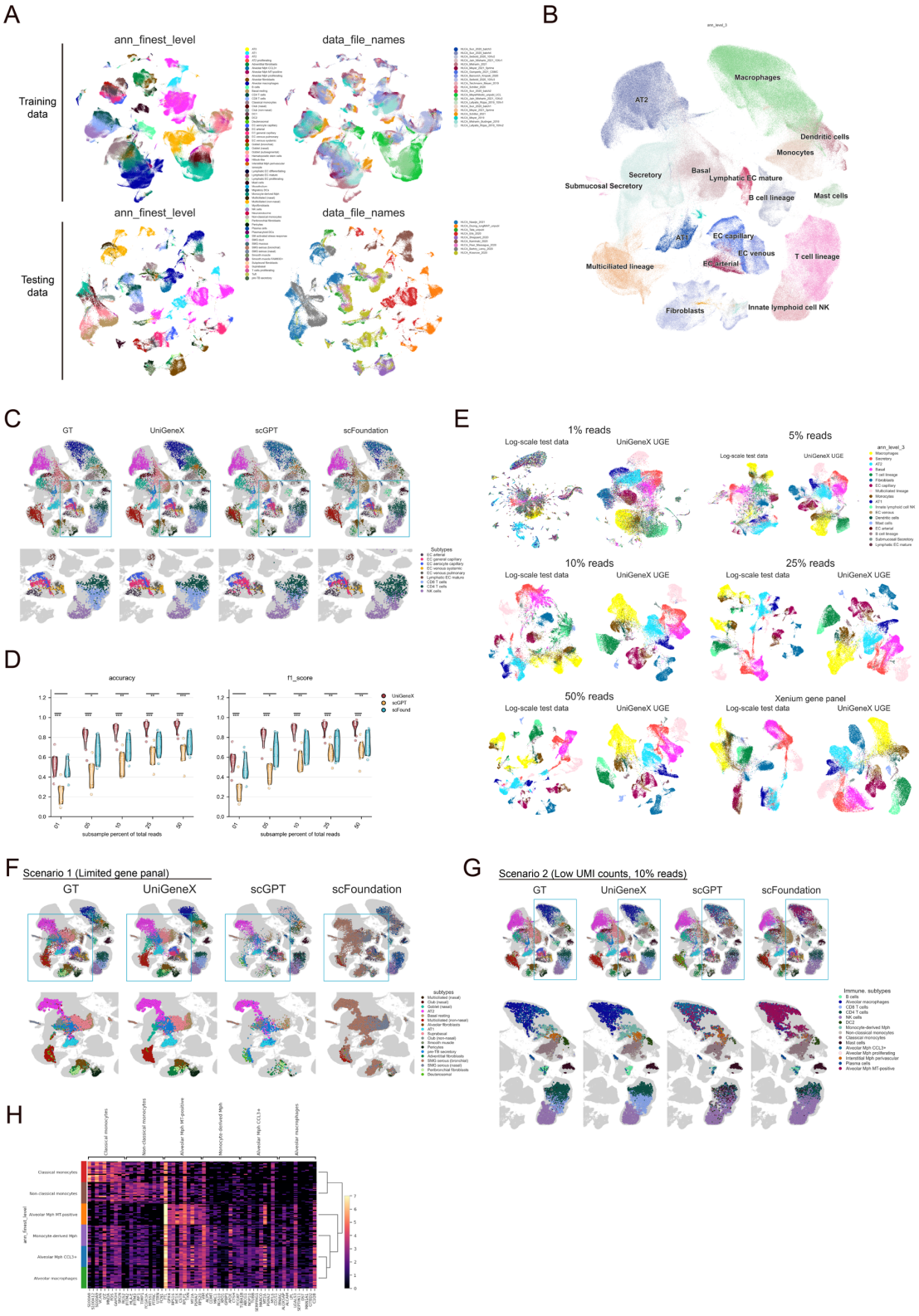


S-Figure 1. Validation of UniGeneX on HLCA lung atlas benchmarks, related to Figure 2.

(A) UMAP overview of standardly preprocessed (normalized and log-transformed) HLCA training and testing datasets colored by fine cell-type annotation and source dataset.

(B) Coarse lineage annotation of the UGE lung atlas UMAP.

(C) Comparison of ground-truth cell-type labels and predicted labels from UniGeneX, scGPT (32-shot), and scFoundation (32-shot). Zoomed panels highlight endothelial and lymphoid lineages. Gray points show the UGE atlas background, and colored foreground points show held-out test cells labeled by cell type.

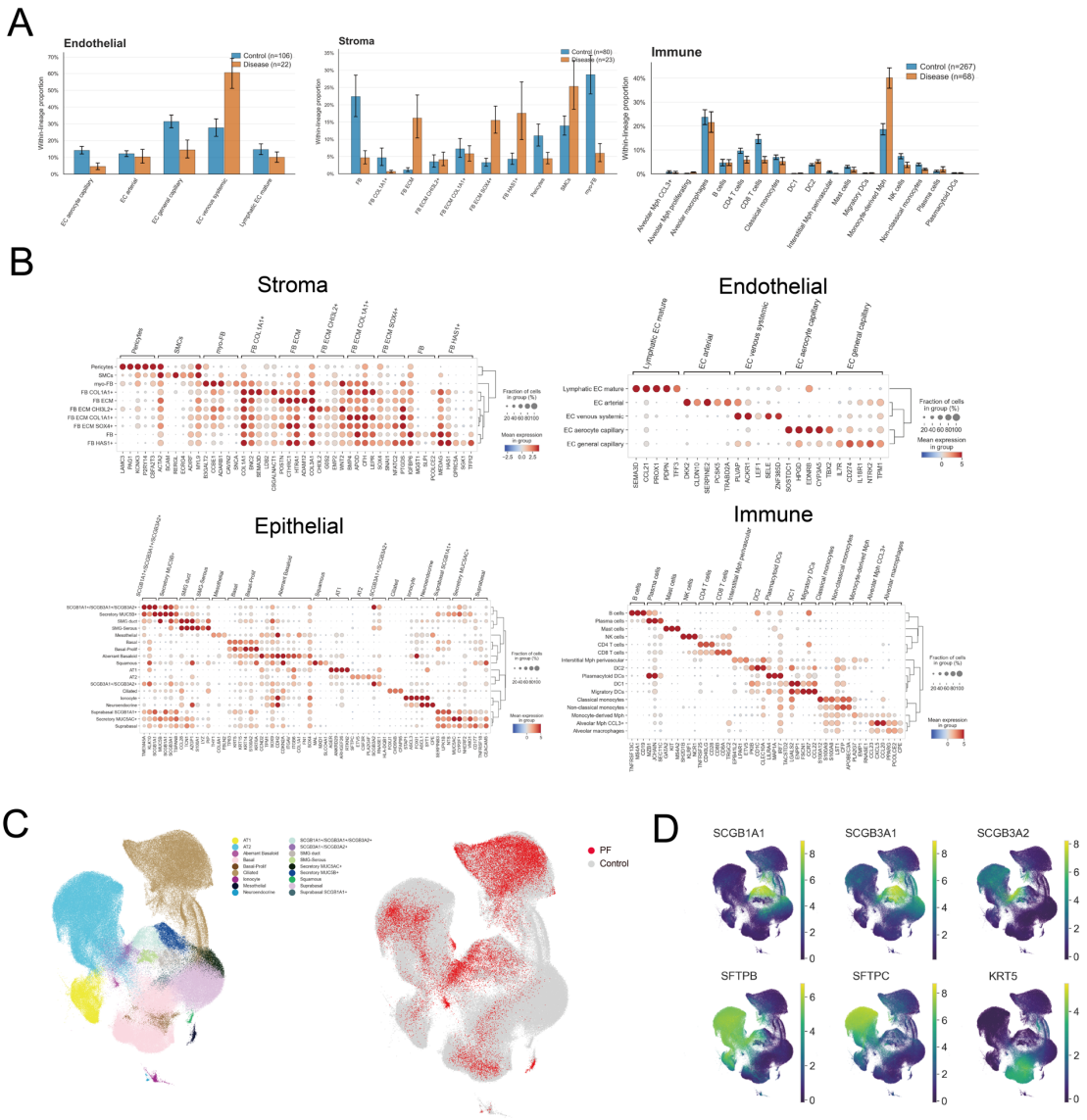
(D) Annotation accuracy and weighted F1 across read-subsampling levels. Each point summarizes a dataset-level benchmark result for UniGeneX, scGPT (32-shot), or scFoundation (32-shot).

(E) UMAP visualization of raw log-scale test data and UniGeneX UGE across read-subsampling levels (1%, 5%, 10%, 25%, and 50%) and the Xenium limited-gene panel. Colors indicate coarse cell labels.

(F) Scenario 1, limited gene panel: comparison of ground-truth labels and predicted labels from UniGeneX, scGPT (32-shot), and scFoundation (32-shot). Zoomed panels highlight epithelial subtypes. Gray points show the UGE atlas background, and colored foreground points show held-out test cells labeled by cell type.

(G) Scenario 2, 10% reads: comparison of ground-truth labels and predicted labels from UniGeneX, scGPT (32-shot), and scFoundation (32-shot). Zoomed panels highlight immune subtypes. Gray points show the UGE atlas background, and colored foreground points show held-out test cells labeled by cell type.

(H) Differentially expressed (DE) gene heatmap for myeloid-lineage subtypes. Rows represent cells, columns represent DE genes. Row colors indicate annotated subtypes, and heatmap color intensity denotes scaled expression.



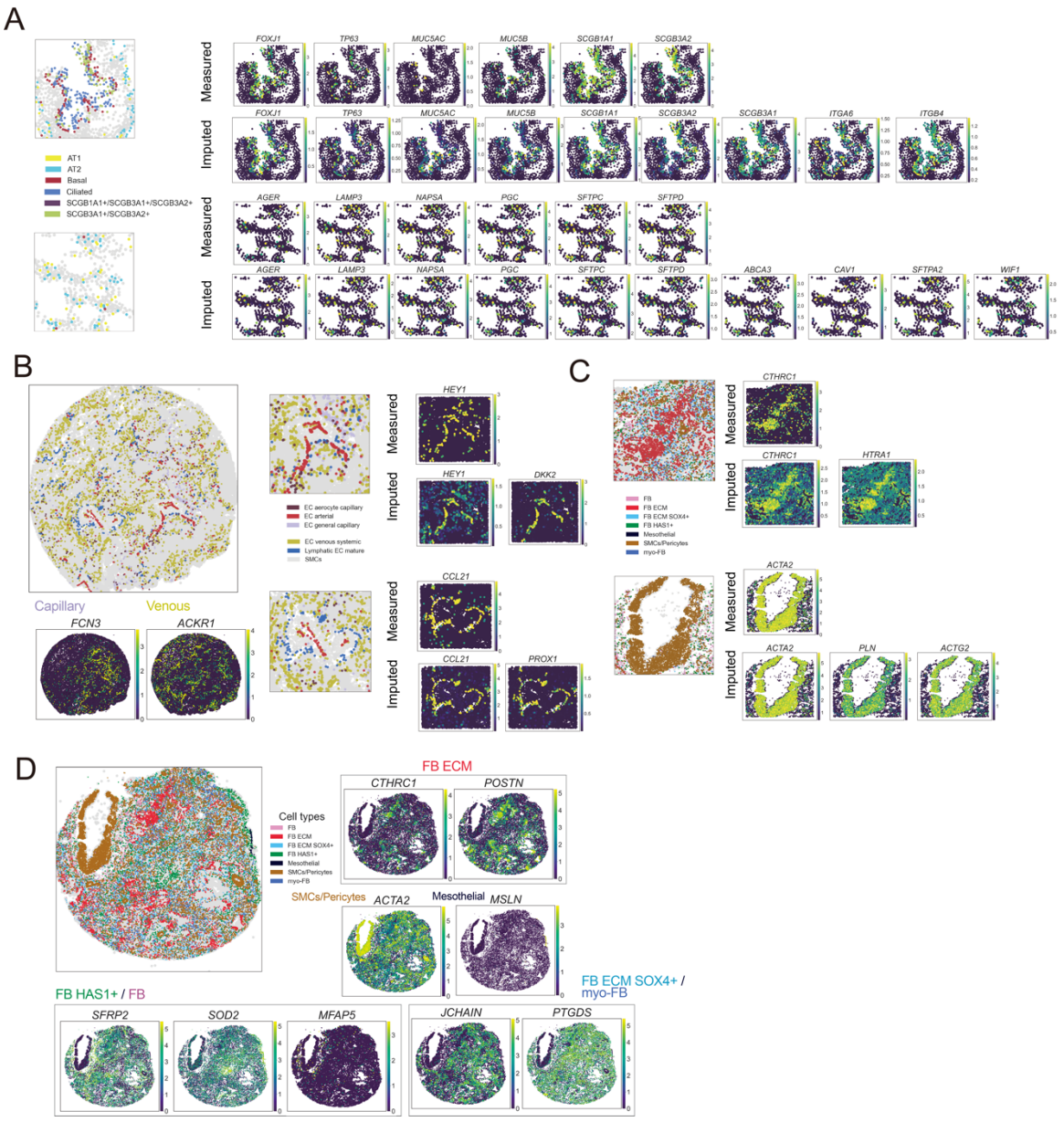
S Figure 2. UGE atlas building, related to Figure 3.

(A) Cell type proportion in endothelial, stroma, immune lineage in PF versus control lungs. Error bars represent the 95% confidence interval of proportion evaluated on n = 128 (endothelial), n = 103 (stroma), and n = 335 (immune) samples.

(B) Relative expression of stroma, endothelial, epithelial, and immune cell type markers is presented as Z-scores within each cell type compartment.

(C) UMAP of epithelial subtypes of UGE atlas. Color reflects difference cell types in left and disease status at right. AT: alveolar type; SMG: submucosal gland.

(D) Expression of cell type markers to identifying subpopulations of RAS cells.



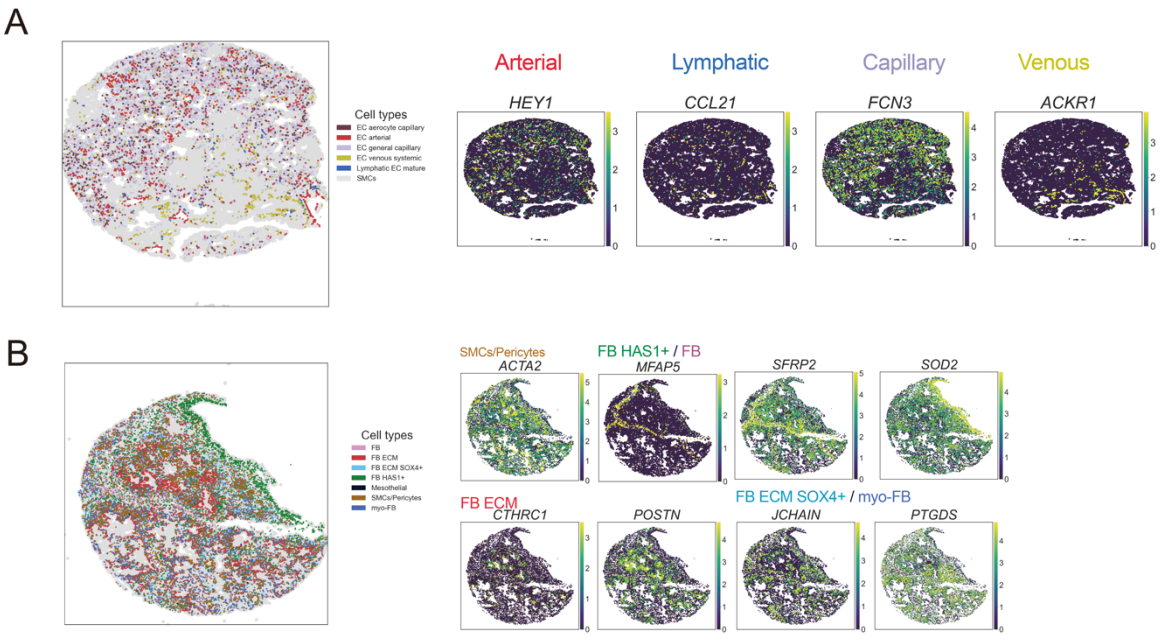
S Figure 3. UniGeneX for ST data cell type annotation, related to Figure 3.

(A) Epithelial cell types (left) and epithelial marker gene expression of measured or UGE pseudo cells (right) in selected regions (VUHD116A; unaffected).

(B) Left, marker gene expression for endothelial subtypes. Right, endothelial marker gene expression of measured or UGE pseudo cells in selected regions (VUILD107MA; IPF).

(C) Mesenchymal marker gene expression of measured or UGE pseudo cells in selected regions (VUILD106MA; IPF).

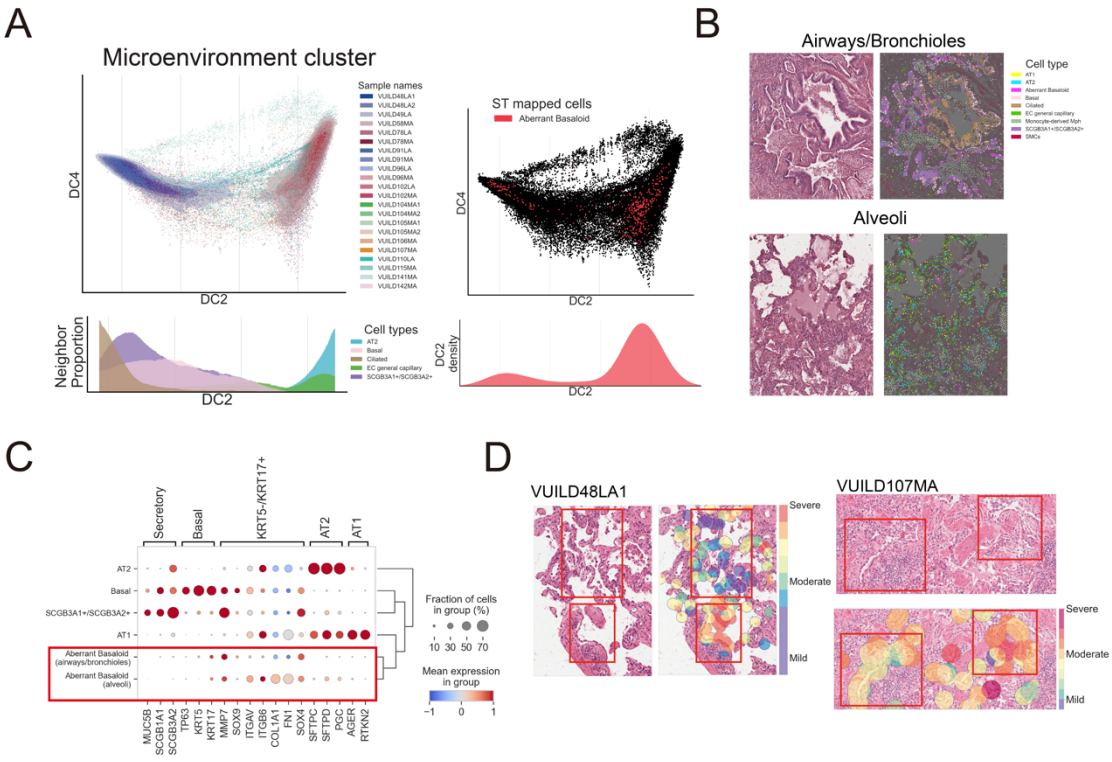
(D) Marker gene expression of fibroblast subtypes (VUILD106MA; IPF).



S Figure 4. UniGeneX for ST data cell type annotation, related to Figure 3.

(A) Marker gene expression of endothelial subtypes (VUILD96LA; sarcoidosis).

(B) Marker gene expression of mesenchymal subtypes (VUILD115MA; cHP).



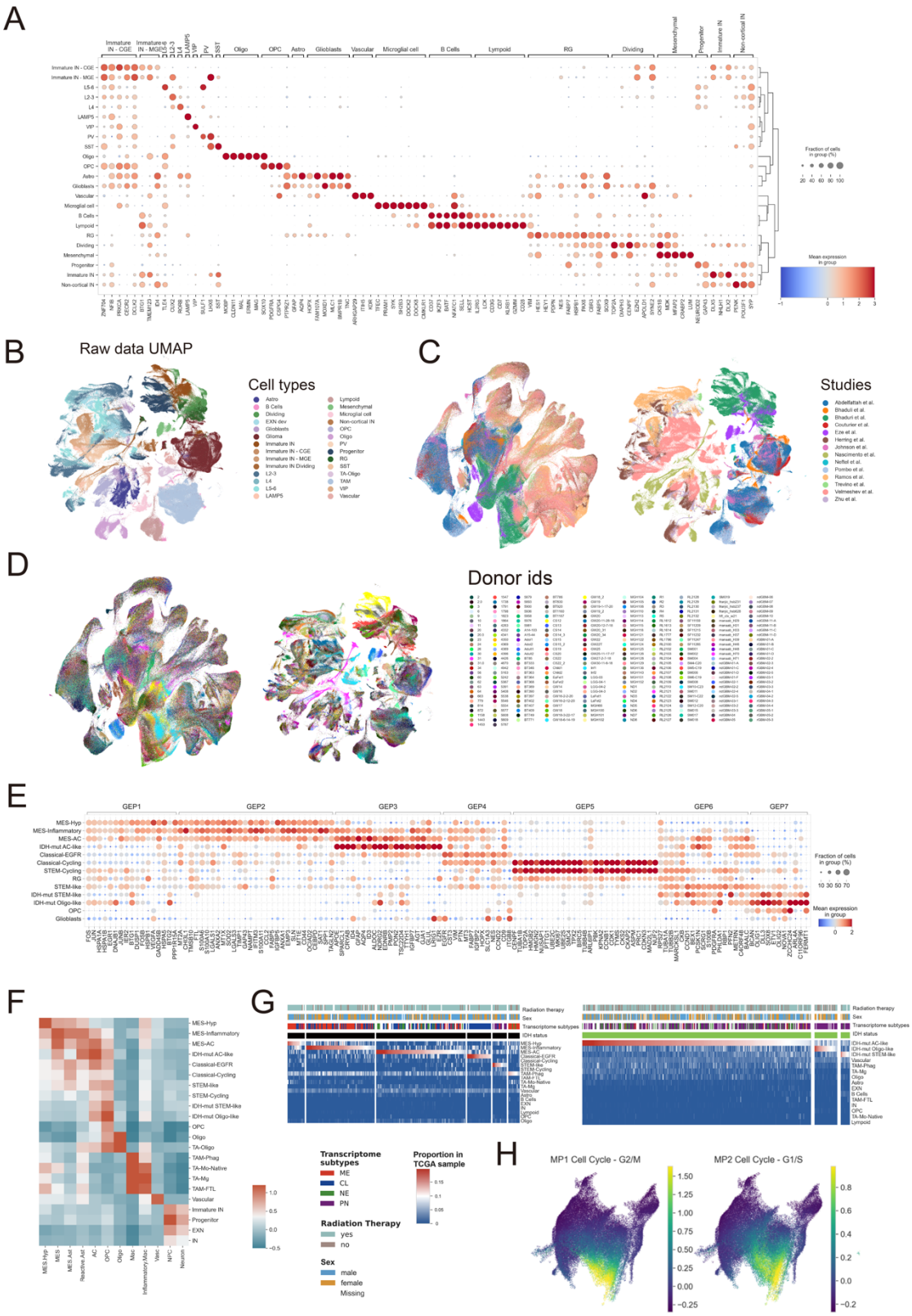
S Figure 5. Spatial pseudo time derived from the integrative analysis of the UGE atlas and ST data, related to Figure 4.

(A) Left top, Top left: Diffusion map (DC2 vs. DC3) of neighboring cell-type composition surrounding AT1, AT2, aberrant basaloid, SCGB3A1+/ SCGB3A2+ and SCGB1A1+/SCGB3A1+/ SCGB3A2+ cells in the ST data. Left bottom, Bar plot of neighboring cell type proportions for different cell types along DCs 2. Right top, Red dots represent the aberrant basaloid from the ST data that mapped to the core of aberrant basaloid in the UGE atlas. Right bottom, Kernel density estimation (KDE) plot illustrating the distribution of DCs 2 for red dots in top.

(B) Examples of airways/bronchioles or alveolar regions in H&E images (left) and H&E images overlaid with selected cell types.

(C) Marker gene expression of Aberrant Basaloid in airways/ bronchioles (red dots in (B) with DCs 2 less than 0) and aberrant basaloid cells in alveoli (red dots in (B) with DCs 2 greater than 0) are highlighted.

(D) More examples of contour map of spatial pseudo time for mapped cells in (D) in selected region of VUILD48LA (NSIP diagnosis) and VUILD107MA (IPF diagnosis).



S Figure 6. Constucting an UGE atlas of glioma and human brain development, related to Figure 5.

(A) Dotplot heatmap showing select genes used to annotate cell types in the dataset.

(B) UMAP visualization of standardly preprocessed (normalized and log-transformed) single-cell data prior to model training, color-coded by cell types.

(C) Comparative UMAP visualization of UGE atlas (left) and untrained raw data, colored by studies.

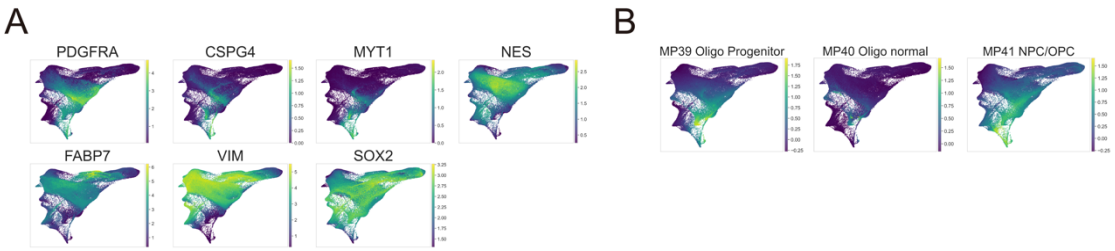
(D) Comparative UMAP visualization of UGE atlas (left) and untrained raw data, colored by donor ids.

(E) Dot plot illustrating the expression profiles of the identified gene expression programs used to define distinct tumor cellular states.

(F) Heatmap matrix displaying the gene module scores of UGE-defined tumor cellular states across marker signatures from published literature.

(G) Heatmap displaying the computational deconvolution of bulk RNA-seq samples from the TCGA cohort using the UGE atlas as a reference. Color intensity directly represents the estimated proportion of each defined tumor cellular state within individual bulk samples, aligned with clinical and molecular metadata including transcriptional subtypes, radiation therapy history, and patient sex.

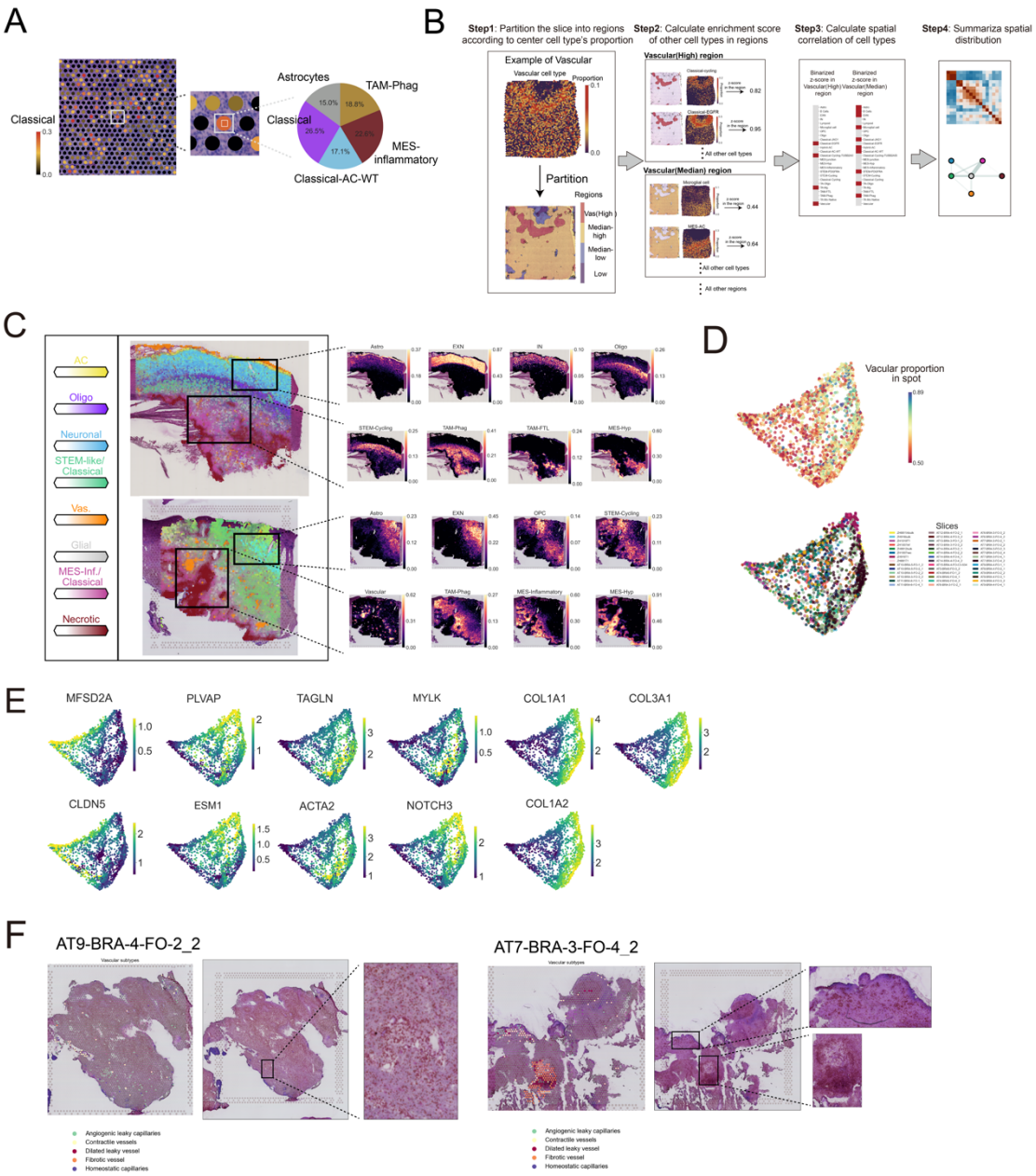
(H) UMAP embeddings indicating gene scores of distinct cell-cycle programs.



S Figure 7. Lineage markers and meta-programs characterizing neurodevelopmental programs in tumor cells, related to Figure 5.

(A) UMAP embeddings showing the expression of canonical RG cells marker genes.

(B) UMAP embeddings showing the gene module scores for distinct transcriptional meta-programs.



S Figure 8. Spatial distribution of GBM and mapping of vascualar zonations, related to Figure 6&7.

(A) Schematic of the deconvolution results illustrating the cell-type composition within individual spots.

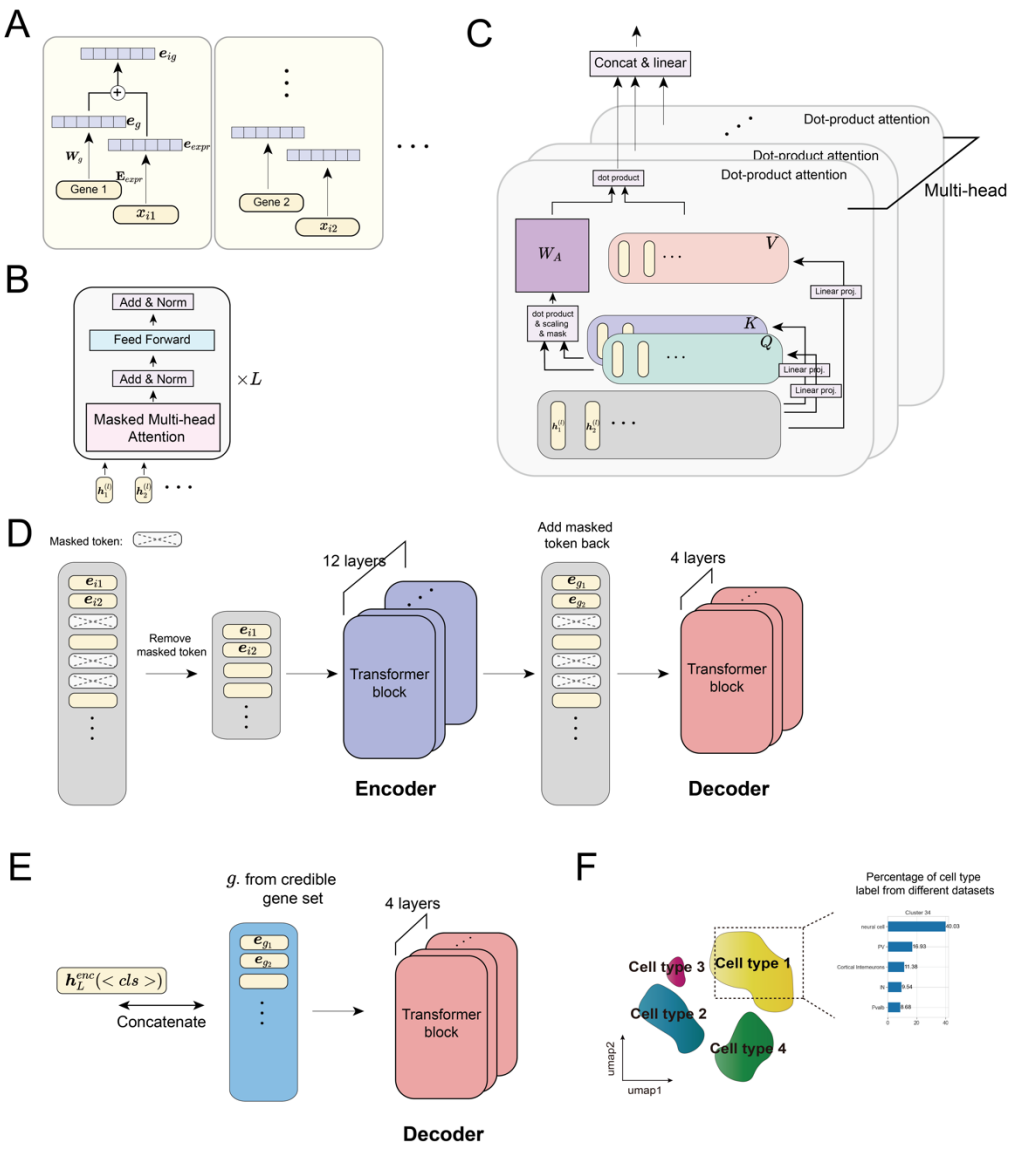
(B) Schematic workflow detailing the computational pipeline used to identify cell-type colocalization patterns and their spatial distributions.

(C) Cell-type abundances across Visium ST sections showing key distribution patterns.

(D) UMAP embeddings of vessel-dense spatial spots based on inferred vascular transcriptomic profiles, colored by vascular proportion within each spot (top) and individual patient tissue slices (bottom).

(E) UMAP embeddings displaying continuous expression profiles of canonical vascular marker genes, with color intensity reflecting relative expression levels.

(F) Spatial histology annotations of distinct vascular subtypes mapped onto representative tumor tissue sections.



S Figure 9. Schematic of the Transformer-based network architecture, as related to Methods.

(A) Gene and expression embedding input module.

(B) Transformer block architecture

(C) Multi-head scaled dot-product attention mechanism.

(D) Masked autoencoder architecture.

(E) Decoding module.

(F) Schematic representation of the majority voting framework used to harmonize heterogeneous annotations across diverse reference datasets.