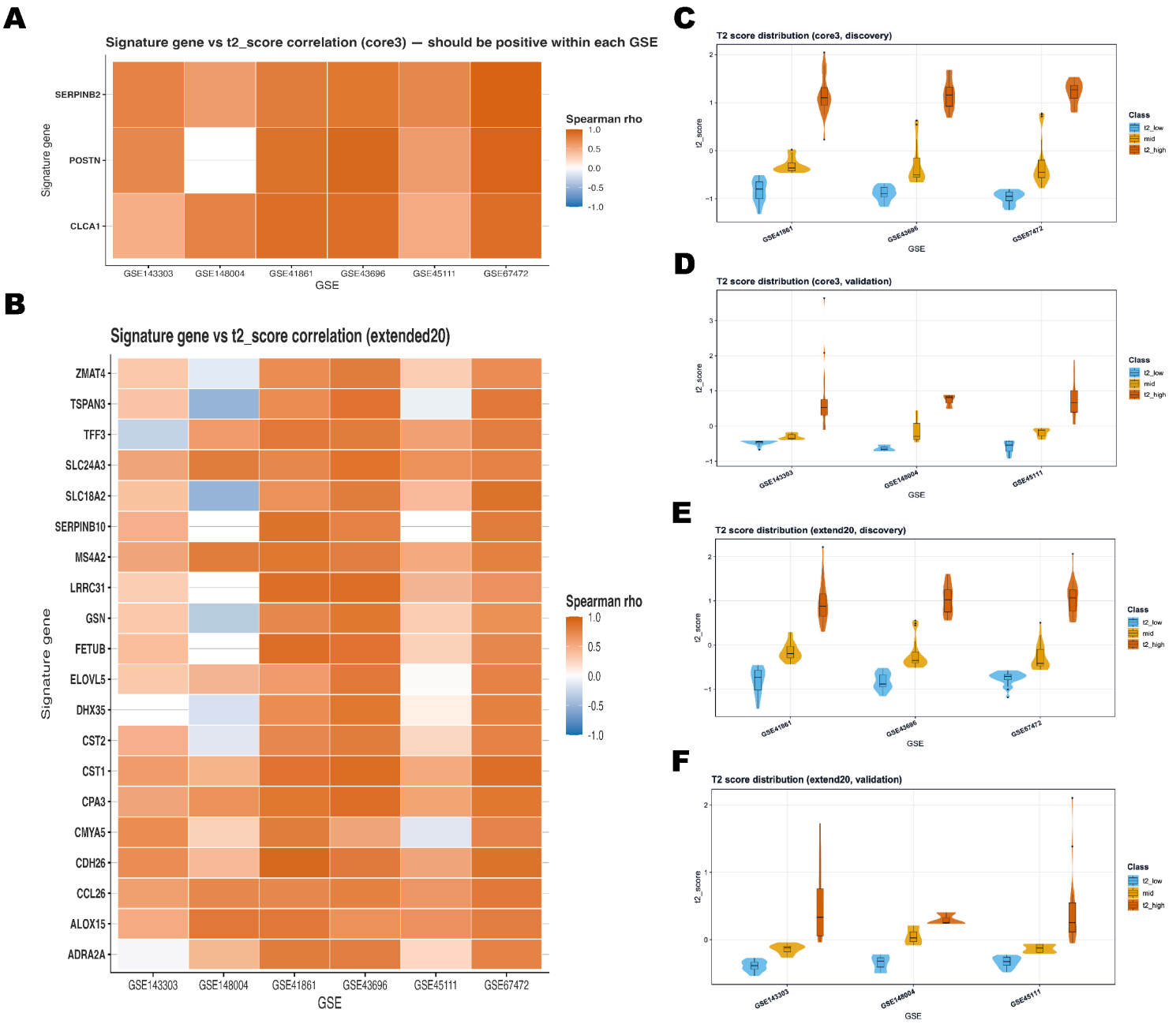


Figure 2

Core3-derived T2 scoring robustly stratifies samples across discovery and validation cohorts. (A) Within-cohort Spearman correlations between the core3-derived T2 inflammation score (T2 score; derived from SERPINB2, POSTN and CLCA1) and expression of each core gene (overall ρ = 0.494–0.954; discovery 0.801–0.954; validation 0.494–0.753). (B) Sensitivity analysis of Spearman correlations between T2 score and genes tested from the extended20 gene set, with most associations positive and directionally concordant with the core T2 signal. (C–D) Violin plots of core3 T2 score across T2-low, T2-mid and T2-high strata in the discovery (C) and independent validation (D) sets. T2 strata were taken from study annotations when available; otherwise, strata were defined by within-cohort T2 score quantiles (see Methods). (E–F) Violin plots of the extended20-derived score in the discovery (E) and validation (F) sets. In C–F, embedded

boxplots indicate the median and IQR (whiskers, 1.5×IQR) and colours denote T2 strata (T2-low, blue; T2-mid, yellow; T2-high, orange).

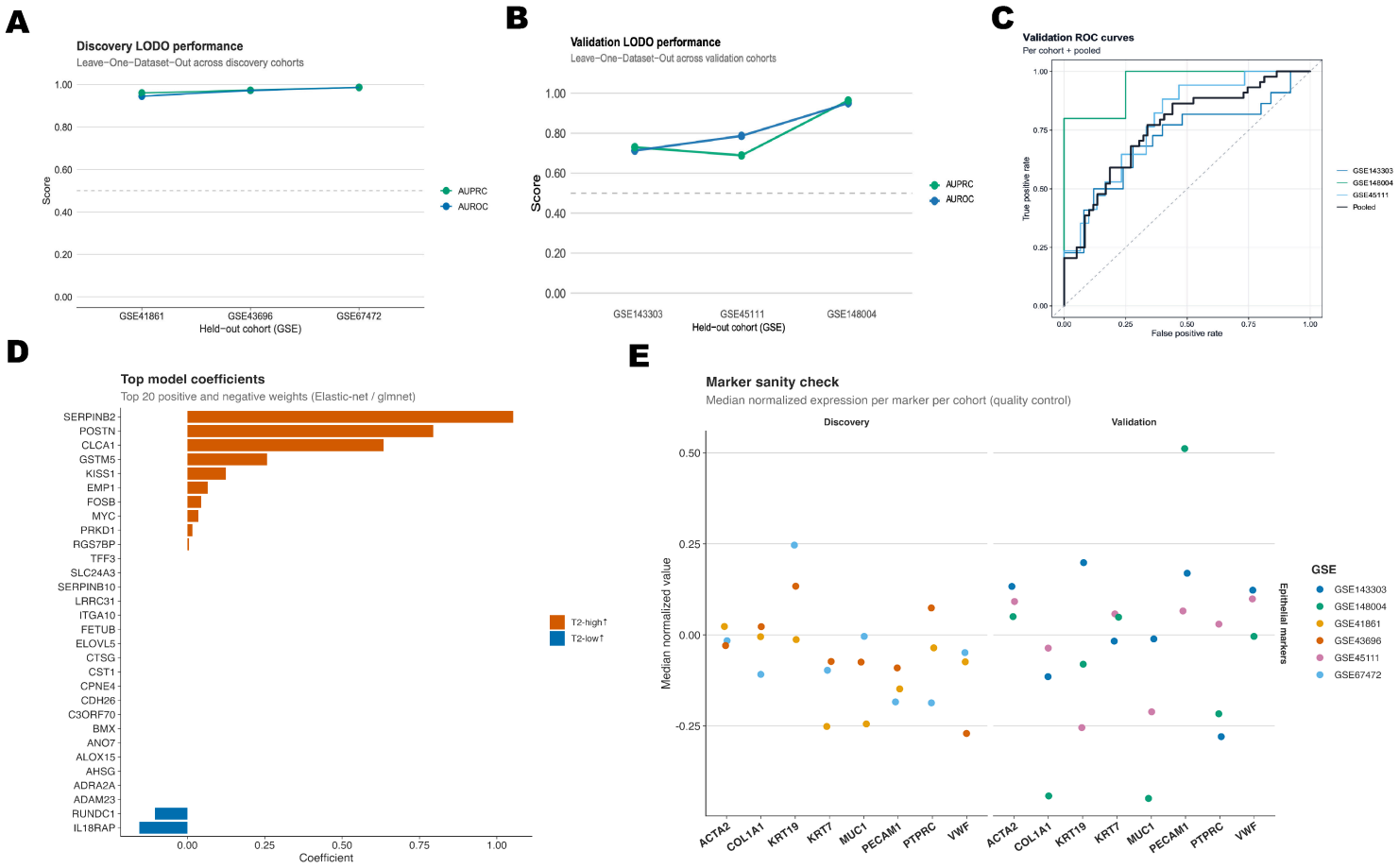


Figure 3

A leakage-controlled elastic-net baseline model generalizes across cohorts. Cohort-wise standardization and a leave-one-dataset-out (LODO) external generalisation framework were used to evaluate an elastic-net logistic regression classifier for discriminating T2-high from T2-low samples. (A) Discovery LODO performance across held-out cohorts (GSE41861, GSE43696, GSE67472), reporting AUROC and AUPRC. (B) External validation performance by cohort (model trained on all discovery data and tested separately in GSE143303, GSE45111 and GSE148004), reporting AUROC and AUPRC. (C) Validation ROC curves shown per cohort and pooled; the diagonal line denotes the no-discrimination reference. (D) Elastic-net coefficient weights (positive vs negative coefficients shown by bar colour), highlighting canonical T2-associated genes (SERPINB2, POSTN, CLCA1) and IL18RAP. (E) Marker sanity check showing median normalized expression of epithelial, stromal, endothelial and immune cell-type markers across discovery and validation cohorts (points colored by GSE).

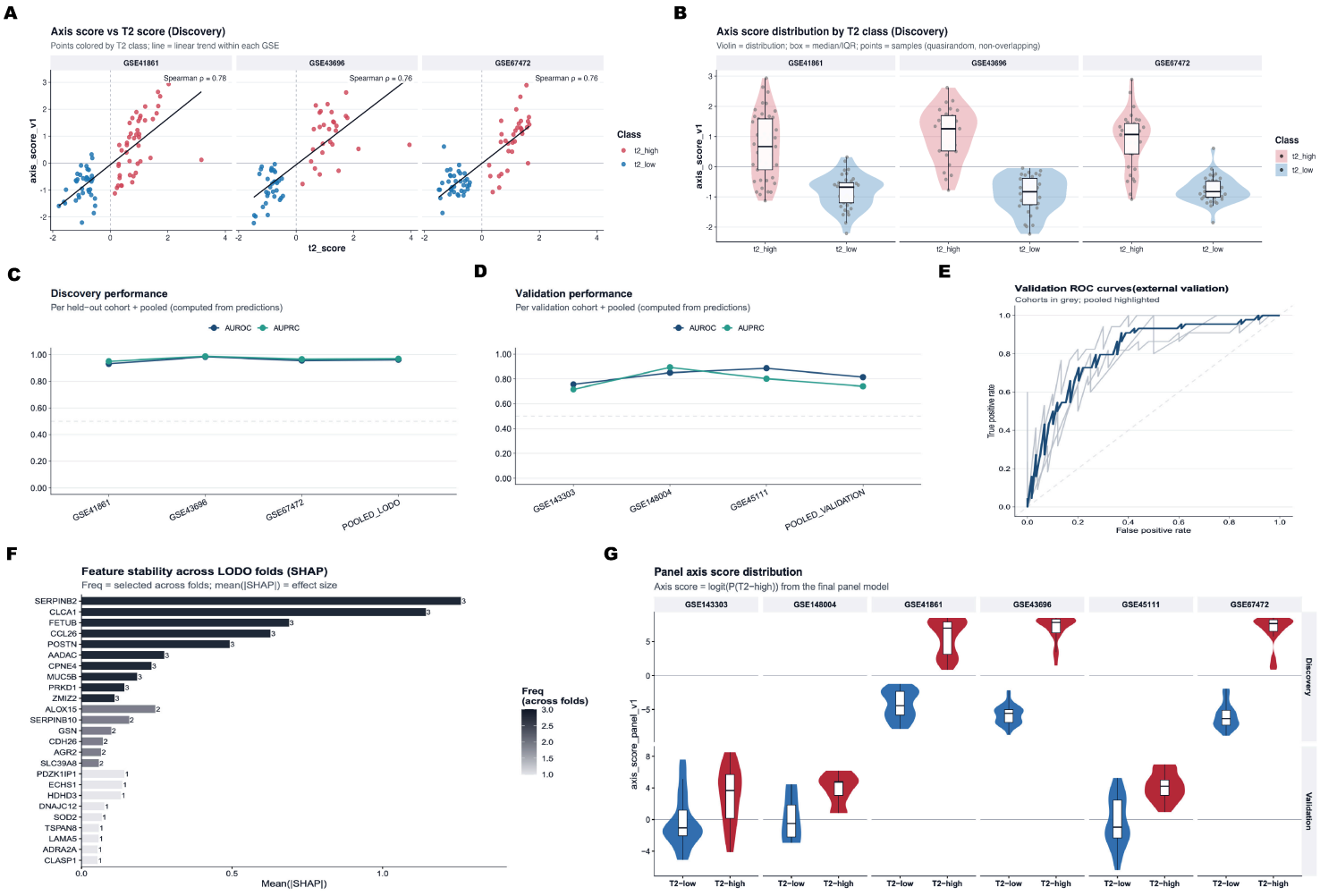


Figure 4

A T2-associated epithelial axis captures T2 biology and retains predictive value in external validation. (A) In discovery cohorts (GSE41861, GSE43696, GSE67472), the T2 epithelial axis score correlates with T2 score; points are colored by T2 class and the black line shows the within-cohort linear trend (Spearman's ρ shown in-panel). (B) Violin plots with embedded boxplots show the T2 epithelial axis score separation of T2-high and T2-low samples within each discovery cohort (box, median and IQR; whiskers, 1.5×IQR; points, samples). (C) Discovery LODO evaluation of an epithelial-axis feature classifier, reporting AUROC/AUPRC per held-out cohort and pooled. (D) External validation performance in GSE143303, GSE148004 and GSE45111, shown per cohort and pooled (AUROC/AUPRC). (E) External validation ROC curves (per cohort, grey; pooled highlighted), with the diagonal no-discrimination reference. (F) SHAP feature stability across discovery LODO folds (selection frequency and mean |SHAP|), highlighting SERPINB2, CLCA1 and POSTN. (G) Final panel-based axis score (logit[P(T2-high)]) across discovery and validation cohorts, shown as violin plots with embedded boxplots (box, median and IQR; whiskers, 1.5×IQR) stratified by phenotype.

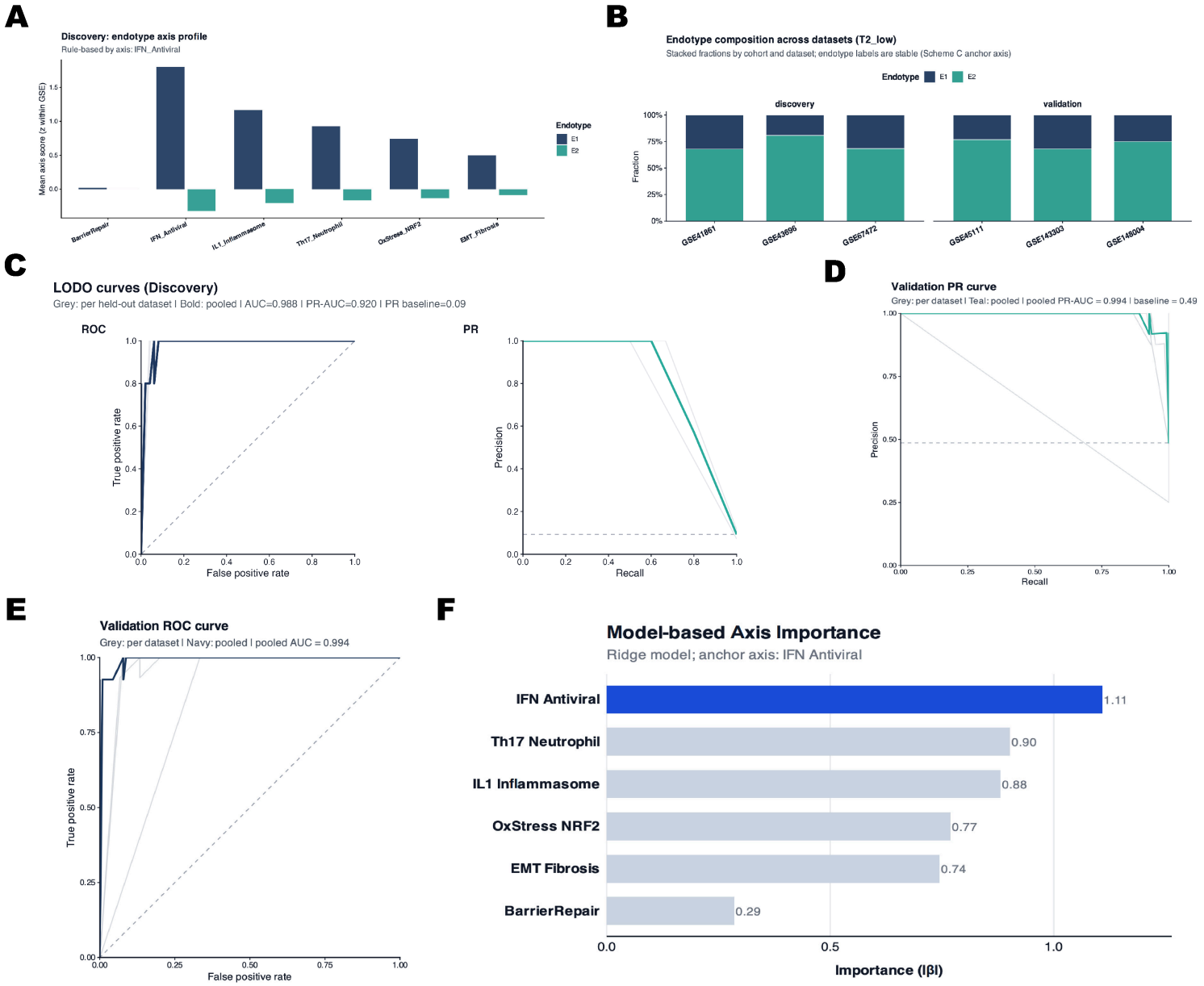


Figure 5

Non-T2 endotypes (E1 and E2) are reproducible across cohorts and can be predicted from axis-level features. T2-low samples were assigned to two non-T2 endotypes (E1/E2) using a rule-based, IFN_Antiviral-anchored clustering scheme applied within each GSE. Discovery cohorts comprised GSE41861, GSE43696 and GSE67472, and validation cohorts comprised GSE45111, GSE143303 and GSE148004. (A) Discovery endotype axis profile showing mean zWithinGSE-standardized scores for six non-T2 axes by endotype (E1 vs E2); the corresponding validation profile is shown in the S8 Fig. (B) Endotype composition across discovery and validation cohorts (stacked fractions). (C) Discovery LODO evaluation of an elastic-net logistic regression trained on the six axis scores, shown as per held-out dataset (grey) and pooled (bold) ROC and PR curves; dashed lines indicate the no-discrimination reference (ROC) and the PR prevalence baseline. (D–E) External validation PR and ROC curves shown per dataset (grey) with pooled curves highlighted (teal for PR; navy for ROC); dashed lines indicate the no-discrimination reference (ROC) and PR prevalence baseline (pooled AUROC = 0.994; pooled AUPRC =

0.994). (F) Ridge-model axis importance (|β|) for endotype prediction, highlighting IFN_Antiviral as the dominant contributor (|β| = 1.11), followed by Th17/Neutrophil and IL1_Inflammasome.

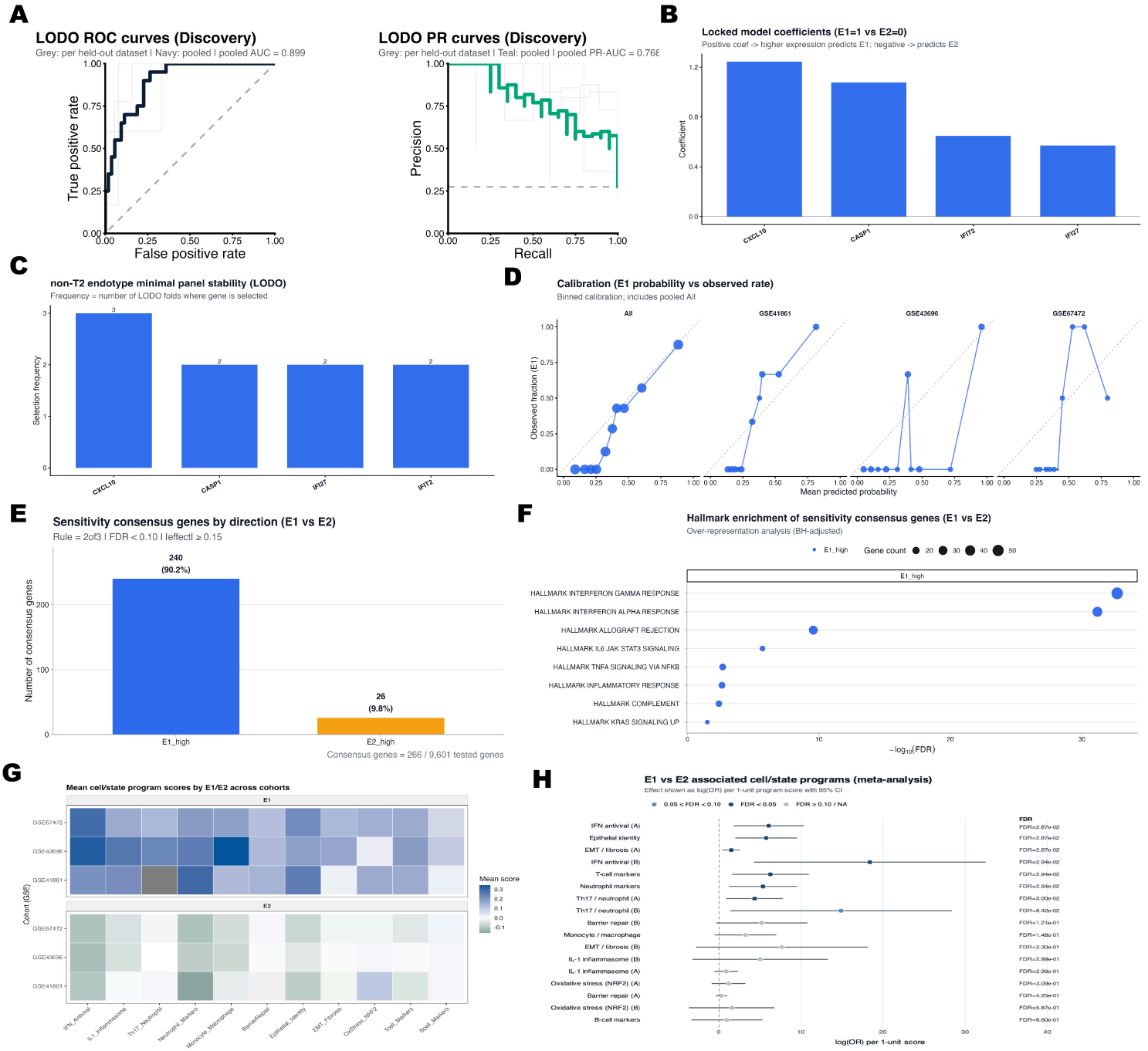


Figure 6

A minimal E1/E2 panel identifies a reproducible inflammatory-like E1 program in T2-low samples. Analyses were restricted to T2-low samples in the discovery cohorts (GSE41861, GSE43696, GSE67472) unless otherwise indicated. (A) LODO performance of a locked minimal-feature classifier discriminating E1 vs E2, shown as ROC and precision–recall (PR) curves for each held-out cohort (grey) with the pooled curve highlighted (navy for ROC; teal for PR); dashed lines denote the ROC no-discrimination reference and the PR prevalence baseline (pooled AUROC = 0.899; pooled AUPRC = 0.766). (B) Locked-model

coefficients (E1 = 1, E2 = 0), highlighting CXCL10, CASP1, IFIT2 and IFI27 as dominant contributors (positive coefficients indicate higher expression predicts E1). (C) Feature selection stability under LODO (selection frequency across folds). (D) Calibration of predicted E1 probabilities (binned mean predicted probability vs observed E1 rate), shown pooled and per cohort. (E) Sensitivity consensus analysis of cross-cohort differential expression direction (E1 vs E2) under the prespecified rule (≥2/3 cohorts supported at FDR < 0.10 and |effect| ≥ 0.15), yielding 266 consensus genes out of 9,601 tested (240 E1-high; 26 E2-high). (F) Hallmark over-representation analysis (BH-adjusted) for E1-high consensus genes (x-axis, −log10[FDR]; point size, gene count). (G) Cohort-level heatmap of mean program scores by endotype. (H) Random-effects meta-analysis of program–endotype associations (effect sizes with 95% CI; multiple-testing correction), showing reproducible E1-biased inflammatory/IFN programs and generally weak or non-significant E2 biases.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile1.xlsx
- SupplementaryFile2.xlsx
- SupplementaryFile3.xlsx
- SupplementaryFile4.xlsx
- SupplementaryFile5.xlsx
- SupplementaryFile6.xlsx
- SupplementaryFile7.xlsx
- SupplementaryFile8.xlsx
- SupplementaryFile9.xlsx
- SupplementaryFile10.xlsx
- SupplementaryFile11.xlsx
- SupplementaryFile12.xlsx
- SupplementaryFile14.docx
- SupplementaryFile13.xlsx
- SupplementaryFigure.pdf