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.

# Generative atlasing in universal gene expression space defines cell types and microenvironment spectra during disease progression

**Angela Wu**

angelawu@ust.hk

The Hong Kong University of Science and Technology https://orcid.org/0000-0002-3531-4830

**Xiaomeng Wan**

The Hong Kong University of Science and Technology

**Lei Yu**

The Hong Kong University of Science and Technology

**Yuheng Chen**

The Hong Kong University of Science and Technology

**Jiashun Xiao**

the Hong Kong University of Science and Technology

**Mu He**

The University of Hong Kong

**Can Yang**

The Hong Kong University of Science and Technology https://orcid.org/0000-0002-4407-3055

**Article**

**Keywords:**

**Posted Date:** September 4th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10838361/v1

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

**Additional Declarations:** There is **NO** Competing Interest.

# Generative atlasing in universal gene expression space defines cell types and microenvironment spectra during disease progression

*Xiaomeng Wan* [1,8], *Lei Yu* [2,8], *Yuheng Chen* [1], *Jiashun Xiao* [3], *Mu He* [7,*], *Can Yang* [1,4,*], *and Angela Ruohao Wu* [2,5,6,*]

[1] Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong SAR, China;

[2] Division of Life Science, The Hong Kong University of Science andTechnology, Clear Water Bay, Kowloon, Hong Kong S.A.R., China;

[3] School of Public Health, Sun Yat-sen University, Guangzhou 518107, China.

[4] Guangdong-Hong Kong-Macao Joint Laboratory for Data-Driven Fluid Mechanics and Engineering Applications, The Hong Kong University of Science and Technology, Hong Kong SAR, China;

[5] Department of Chemical and Biological Engineering, The Hong KongUniversity of Science and Technology, Hong Kong SAR, China;

[6] Center for Aging Science, The Hong Kong University of Science and Technology, Hong Kong SAR, China;

[7] School of Biomedical Sciences, Faculty of Medicine, The University of Hong Kong, Hong Kong

[8] These authors contributed equally to this work.

*Correspondence: muhe@hku.hk (M. H.), macyang@ust.hk (C.Y.), angelawu@ust.hk (A.W.)

## Abstract

Cells execute core physiological functions through intricate molecular networks. Elucidating these network principles is critical for understanding disease progression. However, most current transformer-based single-cell foundation models rely on encoder-only architectures that output abstract latent embeddings lacking explicit biological interpretability. An effective generative framework must quantitatively map these latent coordinates back into the biological real space of gene expression. To achieve this, we present UniGeneX, a generative single-cell foundation model that reconstructs large-scale, multi-dataset transcriptomics data into a quantitative space of Universal Gene Expression (UGE). By constructing the UGE space, UniGeneX effectively disentangles the molecular spectra of cell states across health and disease. This generative framework seamlessly integrates single-cell cohorts and spatial transcriptomics (ST) data, facilitating cell-type identification and niche architecture mapping. Applying UniGeneX to pulmonary fibrosis (PF) and glioma cohorts, we demonstrate its utility in constructing comprehensive tissue atlases, tracking disease progression along a spatial niche pseudotime, and identifying localized vascular zonation programs associated with tumor cellular state transitions.

# Introduction

The cell achieves basic function through molecular interactions that result in different phenotypes under diverse physiological contexts. Deciphering this cellular physiology requires extracting the fundamental rules of operation for cells, or a universal representation of the molecular interactions in cells, from large scale data.

Single-cell RNA sequencing (scRNA-seq) provides high-resolution molecular profiles of cellular identity and states, encoded in complex gene–gene interactions. Importantly, it captures more than the activity of an isolated cell; rather, the overall transcriptomic profile of one cell also reflects its biologically meaningful interactions with other cells, such as those within niche-specific microenvironment or disease-associated programs[1–3]. We conceptualize these latent biological signals within cells as gene programs, which may encode previously non-explicit regulatory factors or contextual influences. Destabilization of these gene programs can drive progressive diseases, such as Pulmonary Fibrosis (PF) or cancer, and these programs undergo continuous changes as the pathology advances[4–6]. Therefore, deciphering the gene program trajectory is essential for understanding disease development and finding potential therapeutic targets. Efforts to understand such underlying cellular programs include recently proposed virtual cell models[7], where one of the objectives is to learn a universal representation of the cell. By leveraging curated, context-specific scRNA-seq datasets at scale, an appropriate model can define a universal gene expression space that transcends individual datasets, while remaining relevant in a specific tissue context. Furthermore, a useful model should produce biologically interpretable and meaningful output to reveal disease-related cell-state transitions and resolve spatial niches.

In theory, the transformer attention mechanism[8] is well-suited to model pairwise gene–gene interactions and capture the structure and dynamics of the underlying gene regulatory network. Multi-head attention further enhances this capacity by enabling each attention head to focus on distinct biological relationships. For example, one head may learn associations related to the cell cycle, while another may capture programs involved in apoptosis, metabolism, or DNA repair, effectively disentangling diverse regulatory programs within the same model[9–11]. Yet existing transformer-based approaches rely heavily on encoder-only architectures[12,13], which output abstract latent embeddings rather than reconstructing physical, quantitative gene expression profiles. These latent embeddings can neither be readily interpreted in the form of gene programs, nor used for downstream mechanistic modeling. Therefore, an effective generative framework must quantitatively map these latent coordinates back into the interpretable biological real-space of gene expression.

Here, we introduce a generative single-cell foundation model, UniGeneX, trained on large-scale transcriptomic data to learn the fundamental rules of cellular transcription (Fig. 1A). Unlike existing models that focus on learning cell embeddings for subsequent fine-tuning to meet task-specific requirements, UniGeneX reconstructs inputs into a universal gene expression (UGE) space (Fig. 1C). Our model design provides real-space reconstruction of data that serves as a robust foundation for integrating diverse cell states across sample cohorts and multimodal data types. UniGeneX achieves this through four key capabilities. First, it constructs a comprehensive UGE atlas from extensive disease- or tissue-matched scRNA-seq training data,

enabling the identification of both conserved and context-specific gene programs across diverse healthy and pathological cell states, providing a high resolution for cellular heterogeneity (Fig. 1B). Second, the UGE atlas harmonizes cell-type annotations across datasets via a majority vote strategy, which leverages abundant labels in the large training datasets. Third, the UGE atlas serves as a comprehensive reference to deconvolute bulk RNA-seq and low-resolution spatial transcriptomics (ST) data (Fig. 1B). Finally, by scaling scRNA-seq and ST data to a common expression space, UniGeneX generates high-resolution pseudocells from ST data based on learned transcriptional patterns. This in turn facilitates the integration of the UGE atlas with ST data to identify disease-relevant cell states and pinpoint cell niches specific to disease progression in a spatial context (Fig. 1B). To demonstrate its utility, we applied UniGeneX to model PF and glioma. Our analyses reveal that UniGeneX effectively characterizes disease-relevant cell states and resolves complex spatial microenvironments, providing a powerful, interpretable tool for elucidating the mechanisms of progressive diseases.

# UniGeneX overview

UniGeneX is a single-cell foundation model designed to construct a comprehensive UGE atlas to uncover the underlying cell state transitions and associated microenvironments in human diseases. The framework consists of two main stages: training and inference; and three novel features: context-specific, interpretable and actionable.

In the training stage, UniGeneX is trained in an unsupervised manner on millions of single cells and thousands of genes, comprising three critical steps: gene selection, transformer-based model training, and cell-type harmonization. Initially, UniGeneX performs gene selection to ensure high-quality input data. Instead of using common highly variable genes (HVGs) across all datasets, it identifies HVGs specific to each dataset. The specific model design of UniGeneX allows it to train only on these dataset specific HVG genes and integrate them to generate a universal gene expression atlas.

Next, UniGeneX employs a novel transformer-based probabilistic autoencoder architecture to construct the UGE atlas. Unlike methods that solely utilize a transformer encoder or a standard encoder-decoder architecture, UniGeneX incorporates a KL regularization term to create a smoother, more continuous latent space, effectively clustering similar cell states together. This strategy confers generative ability and enable the data share information from each other, which allows the model output gene expression profiles while eliminating batch effect.

After training, UniGeneX build an UGE atlas and annotates cell types for cells by harmonizing cell type labels across diverse datasets. The comprehensiveness of UGE atlas provide notable advantages for cell-type harmonization: it does not require every cell to have accurate annotations, as cell identities can be inferred from diverse dataset annotations, and it allows for more detailed labels by referencing the most refined annotations from other datasets.

In the inference stage, the UGE atlas offers notable interpretability and actionability. It identifies shared gene programs across samples by following standard processing strategies of scRNA-seq data, uncovering the spectra of gene programs during cell state transitions. The UGE atlas facilitates label transfer in a zero-shot manner for query datasets; specifically, it reconstructs

the UGE of the query data and assigns cell-type labels from the atlas. UniGeneX's generative capability reconstructs UGE for data with varying numbers of detected genes, making it applicable to highly sparse datasets.

Additionally, UGE is highly actionable, serving as a comprehensive single-cell reference dataset that integrates multiple data type and performs diverse downstream analyses. For example, it enables the deconvolution of both bulk RNA-seq and ST data across various resolutions. By integrating a highly interpretable UGE atlas with ST data, UniGeneX provides a powerful framework for deciphering cell-state transitions during disease progression and their associated microenvironmental changes.

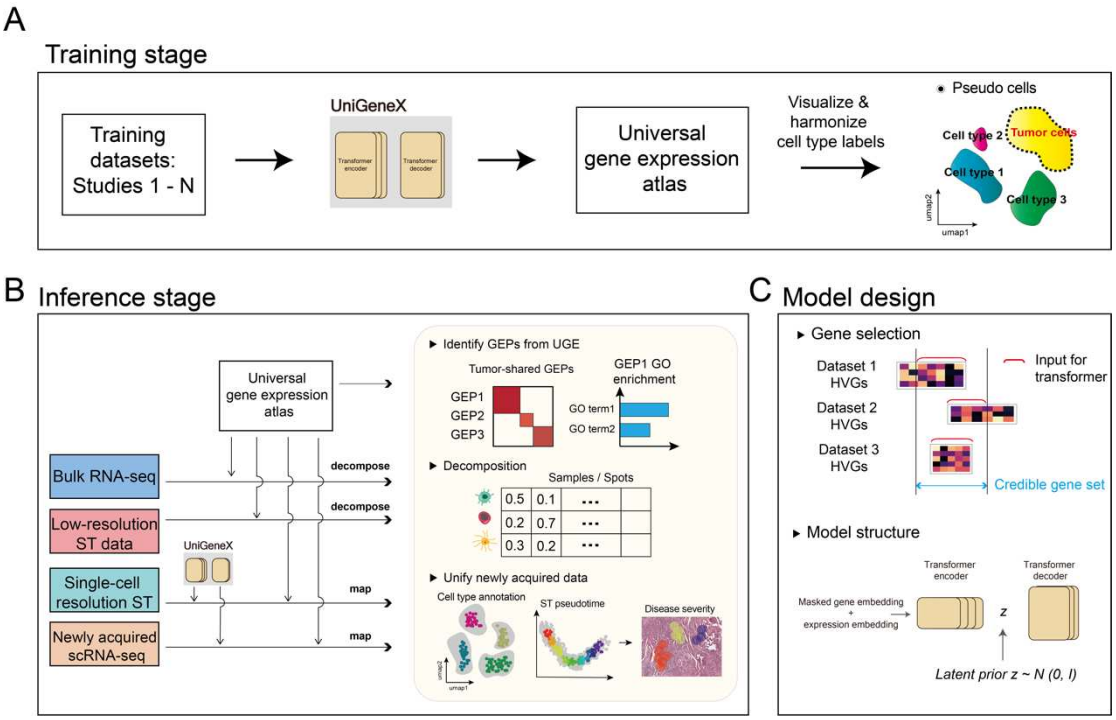



**Figure 1. Overview of UniGeneX.**

(A) In training stage, UniGeneX takes as input a comprehensive collection of single-cell transcriptomic datasets from a specific tissue type. Through the UniGeneX core architecture, the network learns from these datasets to construct a unified universal gene expression atlas.

(B) Leveraging the pretrained universal gene expression atlas, UniGeneX zero-shot decomposes or maps varied transcriptomic data types, including Bulk RNA-seq, low-resolution and single-cell resolution ST data, and newly acquired scRNA-seq. This stage enables the identification of shared gene expression programs (GEPs), spot/sample cell-type decomposition, and integrated scRNA-seq and ST pseudotime tracking—ultimately facilitating disease severity evaluation in tissue pathology.

(C) Highly variable genes (HVGs) across studies are consolidated into a credible gene set for input. Following masked gene expression embedding, a transformer encoder maps inputs into a normalized latent coordinate space, which is subsequently processed by a transformer decoder for expression reconstruction.

## Building a universal gene expression atlas and projecting new transcriptomics data

To assess the accuracy and efficiency of cell-type identification via projecting onto the UGE atlas, we evaluated UniGeneX using the non-disease Human Lung Cell Atlas (HLCA)[14], which provides consensus annotations across diverse lung studies. We accessed 32 datasets from 243 donors and 414 samples, comprising approximately 0.8 million cells. The atlas was split into 21 training datasets and 11 held-out testing datasets, with the held-out datasets treated as newly acquired query data requiring annotation (Fig. 2A, Fig. S1A). Before UGE generation, direct UMAP visualization of conventionally processed expression data (normalized and log-transformed counts) showed fragmented structures and strong dataset-associated variation, making joint clustering, subtype comparison, and cross-study integrative analysis difficult (Fig. S1A). In contrast, the UGE atlas resolved the major lung compartments and fine cell states, and mapped query cells were projected into the same manifold as the training atlas, enabling zero-shot label transfer (Fig. 2B). UniGeneX accurately recovered consensus cell-type labels in the held-out datasets. In addition, lineage-specific confusion matrices showed consistent performance across immune, epithelial, stromal, and endothelial lineages (Fig. 2C).

We further compared UniGeneX with two existing single-cell foundation models, scGPT[12] and scFoundation[13]. Both methods use pretrained representations followed by classifier fine-tuning for cell-type annotation. We evaluated annotation at both coarse and fine cell-type resolutions (Fig. S1B). Notably, these fine-tuning strategies require separate classifiers for the two label resolutions, whereas UniGeneX uses a single zero-shot projection and atlas label-transfer workflow. For a fair comparison, scGPT and scFoundation were evaluated across few-shot settings. Their performance increased as more labeled examples were provided, but UniGeneX consistently achieved higher performance for both coarse and fine labels (Fig. 2D, Fig. S1C). Zero-shot UniGeneX achieved mean weighted F1 scores of 0.969 for coarse labels and 0.914 for fine labels across nine held-out datasets. Even at the 32-shot setting, scFoundation reached only 0.849 and 0.733 for coarse and fine labels, respectively, while scGPT reached 0.733 and 0.657, indicating that atlas-based UGE projection provides stronger annotation performance without additional classifier fine-tuning. UMAP visualization confirmed that UniGeneX predictions aligned more closely with the ground-truth subtype labels, whereas scGPT and scFoundation showed lower accuracy and less clearly separated CD4 T cells, CD8 T cells, and NK cells (Fig. S1C).

We next tested whether this atlas-mapping strategy remains effective when query datasets have limited gene coverage or sparse UMI counts, two common limitations in ST data. We considered two degraded-input scenarios. In the first, held-out cells were restricted to a Xenium lung tissue gene panel to mimic targeted image-based spatial transcriptomics. In the second, held-out counts were subsampled to 1%, 5%, 10%, 25%, and 50% of the original reads to mimic shallow sequencing (Methods). In both scenarios, UniGeneX generated high quality UGE profiles from the degraded inputs and restored marker-gene patterns that were weakened or absent in the measured data (Fig. 2E). UMAP visualization and expression heatmaps showed that the UGE representation recovered information learned from the training atlas (Fig. 2E, Fig. S1E). For example, in the limited-gene-panel setting, UGE profiles recovered the club-cell marker *SCGB1A1*, *SCGB3A1* and the adventitial fibroblast marker *CFD*, which were not directly captured by the restricted test input. In the 10% read-depth

setting, UGE preserved the same marker blocks from shallow inputs and enhanced signals for markers such as the adventitial fibroblast gene *PCOLCE2* and the mural-cell marker *RGS5* (Fig. 2E).

Quantitative benchmarking confirmed that these recovered representations improved cell-type annotation under degraded inputs. In the Xenium-like panel scenario, UniGeneX achieved a median dataset-level weighted F1 of approximately 0.80, compared with approximately 0.33 for scGPT and 0.19 for scFoundation (Fig. 2F). In the 10% read-depth scenario, UniGeneX achieved a median weighted F1 of approximately 0.91, compared with approximately 0.54 for scGPT and 0.68 for scFoundation (Fig. 2F). Overall improvements were significant in all levels of subsample rate (Fig. S1D), and lineage-level weighted F1 score showed clear gains across epithelial and immune compartments (Fig. 2F, Fig. S1F-G). We also observed that, in the subsampling scenario, scGPT and scFoundation often misclassified Monocyte-derived Mph as Alveolar Mph MT-positive, a cell type present only in the training data. Inspection of immune marker heatmap suggested a likely reason: some genes differentially expressed in Alveolar Mph MT-positive cells are also highly expressed in other myeloid lineages (Fig. S1H). Classifier fine-tuning may therefore be biased by a small number of highly expressed marker genes, whereas UniGeneX projection leverages broader transcriptomic variation in the UGE atlas. This global projection strategy likely contributes to the stronger robustness of UniGene under sparse-input conditions.

Together, these results show that UniGeneX provides more than a cell embedding for annotation. By reconstructing a UGE profile in a shared atlas space, it enables zero-shot projection of new datasets, transfer of fine cell-type labels, visualization in a common manifold, and recovery of biologically interpretable marker programs even from low-quality measurements. This makes UniGeneX suitable for integrating ST datasets with a large reference atlas, including datasets with limited gene panels or reduced sequencing depth.

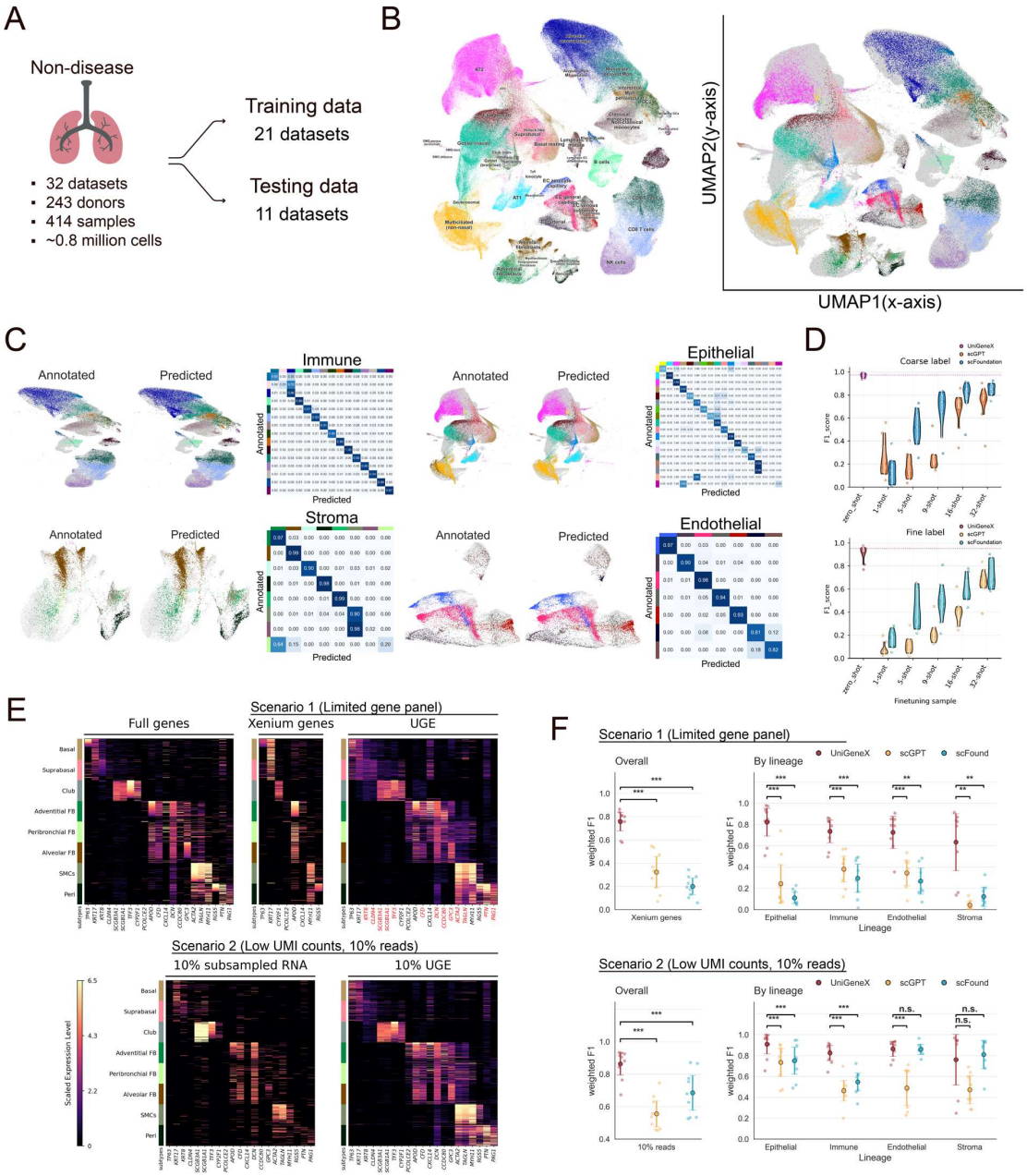



**Figure 2. UniGeneX validation and benchmarking in the HLCA lung atlas.**

(A) Overview of the non-disease HLCA benchmark split.

(B) UMAP visualization of the lung UGE atlas and projection of UGE test data cells.

(C) Cell-type annotation performance of UniGeneX on held-out cells. Annotated and predicted UMAPs and confusion matrices are shown for immune, epithelial, stromal, and endothelial lineages.

(D) Comparison of cell-type annotation performance across different methods. Weighted F1 scores were computed for coarse and fine labels across held-out HLCA test datasets. Points represent dataset-level scores, and distributions summarize performance across dataset. Vertical lines show mean and 95% confidence interval.

(E) Heatmap of scaled marker-gene expression. Heatmap rows are aligned cells, columns are marker genes, side colors indicate subtypes, and color intensity denotes scaled

expression. Genes absent from the Xenium gene panel are highlighted in red within the UGE panel.

(F) Annotation performance under degraded-input scenarios. Weighted F1 scores are shown overall and by lineage for UniGeneX, scGPT, and scFoundation using either the Xenium gene panel or 10% subsampled reads. Points represent held-out datasets; summary points and vertical lines show mean and 95% confidence interval. Significance was assessed by paired t-test across datasets, with $P < 0.05$, *$P < 0.01$, **$P < 0.001$, and n.s., not significant.

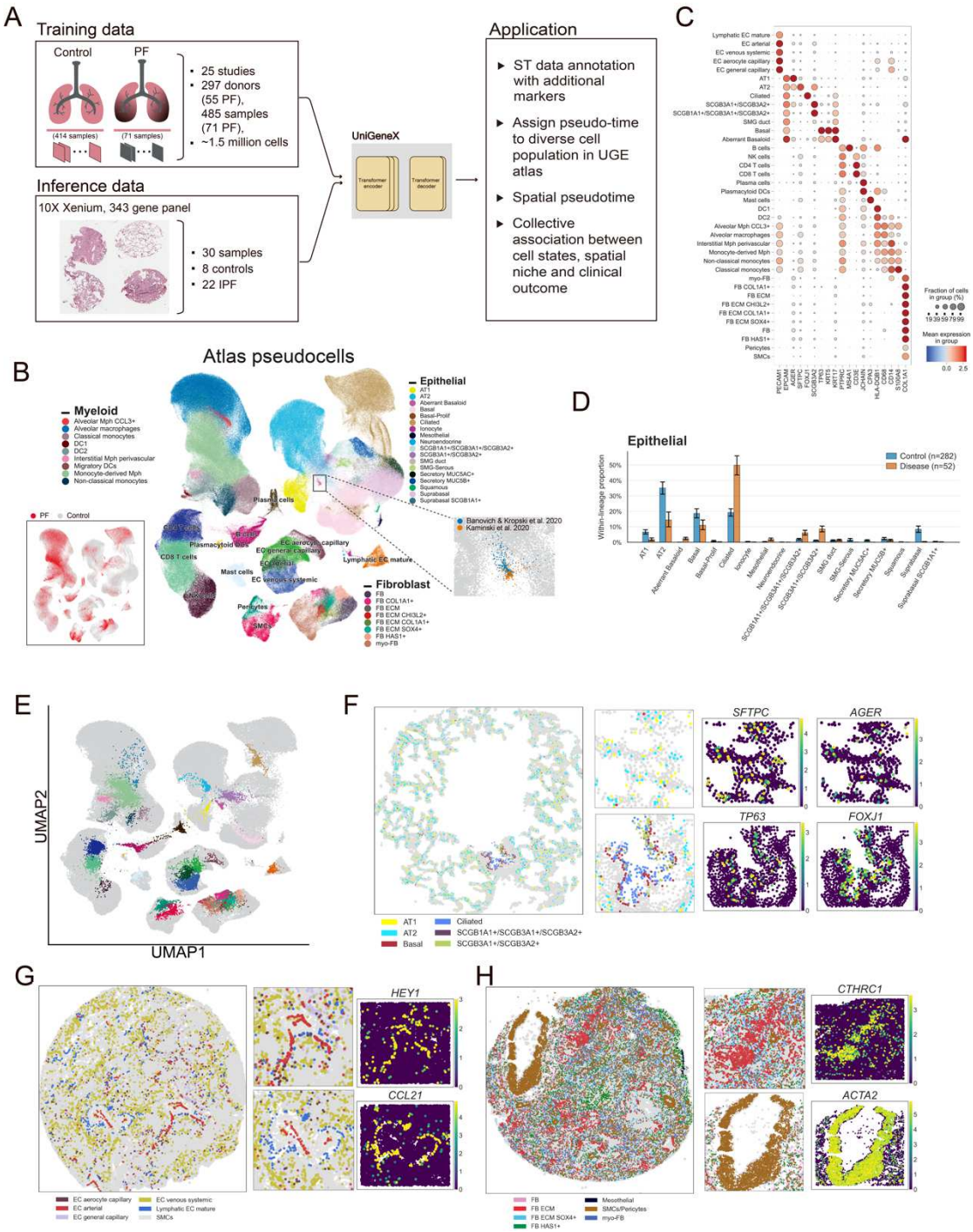
## UGE atlas of human lung in health and disease

Pulmonary fibrosis (PF) is a chronic and progressive lung disease characterized by the accumulation of scar tissue (fibrosis) in the lungs, leading to a decline in lung function and impaired gas exchange[15]. A central unresolved question is how molecular and cellular dynamics within spatially defined niches drive disease progression and impair alveolar regeneration[16–20]. To address this, we used UniGeneX to integrate large-scale scRNA-seq and spatial transcriptomics data across health and disease states, and to perform comprehensive trajectory analysis in different niches (Fig. 3A).

Using >1.5 million cells from 485 samples, UniGeneX generated a UGE atlas spanning healthy and PF lungs (Fig. 1B). Cell types were annotated using canonical markers and cross-referenced with prior studies (Fig. 3B, C, Fig. S2B). As expected, both cell type composition and transcriptional states are profoundly altered in PF samples (Fig. 3D, Fig. S2A, C). Notably, we rediscovered the aberrant basaloid epithelial population previously linked to PF severity[16,17], which is predominantly present in disease samples (Fig. 3B, D, Fig. S2C). The functional canonical alveolar type 1 (AT1) and type 2 (AT2) cells are markedly depleted, indicating impaired alveoli regeneration (Fig. 3D). Concurrently, airway lineages, including ciliated cells and respiratory airway secretory (RAS) cells, expand significantly. Within RAS cells from both healthy and diseased tissues, we identified two subpopulations (SCGB1A1+/SCGB3A1+/SCGB3A2+ and SCGB3A1+/SCGB3A2+) that expand significantly in disease states (Fig. 3D, Fig. S2C-D). These shifts align with prior findings, validating the biological fidelity of the UGE atlas[17].

Alveolar regeneration is classically driven by AT2-to-AT1 differentiation. However, emerging evidence indicates that human RAS cells in respiratory bronchioles and bronchioalveolar stem cells (BASCs) in mice also contribute to alveolar repair[21–23], highlighting diverse regenerative trajectories during alveolar epithelium repair. In pathological PF tissues, the aberrant basaloid cells are thought to represent a stalled differentiation state towards AT1 cells, rather than part of a normal regenerative trajectory[16,17]. To decipher the mechanisms underlying regenerative failure, high-resolution mapping of spatially constrained differentiation processes within distinct tissue niches is demanded. UniGeneX addresses this by generating spatially informed pseudocells from ST data, thereby facilitating the integrative analysis of large-scale UGE atlas and ST data. We align these pseudocells with the scRNA-seq–derived UGE atlas, enabling cross-platform integration on a unified expression scale (Fig. 3E). This approach generates high-resolution spatial annotations and effectively expanding the original Xenium panel to infer additional markers.

Overall, UniGeneX accurately annotates distinct niches, with spatial assignments showing strong concordance with canonical marker expression (Fig. 3F-H). Moreover, UniGeneX enables us to evaluate markers not included in the original Xenium panel. For example, expression of AT1 markers such as *CAV1*[22] and AT2 markers such as *SFTPA2*[24] and *WIF1*[22], which were not directly measured by Xenium, can be robustly imputed, revealing spatial patterns consistent with their canonical biology (Fig. S3A). Similarly, basal cells localize to the inner surface of the airways, where the canonical marker *TP63* and the inferred marker *ITGA6*[25] exhibit corresponding enrichment (Fig. 3F, Fig. S3A). Key pathological niches, including vascular and fibrotic compartments, are precisely resolved (Fig. S3B-D, Fig. S4A-B). Together, these findings confirm the reliability of UniGeneX.




## Figure 3. Building a human lung PF UGE atlas using UniGeneX.

(A) Schematic workflow for analyzing human lung diseases using UniGeneX. UniGeneX is trained on millions of human lung scRNA-seq cells across health and disease. During the inference stage, UniGeneX expand ST data gene detection number by inferring based on UGE, facilitating comprehensive integrative analysis.

(B) Uniform Manifold Approximation and Projection (UMAP) of the UGE atlas, trained on millions of cells. Colors denote distinct cell types. Left panel: cells colored by disease status. Right panel: zoomed-in view of aberrant basaloid cells, colored by study of origin. Abbreviations: AT, alveolar type; FB, fibroblast; SMG, submucosal gland; DC, dendritic cell; Mph, macrophage; NK, natural killer cell; SMCs, smooth muscle cells; EC, endothelial cell.

(C) Dotplot showing expression of selected marker genes of different cell types.
(D) Barplot showing cell type proportion changes in epithelial lineage in PF versus control lungs. Error bars represent the 95% confidence interval of proportion evaluated on n = 334 samples.
(E) UMAP of aligned pseudocells from ST data and the UGE atlas. Gray dots represent cells from UGE atlas. Colored dots represent pseudocells from ST data, with color indicating the annotated cell type.
(F) Spatial annotation plot showing the spatial distribution of epithelial cell types (left) and log-normalized marker gene expression (right) in selected regions (VUHD116A; unaffected).
(G) Spatial annotation plot showing the spatial distribution of endothelial cell types (left) and log-normalized marker gene expression (right) in selected regions (VUILD107MA; IPF).
(H) Spatial annotation plot showing the spatial distribution of fibroblast cell types (left) and log-normalized marker gene expression (right) in selected regions (VUILD106MA; IPF).

## Mapping spatial pseudotime via the UGE atlas resolves disease progression

Studies have increasingly shown that both fibrosis and dysfunctional alveolar epithelium regeneration contribute synergistically to the progression of PF. Therefore, understanding impaired alveolar regeneration and its spatial relationship with fibrotic lesions is critical for elucidating disease progression. To this end, we mapped UGE from ST data onto a lung PF UGE atlas, aligning the spatiotemporal dynamics of diverse cell populations across a continuum of tissue states from healthy to severe disease. We first examined the regeneration of alveolar epithelium using diffusion map analysis (Fig. 4A). In healthy samples, AT2 cells exhibited trajectory connections to both AT1 cells and SCGB3A1+/SCGB3A2+ RAS cells, indicating their regenerative capacity. This aligns with recent studies by Murthy et al. and Basil et al. confirming the role of RAS cells in alveolar repair[17,22,26]. However, this regenerative trajectory is disrupted in pathology; diseased samples showed a marked reduction in AT1 cells and an increased AT2 to AT1 ratio, indicating a blockade in AT2 to AT1 differentiation axis (Fig. 4B). These observations are consistent with prior reports and suggest that RAS or AT2 cells fail to terminally differentiate into AT1 cells, instead adopting an aberrant basaloid fate in disease[16,17]. To determine the origin of aberrant basaloid cells, we first examined their spatial distribution and found that they were predominantly localized within alveolar niches enriched with AT2 and general capillary cells (Fig. S5A–B). A small fraction of these cells was detected in airway niches enriched with ciliated and basal cells; however, they consistently lacked expression of *COL1A1* and *FN1* and are therefore unlikely to represent the typical aberrant basaloid population associated with fibrosis (Fig. S5C). This supports the conclusion that pathogenic aberrant basaloid cells arise specifically within alveolar niches. Notably, by ranking SCGB3A1+/SCGB3A2+ RAS cells based on their proximity to AT2 cells, we observed co-expression of canonical AT2 markers and *SCGB3A2*, providing evidence for a regenerative axis from RAS cells to AT2 cells (Fig. 4C). This finding aligns with prior studies on RAS cell plasticity[23]. Collectively, based on these results, we suggest that, in pathological alveolar niches, both resident AT2 cells and RAS-derived AT2 cells fail to complete differentiation into AT1 cells and instead adopt an aberrant basaloid fate affected by the microenvironment changes.

This assumption suggests a potential regenerative axis linking RAS cells, AT2 cells, and aberrant basaloid cells. Integrating these observations, we propose that the trajectory from aberrant basaloid cells toward AT1 fate represents an abnormal continuum of transcriptional states in disease progression: transcriptional programs shifting toward AT1 identity reflect successful regeneration, whereas diversion toward the aberrant basaloid state signifies pathological remodeling. To visualize this trajectory, we leveraged the UGE atlas in conjunction with ST data to reconstruct a spatiotemporal map of disease progression using diffusion map (Fig. 4D). Importantly, after mapping the pseudotime data from the UGE atlas onto the ST cells and projecting it back into the spatial context, we surprisingly identify a region that exhibits very early pseudo time (Mild) even in a sample where most regions display severe fibrosis (Fig. 4E). Correlative H&E staining confirmed the presence of healthy alveolar architecture in early-stage regions (Fig. 4H), while mid-stage (moderate pseudotime) regions exhibited epithelial and stromal remodeling, and late-stage (severe pseudotime) regions displayed advanced fibrosis. We also confirmed these findings in other slices (Fig. S5D). Theses observation is consistent with the spatially resolved pseudotime trajectory (Fig. 4D-E, 4I). The strong concordance between histopathological severity and the spatial pseudotime assigned through UGE atlas and ST integration validates our method for recapitulating disease progression in situ (Fig. 4I).

To further characterize the microenvironment during disease progression, we investigated changes in the composition of neighboring cells along pseudotime. We observed an increase in fibroblast ECM-producing cells and monocyte-derived macrophages, while AT2 cells remained relatively stable initially before declining rapidly (Fig. 4F). Additionally, alveolar capillary endothelial cells exhibited a progressive decrease, indicating a loss of alveolar function (Fig. 4F).

Next, we examined how the cellular microenvironment evolves along the spatiotemporal trajectory of disease progression. Macrophages in the mild stage exhibit a normal immune response, while in the moderate and severe stages, they activate pro-inflammatory pathways (such as *PTGDS*, *IL1B*, and *SPP1*) (Fig. 4G). This transition may indicate a shift from a balanced immune function to a dysregulated, pro-inflammatory state. Fibroblasts in the mild stage express genes associated with tissue remodeling and repair, such as *ACTA2* and *CCN2*. In contrast, during the moderate and severe stages, they upregulate genes related to the recruitment of inflammatory cells, indicating potential interactions with macrophages (Fig. 4G). Collectively, our integrative analysis of the UGE atlas and ST data suggests that microenvironmental shifts are not isolated to individual cell states but emerge from coordinated, spatially organized cellular interactions within niches. This implies that clinical outcomes are shaped by the collective functional architecture of cell communities, rather than by the isolated phenotypes of single cell types.

Leveraging UniGeneX to characterize the formation of aberrant basaloid cells, we conclude that the regenerative axis linking RAS cells, AT2 cells, and AT1 cells is disrupted in PF. Importantly, we defined disease severity scores to capture the spatiotemporal evolution of alveolar niches and used these scores to identify concurrent changes in microenvironment composition and associated shifts in cellular states.

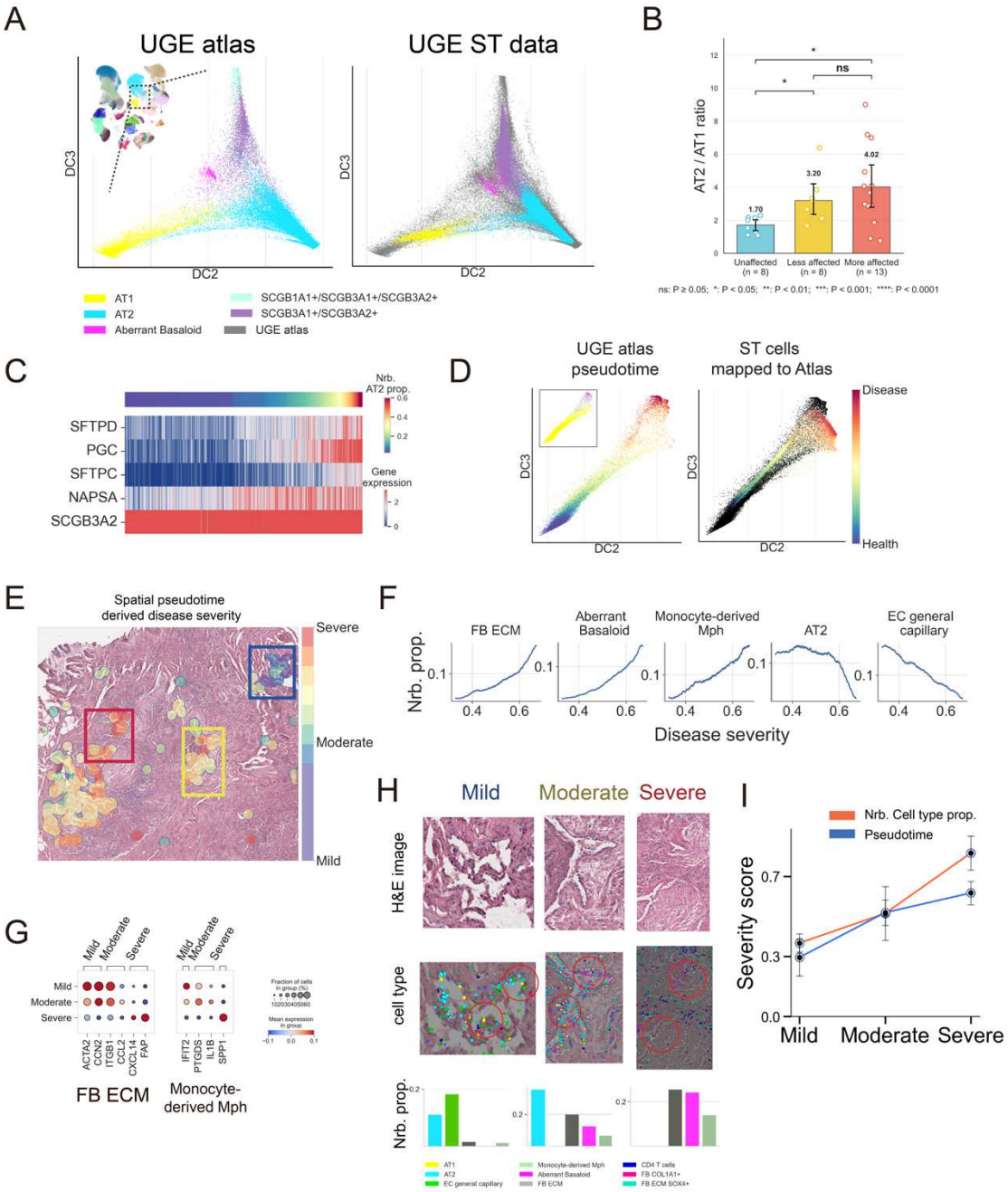



**Figure 4. Defining disease progression through the integrative analysis of the UGE atlas and ST data using UniGeneX.**

(A) Diffusion components (DCs) maps showing the trajectory relationship between 5 epithelial subtype cells from UGE atlas (left). Pseudocells from ST data were also aligned to the UGE atlas, colored by their assigned cell type (ST data), while UGE atlas cells are shown in gray.

(B) Bar plot showing the ratio of AT2 to AT1 cell counts across samples, stratified by disease status: unaffected, less affected, and more affected. Error bars represent the 95% confidence interval of the ratio, based on n = 8 (unaffected), n = 8 (less affected), and n = 13 (more affected) samples. Pairwise comparisons used two-sided Mann–Whitney U tests with Holm correction for three comparisons. Significance symbols denote Holm-adjusted P-values: ns, P ≥ 0.05; *, P < 0.05; **, P < 0.01; ***, P < 0.001; ****, P < 0.0001.

(C) Heatmap showing co-expression of AT2 markers and SCGB3A1$^{+}$/SCGB3A2$^{+}$ cell-specific markers in SCGB3A1$^{+}$/SCGB3A2$^{+}$ cells, ordered by the proportion (prop.) of neighboring (nrb.) AT2 cells. Color intensity denotes log-normalized expression.

(D) Disease severity is quantified using diffusion pseudotime along the trajectory from AT1 to aberrant basaloid cells in the UGE diffusion map (left). Spatial transcriptomics (ST) pseudocells were aligned to the UGE diffusion map and colored by transferred diffusion pseudotime values; UGE atlas cells are shown in gray (right).

(E) Contour map of spatial pseudotime for ST-derived cells mapped in (D), overlaid on an H&E-stained tissue section (sample VUILD106MA; IPF) in a representative region.

(F) Proportion of neighboring cell types, ordered by spatial pseudotime.

(G) Relative expression of stage DE genes in FB ECM and Monocyte-derived Mph represented as Z-scores. Relative expression (Z-scores) of stage-associated DE genes in FB ECM and monocyte-derived macrophage populations.

(H) Examples illustrating niche heterogeneity across regions of varying disease severity. Top row: H&E images of the boxed regions in (E). Middle row: H&E images overlaid with selected cell types mapped from ST data. Bottom row: mean proportion (prop.) of neighboring (nrb.) cell types for cells aligned to the UGE atlas in (D). The red circle in the middle row highlights representative ST spots mapped to the UGE atlas; the surrounding area reflects the local cellular composition associated with increasing disease severity.

(I) Severity scores of mapped cells in (D) for the three boxed regions in (E). The severity score is defined by both spatial pseudotime and the proportion (prop.) of neighboring (nrb,) cells of FB ECM, Monocyte-derived Mph and Aberrant Basaloid, which significantly increase with spatial pseudotime.

## An UGE atlas of glioma and human brain development deciphers gene programs of transitions in tumor cellular states

Gliomas are among the most lethal tumors of the central nervous system. Molecular classification has redefined glioma classifications, with mutations in isocitrate dehydrogenase (IDH) genes serving as a key distinguishing feature[27,28]. IDH-mutant gliomas are typically low-grade, grow slowly, and exhibit less aggressive behavior, whereas IDH-wildtype (IDH-wt) gliomas are frequently diagnosed as glioblastoma (WHO grade IV) and have a much worse prognosis[28]. A long-lasting question remains about the cellular origins of these tumors with different IDH genetic background. Several studies suggest that glioma stem cells possess radial glial-like features[29,30]. However, these analyses have largely focused on glioblastoma, or without considering the IDH genetic background[30,31]. More importantly, recent advances using genomic analysis like scRNA-seq and spatial transcriptomics have provided comprehensive views of tumor cellular states and microenvironment (TME) composition, clearly delineating differences between glioma with different IDH genetic background[32–35]. Therefore, dissecting how IDH-mut and IDH-wt tumors differentially recapitulate neurodevelopmental programs could uncover type-specific gene regulatory programs, which could serve as precision therapeutic targets. UniGeneX, as we demonstrated in PF disease progression analysis, has the ability to uncover the developmental trajectories that glioma tumors recapitulate by construct an UGE atlas of both developing human brain tissues and glioma tumor samples.

By leveraging a large-scale collection of publicly available scRNA-seq datasets spanning developing human brain tissues and glioma samples (Fig. 5A), we constructed and annotated a joint UGE atlas (Fig. 5B, Fig. S6A). Batch effects were effectively mitigated within this joint atlas (Fig. S6B-D). Utilizing this framework, we first identified conserved tumor cellular states shared across samples (Fig. 5C). As expected, glioma cells with different IDH genetic backgrounds show distinct molecular phenotypes (Fig. S6E). For IDH-wt cells, we defined 5 cellular states: Classical tumor cells (Classical-EGFR), which show high *EGFR* expression and high TCGA Classical signature[36]; Stem-like tumor cells (Stem-like), which enriched for neural stem cell markers such as *SOX2* and *NES*; Astrocyte-like tumor cells (MES-AC), which enriched for astrocyte markers such as *APOE* and *GFAP*; Mesenchymal-hypoxia tumor cells (MES-Hyp), which enriched for hypoxia responses signatures such as *HILPDA* and *EPAS1* expression; Mesenchymal-inflammation tumor cells (MES-Inflammatory), which enriched for inflammation related markers such as *CARD16* and *TMEM176B*. IDH-mutant tumor cells were resolved into three distinct states: an Astrocyte-like state reflecting astrocytic phenotypes (IDH-mut AC-like); an oligodendrocyte-like state characterized by the expression of myelin-associated genes such as PLP1 and MBP (IDH-mut Oligo-like); and a stem-like population that exhibited an enriched signature of stemness-related markers, including *SOX4* and *STMN1* (IDH-mut Stem-like).These annotations are consistent with previous single cell or bulk level studies (Fig. S6F-G)[32,33,37]. In addition, we also noted that tumor cycling cells are dominantly classified as classical or stem-like states, indicating the major proliferation roles in GBM (Fig. S6H).

Next, to investigate the divergent neurodevelopmental programs associated with distinct IDH genetic backgrounds, we performed Partition-based graph abstraction (PAGA)[38] on cells developmentally related (Fig. 5D). Strikingly, IDH-wt glioma cells align along a trajectory from radial glia, which agree with several previous studies[31]. We also performed correlation analysis,

which indicated higher similarity between radial glia (RG) and IDH-wt cycling and Stem-like cells (Fig. 5E). Several canonical markers for RG cells are also observed to have higher expression in these cellular states, including *NES*, *FABP7*, *VIM*, and *SOX2* (Fig. S7A). IDH-mut tumor cells shows reduced similarity with RG cells, while oligo-like state tumor cells enriched with higher similarity to oligodendrocyte precursor cells (OPCs), and expressed several markers of OPC, including *PDGFRA*, *CSPG4*, and *MYT1*(Fig. S7B). Together, these analyses suggest that IDH mutation status is linked to the degree of glioma cells resembling radial glia–like phenotypes.

We next sought to identify the hijacked gene programs that drive IDH-wt GBM to mimic the features of RG. Compared to IDH-mut cells, IDH-wt cells upregulated several pathways shared with RG identity, including EMT, hypoxia, angiogenesis, and increased fatty acid transport (Fig. 5F). Furthermore, IDH-wt GBM exhibited a significantly higher proliferative capacity. While IDH-mut stem-like cells do share certain developmental programs with RG, such as the regulation of glia cell differentiation, they more closely resemble an OPC-like state, characterized active WNT signalling and a bias toward myelination (Fig. 5F).

The UniGeneX model reveals that IDH-wt gliomas co-opt an early, multipotent radial glia–like developmental program, whereas IDH-mut gliomas adopt later, lineage-restricted OPC programs. The UGE atlas reveals a transcriptomic continuum across malignant states, suggesting that these cellular states are highly plastic and responsive to external signals. Given that the tumor microenvironment drives cellular plasticity in various cancers[39], we next investigate how these cellular states interact with their localized microenvironmental niches by utilizing this high-resolution UGE atlas for GBM.

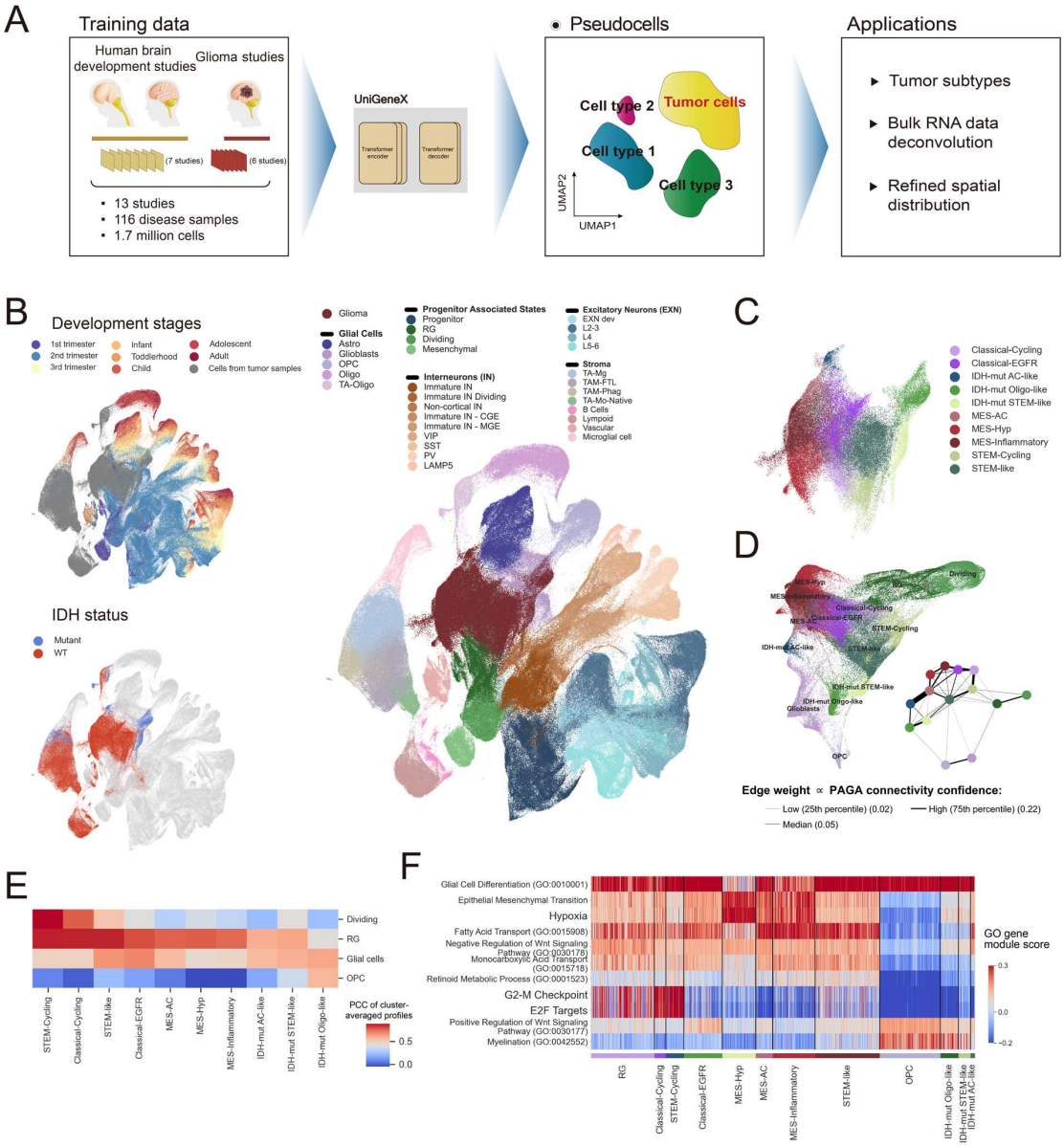



**Figure 5. UniGeneX distinguishes glioma cellular states across IDH mutation statuses.**

(A) UniGeneX workflow for glioma analysis. UniGeneX is trained on millions of human scRNA-seq cells from healthy brain and glioma samples. The model performs unbiased deconvolution for bulk RNA-seq and spatial transcriptomic analysis.

(B) UMAP of the UGE glioma atlas. (Left) Cells colored by developmental stage and tumor IDH status. (Right) Cells colored by cell type, including glioma, glial, progenitor, interneuron, excitatory neuron, and immune lineages. Astro, astrocytes; Oligo, oligodendrocytes; TA, tumor-associated; ‘‘dev’’ indicates early developing cells. MGE, medial ganglionic eminence; CGE, caudal ganglionic eminence; GE, ganglionic eminence. L2-3, upper-layer neurons; L4, layer 4 neurons; L5-6, deep-layer neurons; TAM, tumor-associated macrophage; Mg, microglia; Mo, monocytes; TAM-FTL, *FTL*-expressing tumor-associated macrophages; Phag, phagocytic; TAM-Phag, phagocytic tumor-associated macrophages; Mo, monocytes; TA-Mo-Native, tumor-associated native-like monocytes.

(C) UMAP showing glioma cellular states. Tumor cells are categorized into ten states: 3 IDH-mut and 7 IDH-wt. Cycling cells are primarily concentrated within IDH-wt Stem-like and Classical states.

(D) PAGA-initialized UMAP of glioma cells and RG-lineage cells. Line weight indicates connection strength between clusters. RG cells connect with IDH-wt cycling and STEM-like cells.

(E) Correlation heatmap of RG, globalists, and OPCs with glioma cellular states. RG correlate strongly with IDH-wt glioma cells, particularly IDH-wt cycling and Stem-like states. Color intensity represents Pearson correlation coefficients (PCC) between mean expression profiles of glioma cellular states (columns) and normal reference cell types (rows).

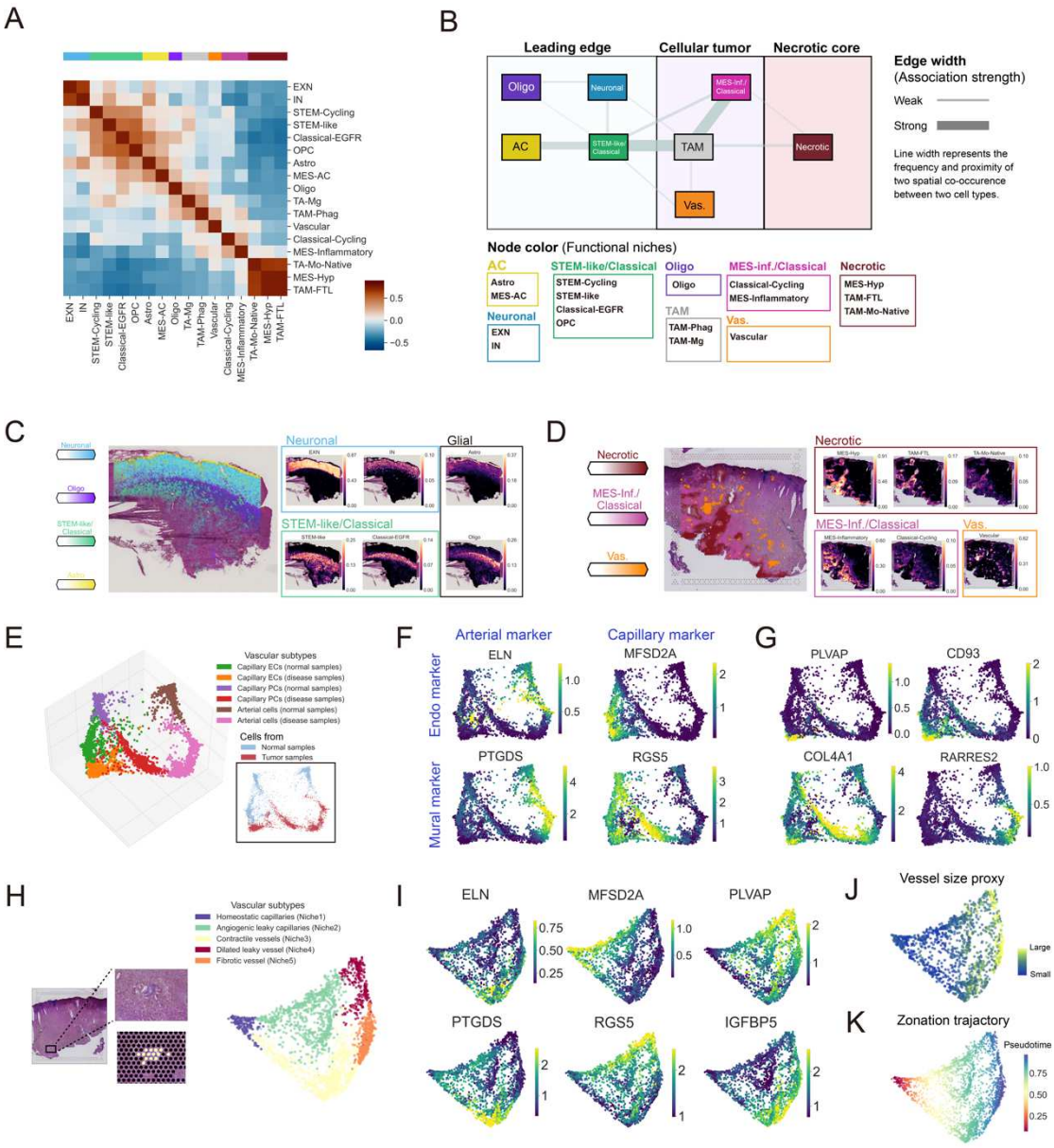
(F) GO term enrichment heatmap. Heatmap displays GO terms shared between RG and IDH-wt glioma, and between OPCs and IDH-mutant glioma cells.

## Vascular zonation associate with tumor cellular state transitions and niche architecture in GBM

Spatial transcriptomics has revolutionized our understanding of GBM heterogeneity, elucidating the complex architecture of the TME. A central feature of GBM architectural organization is the necrotic core, a histological phenotype strongly correlated with patient survival[40]. Previous studies, such as IVY-GAP, established that microvascular regions are enriched for proliferating cells, supporting the role of oxygen gradients in dictating cell state[41]. However, this model often overlooks the highly organized artery-capillary-venous zonation inherent to both normal brain and tumor vasculature. While this zonation is critical for normal brain homeostasis[42], its contribution to tumor development remains under-investigated, largely due to a lack of high-resolution spatial data and computational frameworks capable of dissecting these complex trajectories.

To address this gap, we integrated over 50 GBM tissue sections profiled by the Visium spatial transcriptomics platform from public repositories[33,35]. First, using the UGE atlas as a reference, we performed an unbiased deconvolution analysis on these spatial sections (Fig. S8A). Coupling these deconvolution results with spatial coordinates allowed us to identify distinct cellular niches defined by the local colocalization of specific cell types (Fig. S8A-B, Methods). We summarized these colocalization patterns using a heatmap, revealing stratified tissue regions with distinct cell-type compositions (Fig. 6A). This spatial architecture aligns with findings from IVY-GAP and other spatial omics studies (Fig. 6B)[33]. Within the GBM spatial landscape, Stem-like and Classical cellular states comprised the primary tumor mass and exhibited significant spatial overlap. Notably, Stem-like cells also co-localized with neuronal and glial signatures, suggesting a role in the tumor's invasive integration into the normal parenchyma (Fig. 6C, Fig. S8C). In contrast, the necrotic region remained spatially segregated from other areas and was characterized by MES-Hyp GBM cells infiltrated by two distinct TAM populations (Fig. 6A). Interestingly, the MES-Inflammatory state consistently localized adjacent to this hypoxic niche (Fig. 6D). Next, we evaluated the spatial distribution of vascular signatures, which exhibited moderate overlap with nearly all cell types, reflecting the ubiquitous presence of vasculature across GBM niches (Fig. 6A). However, it remains unclear whether these vessels maintain consistent phenotypes across tissues or adapt to specific niches.

Given the canonical artery-to-vein trajectory, we next sought to investigate vascular zonation across tumor tissues and evaluate its association with niche specialization. We first re-annotated vasculature-related cells in the UGE atlas (Fig. 6E). With established gene markers for arterial (*ELN* for endothelial cells, *PTGDS* for SMCs) and capillary (*MFSD2A* for endothelial cells, *RGS5* for pericytes) zones, we assigned vascular identities to these cell types (Fig. 6F). Furthermore, our analysis revealed a clear transcriptomic separation between normal and GBM vascular states in both mural and endothelial lineages, driven by the expression of GBM-specific markers such as *PLVAP*, *CD93*, *COL4A1*, and *RARRES2* (Fig. 6G). Elevated expression of these genes in GBM is frequently linked to increased vessel permeability and fibrosis. Although the UGE atlas lacks vein-specific data, its resolution was sufficient to distinguish arteries from capillaries for the purpose of zonation analysis.




## Figure 6. Spatial architecture and vasculature zonation in GBM.

(A) Spatial colocalization heatmap. Heatmap correlates the spatial distributions of cell types; cells with similar distributions cluster into three distinct regions: necrotic, cellular tumor, and leading edge.

(B) Schematic of GBM spatial architecture. Lines indicate association strength between niches, highlighting the necrotic, cellular tumor, and leading-edge layers.

(C) A Visium example. Cell-type distributions near the cellular tumor/leading edge.

(D) A Visium example. Cell-type distributions near necrotic core.

(E) PAGA-initialized UMAP of vasculature-related cells. Endothelial and mural cells are colored by sample source, with six major cell types annotated.

(F) Vasculature zonation markers. UMAP displaying expression of ELN (arterial endothelial cells), MFSD2A (capillary endothelial cells), PTGDS (Smooth muscle cells), and RGS5 (pericyte) to define vessel zonation. Color intensity represents log-normalized expression.

(G) Log-normalized expression of GBM-specific pathological markers in vasculature-related cells.

(H) UMAP of vasculature-containing Visium spots. Five niches are annotated based on transcriptional profiles.
(I) Log-normalized vasculature zonation marker expression across the same UMAP.
(J) Diffusion-map trajectory colored by local vascular abundance. Spots are colored by a local vascular abundance proxy, defined as the 30th percentile of the vascular proportion within a co-localized neighborhood comprising the target center spot and its six nearest spatial neighbors.
(K) Diffusion-map trajectory vascular Visium spots colored by velocity pseudotime, representing the inferred spatial zonation continuum.

We next sought to investigate the spatial vascular states and their associated neighboring cellular niches within the tissue microenvironment. However, due to the limited resolution of the Visium platform, we leveraged the spot cell-type compositions inferred from our UGE atlas, in combination with a novel computational model (Methods), to resolve the vascular transcriptomic profiles within vessel-dense spots (Fig. 6H, Fig. S8D). Based on established markers, the inferred vascular transcriptomic profiles tightly aligned with specific vascular zones, effectively establishing a spatial map of vascular niches (Fig. 6I). Capturing these spatially resolved vascular zones in situ represents a significant advancement, allowing us to map the functional heterogeneity of the tumor vasculature directly onto the intact tissue architecture. We then characterized the molecular properties of these niches: Niche 1 predominantly represented capillaries enriched for BBB integrity markers (e.g., *MFSD2A*, *CLDN5*), suggesting preserved normal function. Niches 2 represented angiogenic capillaries with increased leakage capacity, denoted by *PLVAP* and ESM1 expression. Niches 3 were enriched for contractile genes (*TAGLN*, *ACTA2*, *MYLK*). Niche 4 co-expressed leakage markers alongside smooth muscle markers (e.g., *NOTCH3*), indicative of the luminal enlargement of permeable vessels (Fig. 6I, Fig. S8E). Finally, Niche 5 represented regions with fibrotic features, characterized by high collagen expression (*COL1A1*, *COL1A2*, *COL3A1*; Fig. S8B). Beyond molecular features, vessel size, approximated by spatial spot density, differed significantly across these niches (Fig. 6J). Consistent with this, larger vessels often spanned multiple distinct transcriptomic states, reflecting a highly complex cellular composition (Fig. S8F). Together, these findings defined a spatial vessel-state trajectory in GBM marked by progressive leakage, vessel enlargement, and fibrosis, providing a framework for modeling vascular state progression in situ (Fig. 6K).

We next investigated whether specific vascular zonation states are associated with distinct tumor or normal cell populations. We observed significant compositional changes along the vascular zonation trajectory (Fig. 7A). We describe this vessel state progression in spatial through two stages (Fig. 7B). The first stage represented a progression of capillary leakage, transitioning from Niche 1 toward Niche 2. At the onset of this stage, Niche 1 vessels exhibited molecular features associated with BBB integrity and were surrounded by homeostatic glial cells (Fig. 7B). Notably, despite the maintained BBB integrity in these regions, Classical-EGFR tumor cells predominated around these vessels, suggesting a niche preference for healthy vasculature. Subsequently, along the transition from Niche 1 to Niche 2, we observed a progressive increase in the leakage marker *PLVAP* (Fig. 6I). This transition was accompanied by an increased proportion of MES-AC and Stem-like cells, alongside a TAM-Phag population characterized by strong phagocytic signatures. Ultimately, the accumulation of MES-

Inflammatory tumor cells suggested a link between TAM-derived inflammatory cues and tumor state transitions. The second trajectory represented pathological vessel enlargement. In contrast to the normal brain, GBM exhibited vessel expansion concurrent with progressive fibrosis (transitioning from Niche 3 toward Niches 4 or 5). Along this progression, cell-type proportions shifted dramatically, transitioning from an abundance of Stem-like and MES-AC tumor cells toward MES-Inflammatory or MES-Hyp states (Fig. 7B). Notably, we observed that fibrotic features were strongly correlated with MES-Hyp states, suggesting that niche factors beyond TA-Mo-Native or TAM-FTL contribute to this microenvironment.

Synthesizing these findings, we propose a model detailing how pathological vessel state progression drives tumor cellular state transitions in GBM (Fig. 7C). In this process, compromised, leaky vessels likely drive niche transitions by facilitating the infiltration of peripheral immune cells, particularly myeloid populations. During the earlier stages of this trajectory, infiltrating myeloid cells appear to differentiate into antigen-presenting macrophages, which we found to be closely associated with the transition of tumor cells from Classical to Stem-like, MES-AC, and ultimately MES-Inflammatory states. Ultimately, the necrotic core was strongly associated with vessel fibrosis and the presence of two distinct myeloid populations: TAM-FTL and native TA-Mo-Native. While it remains to be determined whether hypoxia induces the formation of these specific TAM populations or if these TAMs actively contribute to the establishment of the hypoxic niche, their spatial co-localization is undeniable. To our knowledge, this represents the first spatial framework to integrate highly resolved vascular zonation with tumor cellular state transitions in the field.

Most GBM studies place tumor cells at the center of analysis, although tumor cell states likely serve multiple functions within the microenvironment. Here, we instead used the vasculature as the organizing axis and asked how pathological vascular zonation relates to neighboring tumor and immune populations. Strikingly, vessel states were closely associated with tumor cellular states. In addition to tumor cells, myeloid populations also varied along the zonation trajectory, supporting a model in which reciprocal interactions among vessels, tumor cells, and immune cells shape distinct niches during tumor progression.

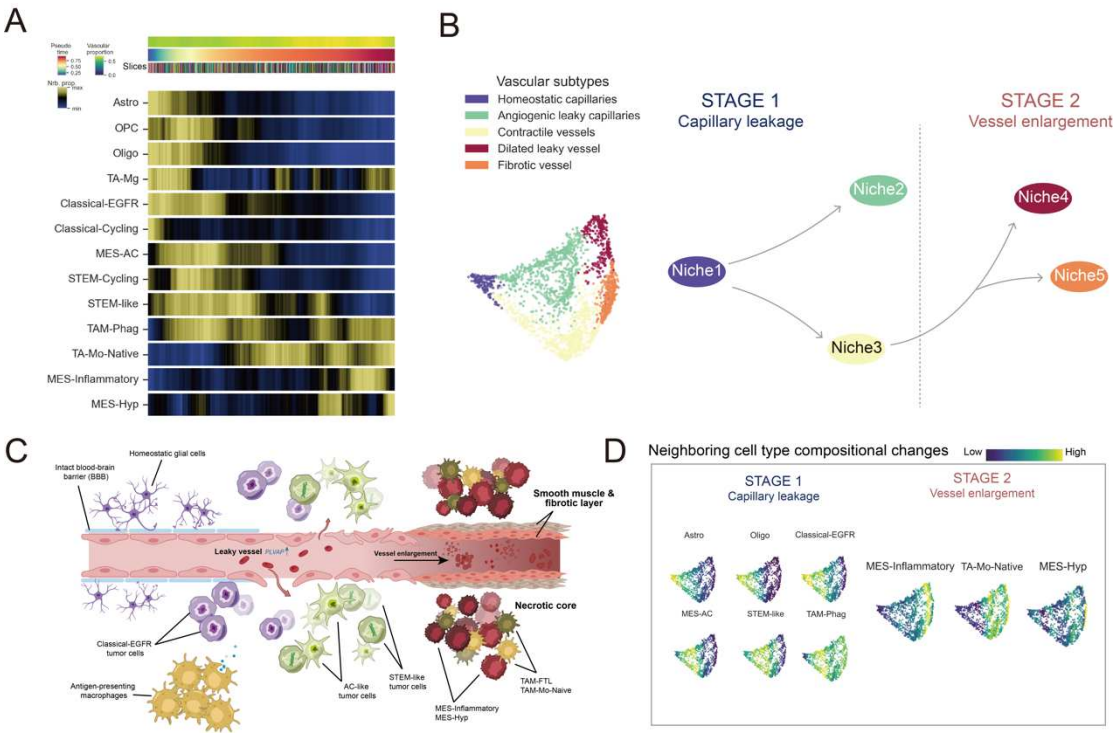



**Figure 7 GBM Vessel zonation correlated with tumor microenvironment changes.**

(A) Heatmap displaying shifting neighborhood proportions (Nrb. prop.) of cell types (rows) across individual spots of Visium ST data (columns), ordered along a vascular-state trajectory. Distinct cell types exhibit localized enrichment across different stages of vascular zonation. The top metadata tracks indicate spot pseudotime values (top), relative blood vessel size (middle), and tissue sections (bottom; "Slices").

(B) Model of pathological GBM vascular progression. We propose a two-stage progression of GBM vessel states: stage 1 corresponds to increasing capillary leakage, whereas stage 2 is characterized by vessel enlargement and fibrosis.

(C) Schematic summarizing pathological vascular progression in GBM and its associated changes in tumor cellular states.

(D) Representative neighboring cell-type compositional patterns across vascular-state progression. Color indicates the relative neighborhood proportion of cell types.

## Discussion

Human disease progression is a dynamic, spatially organized process. Because individual clinical samples inherently represent isolated temporal snapshots, mapping continuous disease trajectories requires the integration of large, heterogeneous cohorts. Although scRNA-seq has empowered the construction of high-resolution cellular atlases, computationally reconstructing long-range disease progression remains challenging. True biological trajectories construction is frequently obscured by technical artifacts, inter-patient variation, and differences in developmental stages[43]. Furthermore, bridging multimodal datasets, such as single-cell, bulk and spatial transcriptomics, is essential for a holistic understanding of disease. However, integrating these disparate platforms at scale remains a major bottleneck.

To address this gap, we developed UniGeneX, a generative atlasing framework. By combining transformer-based self-attention with a variational auto-encoder (VAE) architecture[44,45], UniGeneX captures global biological relationships across massive cohorts and maps them into an interpretable, universal gene expression (UGE) space. Unlike models constrained to learning abstract cell-level embeddings, a key feature of UniGeneX is its lightweight, flexible generative reconstruction capability, which enables context-specific interpretation of tissue-level data that is highly interpretable and actionable. Specifically, it facilitates reconstruction of molecular-level, long-range cell-state progressions. These trajectories can then be directly mapped onto spatial data for niche modelling or integrated with bulk cohorts to reveal clinical associations. This cross-modal integration reveals a continuous spectrum of cellular dynamics across diverse biological dimensions. To demonstrate the biological utility of this framework, we applied UniGeneX to dissect cell states across disease progression and resolve microenvironmental changes in two distinct pathologies: IPF and glioma.

In IPF, UniGeneX provided critical insights into the aberrant regenerative axes driving fibrogenesis. Our analysis revealed a disrupted developmental continuum linking RAS cells, AT2 cells, AT1 cells, and aberrant basaloid cells. We propose that the AT1 regenerative trajectory reflects a critical balance: transcriptional drift toward AT1 identity correlates with productive repair, whereas a diversion toward the aberrant basaloid state marks irreversible fibrotic remodeling. Beyond cell-state transitions, our severity scoring framework captured the continuous evolution of the alveolar microenvironment across tissues, revealing coordinated changes in local niche compositions and cellular states. Notably, even within individual clinical specimens, niche severity remained highly heterogeneous, underscoring that IPF progression is spatially asynchronous and locally dynamic.

In GBM, UniGeneX enabled us to construct what is, to our knowledge, the first spatial framework integrating high-resolution vascular zonation with tumor cellular state transitions. Our analyses suggest that vascular zonation is closely associated with shifts in tumor cellular states and immune composition, supporting a model in which pathological vessel progression interacts and could shape the surrounding tumor niche. In particular, compromised and leaky vessels correlated with transitions toward more inflammatory and aggressive tumor programs, potentially through enhanced infiltration of peripheral immune cells, especially antigen-presenting macrophage populations. Furthermore, the hypoxic necrotic core was associated with vessel fibrosis and with the spatial co-localization of TA-Mo-Native populations. Together, these findings provide a spatial framework connecting vascular remodeling with tumor-state transitions in GBM. However, as these relationships remain primarily associative, the causal

dynamics linking vessel zonation, hypoxia, immune remodeling, and tumor-state shifts will require direct experimental validation in the future.

Several limitations should be acknowledged. As with any atlas-based framework, rare or transient cell states may be underrepresented in the training data[46], which could constrain the model's ability to fully resolve the complete spectrum of disease progression. For example, in our GBM vasculature analysis, the limited number of vascular cells allowed us to resolve only part of the vascular zonation continuum, primarily capturing the artery-to-capillary transition. In addition, since the inferred progression structure is inherently influenced by cohort selection, unrepresented clinical stages may introduce bias into the learned continuum. Our GBM spatial analysis was similarly restricted to regions with sufficient vascular transcript coverage, potentially leaving out some specialized vessel states. This is particularly relevant in hypoxic core regions, where more advanced stages of vascular pathology may exist but remain unresolved at the current spatial resolution. Finally, while UniGeneX successfully identifies coordinated transitions across cell states and tissue niches, further mechanistic studies will be necessary to determine how these interactions arise and to evaluate their therapeutic potential.

Overall, UniGeneX represents a significant conceptual advance in applying attention mechanisms to decode complex disease progression. By integrating large-scale single-cell data with spatial context, the framework enables the reconstruction of disease-associated cellular spectra together with their accompanying microenvironmental changes. As high-resolution spatial technologies mature, improved training strategies and broader multimodal datasets will further refine model performance. Our analyses in IPF and GBM illustrate how this strategy can generate interpretable biological hypotheses and establish a foundation for future mechanistic and translational studies.

# Methods

As scRNA-seq technologies matures, a wealth of data has been accumulated. To build a highly interpretable and actionable atlasing method that can be used for integrative analysis with various other data types, including bulk, single-cell, spatial transcriptomics data, we develop UniGeneX. It is a generative atlasing method trained on millions of scRNA-seq data to construct a comprehensive UGE atlas.

## Training

***Gene selection*** We utilize a unique gene selection method to ensure high-quality input data. With training data encompassing tens of studies and millions of cells, instead of selecting an overall HVGs that could introduce significant noise and batch effects, we choose HVGs specific

to each dataset. Our novel model design allows the model only train on these dataset specific HVGs while still enable the construction of a UGE atlas that integrates all datasets.

Dataset specific HVGs are selected following standard strategy of HVG selection. First, raw count cell by gene transcriptional matrix is filtered and normalized to counts per million and then log+1 transformed within each dataset. Subsequently, 5,000 and 10,000 HVGs are separately selected for each dataset, with only these HVGs considered as candidates for the credible set of genes. Next, we create a credible set of genes that UniGeneX will generate for cells in UGE atlas. For each gene that is selected as HVG at lease one of datasets, the frequency of its occurrence in the HVG sets of 5,000 genes across different datasets is recorded. Genes that appear in the HVG sets of a greater number of datasets are selected for inclusion in the credible set (Fig. 1). The specific threshold for the number of datasets depends on computational costs and the diversity of the pretraining data. In our experiments with PF and GBM, we selected 2,000 to 3,000 genes. We then assess the overlap between each dataset's 10,000 HVGs and the credible set to serve as input for the transformer.

Note that this process leads to distinct input gene sets for each dataset. However, the transformer architecture, with its token design and attention mechanisms, enables flexible input lengths for different input samples while effectively integrates information from diverse datasets by relying on their overlapped genes, which we will introduce next. After training, UniGeneX constructs a UGE atlas from the extensive training data. For each cell, UniGeneX generates its universal gene expression profiles for all genes from credible gene set, including those unselected or unmeasured genes for that cell. This generative capability also mitigates batch effects during generation, integrating information across the large training dataset. The UGE construction is also applicable for newly coming data and will significantly enhances expression signals for inputs that only measure hundreds of genes by leveraging information learned from the training data.

Finally, a vocabulary is constructed for the credible gene set, treating each gene as a token and assigning a learnable gene embedding.

### *Transformer architecture*

*Embedding* In our model, analogous to how words function as tokens in a large language model, the token of a single cell consists of two parts, gene tokens and expression tokens. Each gene in the vocabulary is assigned a unique index. The gene token embedding is defined by a weight matrix $\boldsymbol{W_g} \in \mathbb{R}^{|\mathcal{V}| \times d}$, where $|\mathcal{V}|$ is the length of gene vocabulary and $d$ is the dimension of the output embedding. For gene with its index $id(g)$ in the vocabulary, $\boldsymbol{W_g}$ embeds it to the $id(g) - th$ row of $\boldsymbol{W_g}$, denoted as $\boldsymbol{e}_g = \boldsymbol{W_g}(id(g))$. The log-transformed expression value $x_{ig}$ of $i - th$ cell and $g - th$ gene is embedded by the expression embedding network $\boldsymbol{E}_{expr}: \mathbb{R} \to \mathbb{R}^d$: $\boldsymbol{e}_{expr} = \boldsymbol{E}_{expr}(x_{ig})$. The final embedding for $x_{ig}$, $i = 1, \cdots, N$, $g = 1, \cdots, G$, where $N$ is the number of cells and $G$ is number of genes is as following (Fig. S9A):

$$\boldsymbol{e}_{ig} = \boldsymbol{e}_g + \boldsymbol{e}_{expr} = \boldsymbol{W_g}(id(g)) + \boldsymbol{E_{expr}}(x_{ig})$$

With difference length of genes of cells in one batch, $< pad >$ token will be added to pad the input to a fixed length and these tokens will be ignored when applying attention mechanisms. A $< cls >$ token will be padded at the beginning of each cell's input to aggregate information of cells. Both $< pad >$ token and $< cls >$ token have their own unique index in vocabulary and their learnable embeddings $\boldsymbol{W_g}(id(< pad >))$ and $\boldsymbol{W_g}(id(< cls >))$.

*Transformer blocks* Network blocks in UniGeneX follows standard transformer layer architecture[8]. One block comprises of masked multi-head attention, a residual connection with layer normalization, a two-layer feed-forward network, and a second residual connection with layer normalization (Fig. S9B). The output of the block is as following,

$$\boldsymbol{h}^{(l)} = transformer_{block}(\boldsymbol{h}^{(l-1)}).$$

*Attention mechanisms* Given a hidden layer $\boldsymbol{h}^{(l)}$, three learnable linear projection transform $\boldsymbol{h}^{(l)}$ to the key matrix $K \in \mathbb{R}^{S \times d}$, the query matrix $Q \in \mathbb{R}^{S \times d}$ and the value matrix $V \in \mathbb{R}^{S \times d}$, where $S$ is the sequence length in this batch (Fig. S9C). The dot-product of $Q$ and $K$, following proper scaling function and masking of padding positions, yields a

weight matrix $W_A \in \mathbb{R}^{S \times S}$. The dot-product between $W_A$ and the value matrix $V$ will give the output for this dot-product attention:

$$Attention(\boldsymbol{h}^{(l)}) = scaling_and_mask\ (QK^T) \cdot V$$

Multiple dot-product attentions form a multi-head attention block, facilitating learning different aspects of the data. The outputs of the individual heads are concatenated and pass through a final linear layer to get final output of the attention block.

*Masked autoencoder* Inspired by the success of masked autoencoder in computer vision[47], we randomly mask genes from the input cell and reconstruct masked genes in decoder. An asymmetric encoder-decoder architecture is applied (Fig. S9D), with 12 transformer blocks in the encoder and 4 in decoder. Instead of inserting a $< mask >$ token in the encoder like other transformer-based models, we remove masked tokens, so the encoder only operates on the "visible" subset of genes (Fig. S9D). This strategy combined with the asymmetric architecture, where encoder is deep but only process half of the token (mask ratio is 0.5 in our case) and decoder operates on full set of tokens but is shallow, resulting in large reduce in training time while still optimizing accurately.

The input to the decoder is the full set of gene tokens embeddings, including masked ones in the encoder. To also incooperate cell context information, the output of $< cls >$ token from encoder is concatenated with all gene tokens and pass through two linear layers to get the input for decoder.

$$\boldsymbol{h}_0^{(dec)} = MLP\big(\big[\boldsymbol{h}_L^{(enc)}(< cls >),\ \boldsymbol{W_g}\big(id(g)\big)\big]\big),$$

where $\boldsymbol{h}_L^{(enc)}(< cls >)$ represent the encoder output of $< cls >$ token and $\boldsymbol{h}_0^{(dec)}$ represent the input for the decoder.

***Transformer-based probabilistic model*** UniGeneX is a generative atlasing method that can construct an UGE atlas. It achieves generative ability by applying a variational autoencoder structure where a normal prior is added to the latent variable $\boldsymbol{z}$, which is $\boldsymbol{h}_L^{(enc)}(< \boldsymbol{cls} >)$ is our model,

$$\boldsymbol{z} = \boldsymbol{h}_L^{(enc)}(< cls >) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}).$$

In generative process, the normalized expression observation $Y_{ig}$ is model as Poisson distribution and reconstruction loss follows negative log-likelihood of Poisson distribution:

$$\mathcal{L}_{expr} = \sum_{i,g} \left( \hat{Y}_{i,g} - Y_{i,g} \log \hat{Y}_{i,g} \right),$$

where $\hat{Y}_{i,g}$ is the reconstructed expression from the network and we model it as the following:

$$\log \hat{Y}_{i,g} \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mu, \sigma^2 = f_\theta(\boldsymbol{z})$$

Denote the posterior distribution as $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$, we fit the model by maximize the following variational lower bound,

$$\log p_\theta(\boldsymbol{x}) \geq \mathcal{L}(\theta, \phi; \boldsymbol{x}) = -D_{KL}\left( q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,||\, p_\theta(\boldsymbol{z}) \right) + \mathbb{E}_{q_\phi(\boldsymbol{z} \mid \boldsymbol{x})}(\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})).$$

Empirically, a hyperparameter $\beta$ is added in front of KL regularization term[45], resulting in the final objective function:

$$l(\theta, \phi) = \beta D_{KL}\left( q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,||\, p_\theta(\boldsymbol{z}) \right) + \mathcal{L}_{expr}$$

Notably, setting $\beta$ to a non-zero value is crucial for constructing a UGE atlas. When $\beta = 0$, the model reduces to maximum likelihood estimation and tends to memorize the data, preserving batch effects across datasets. A non-zero $\beta$ imposes a prior on the latent variables, forcing the model to integrate information across datasets and learn the underlying gene-expression patterns; this yields a smoother, more continuous latent space in which similar cell states from different samples lie close together, producing similar reconstructions in UGE atlas.

## Inference

Newly acquired data, either be single-cell data or spatial data with limited measurement of genes, undergoes a similar gene selection method as in the training stage. This includes transforming the data to a log scale, selecting HVGs, and ensuring overlap with credible gene set. We eliminate mask operations and the input for decoder is all gene tokens from the credible gene set concatenated with cell specific latent variable (Fig. S9E). This enables generation of the UGE for newly acquired data by leveraging information from the training set. Integrative analysis between these UGEs and the UGE atlas is straightforward. For example, cell type annotation, pseudotime assignment.

***Cell type harmonization*** To harmonize cell type labels across diverse datasets, we employed a majority voting method. We applied standard dimensionality reduction and unsupervised clustering to UGE atlas, then counted the annotation labels from the original datasets within each cluster (Fig. S9F). Cell type labels were ranked by their percentage within each cluster. Since UniGeneX reconstruct training data in UGE space, different labels often correspond to the same cell states within a cluster. This strategy also benefits from the comprehensiveness of UGE atlas, cell's state can reference to diverse studies. We can select the final label as the top-ranked label for the cluster or make modifications based on the subsequent ranked labels.

## Building two scenario data

***Scenario 1: Xenium gene panel*** Human lung ST data from Xenium platform was downloaded from 10x Genomics (xenium-human-lung). Held-out testing datasets from HLCA' gene was intercept with Xenium human lung panel.

***Scenario 2: Read-depth subsampling*** To evaluate low-depth robustness, raw test counts were subsampled to fractions of the original read depth (1%, 5%, 10%, 25%, 50%). For cell $i$, with count vector $x_i = (x_{11}, x_{12}, \cdots, x_{iG})$ and library size $L_i = \sum_g x_{ig}$, the retained library size was

$$M_i = floor(rL_i), \quad with\, r \in [0.01, 0.05, , 0.10, 0.25, 0.50]$$

Subsampling was performed at the UMI level rather than by multiplying counts. We sampled $M_i$ UMI identities from the total $L_i$ uniformly without replacement to obtain the subsampled profile $x_i^{(r)}$:

$$x_i^{(r)} \sim Multivariate\ Hypergeometric\ (x_i, M_i)$$

$$\sum_g x_{ig}^{(r)} = M_i$$

This preserves the empirical gene-composition structure of each cell while reducing its library size to approximately $r$ of the original depth.

## Cell type annotation and atlas mapping

***UniGeneX*** The UGE profiles were generated with the retrained UniGeneX training checkpoint without any fine-tuning. Specifically, normalized and log-transformed data intercect with credible gene set and then to tokenized for both expression and gene identify and pass though the trained UniGeneX model. UniGeneX generates UGE for each cell with gene space of credible gene set. To obtain annotation predictions, each UGE profile was projected into the HLCA atlas PCA space and matched to atlas cells by cosine k-nearest neighbors (k = 30). For query cell $q$, predicted fine annotation was assigned by neighbor majority vote:

$$score_q(c) = \frac{1}{k} \sum_{a \in N_k(q)} I(y_a = c)$$

$$\hat{y}_q = \arg\max_c score_q(c)$$

$$P_q = \max_c score_q(c)$$

where $N_k(q)$ is the atlas-neighbor set, $y_a$ is the atlas cell type annotation of cell $a$, and $p_q$ is the mapping confidence. In benchmarking, cells of which confidence lower than 0.5 are considered low quality and filtered.

## Compared methods

In few-shot settings, for each cell type, we randomly sampled at most *K* labeled cells, where *K* is the specified shot number, and used this balanced subset to train the classifier. The trained classifier was then evaluated on the matched held-out query cells to measure annotation performance under limited-label conditions.

***scGPT*** scGPT was evaluated using the published cell-type annotation fine-tuning workflow with the pretrained human scGPT checkpoint. Expression matrices were preprocessed as binned inputs (n_bins = 51), matched to the pretrained gene vocabulary, tokenized with a leading <cls> token, and padded to a maximum sequence length of 3001; zero-count genes were excluded from the input sequence. The transformer classifier was fine-tuned for 10 epochs using the cell-type classification objective with cross-entropy loss. Optimization used Adam with learning rate 1e-4, mixed precision training, dropout 0.2, and a step learning-rate scheduler with decay factor 0.9.

***scFoundation*** scFoundation was evaluated using the published model checkpoint and linear-probing annotation workflow. Input matrices were first aligned to the scFoundation 19,264-gene vocabulary (OS_scRNA_gene_index.19264.tsv); genes absent from a degraded input were zero-filled, and all genes were ordered according to the required vocabulary. Counts were library-size normalized, log-transformed, and augmented with read-depth features before being passed to the pretrained scFoundation cell encoder (models.ckpt). A linear-probing classifier was trained on the resulting cell representations to predict cell-type labels using cross-entropy loss, Adam optimization with learning rate 1e-4, mixed precision, gradient clipping, 10 training epochs, and the same scheduler decay factor of 0.9. The fine-tuned classifier was then used to assign fine cell-type labels and class scores to the matched held-out query cells.

## Benchmarking metrics

All methods were evaluated with the same post-processing rules. Cells with missing or "Unknown" true labels were removed. The held-out datasets HLCA_Schultze_unpubl and HLCA_Shalek_2018 were excluded from metric summaries because of low quality. To reduce instability from extremely rare labels, fine cell types were retained only if they had at least 25 cells within a given dataset group.

Metrics were computed independently for each held-out dataset and then summarized across datasets. For each class $c$, $TP_c$ denotes cells truly belonging to class $c$ and predicted as $c$; $FP_c$ denotes cells from other classes incorrectly predicted as $c$; $FN_c$denotes cells from class $c$ incorrectly predicted as another class; and $TN_c$ denotes cells from other classes correctly not predicted as $c$. Precision, recall, and F1 score were then defined as:

$$Precisoin_c = \frac{TP_c}{TP_c + FP_c}$$

$$Recall_c = \frac{TP_c}{TP_c + FN_c}$$

$$F1_c = 2 * \frac{Precisoin_c * Recall_c}{Precisoin_c + Recall_c}$$

Weighted F1 was used as the primary multiclass score:

$$Weighted\ F1 = \sum_c \frac{n_c \cdot F1_c}{\sum_c n_c}$$

Where $n_c$ is the number of cells in cell type $c$.

Accuracy was computed as:

$$Accuracy = \frac{\#\ correctly\ predicted\ cells}{\#\ evaluated\ cells}$$

## PF ST data cell type annotation with UniGeneX

The raw cell-by-gene transcriptional matrix is filtered and normalized to counts per ten thousand due to the low measurement depth and limited number of measured genes inherent to Xenium technology. Then log+1 transform is applied. With UniGeneX model trained on millions of lung single-cell data, the normalized spatial cell-by-gene data is tokenized for both expression and gene identify and pass though the trained UniGeneX model. The model output the UGE for each cell in ST data which contain the expression of credible gene set with nearly two thousand genes. With UGE atlas and UGE of ST data, the cell type label from UGE atlas can be transferred to ST data using nearest neighbor search algorithm *sklearn.neighbors.NearestNeighbors.* More specifically, the principal components (PC) loadings are calculated using UGE atlas, the corresponding PCA representations are fitted to a nearest neighbor model with *n_neighbors = 30.* The UGE of ST data is projected to PC loadings of UGE atlas and 30 nearest neighbors from UGE atlas are found for each cell in ST data using trained nearest neighbor model. The cells which the composition of dominated cell type in nearest neighbor of UGE atlas smaller than 0.3 are considered low quality and filtered. The other cells are assigned label of dominated cell type in nearest neighbor of UGE atlas.

## Variation of PF cell gene expression with neighboring cell types

Given a target cell type, for example, SCGB3A1+/SCGB3A2+ in our case, 30 neighboring cells in spatial context are selected using *sc.pp.neighbors*. Cell type composition $Z_{nrb} \in \mathbb{R}^{N_{ST} \times M}$, where $N_{ST}$ is the number of ST cells and $M$ is the number of cell types, are calculated for each cell which is considered as cell's microenvironment. To investigate cell state variation along with variation of microenvironment, for example, the molecular variation of SCGB3A1+/SCGB3A2+ along its proximation to AT2 cells, cells are order by the neighboring AT2 composition and genes of interest are visualize after smoothing.

## Spatial pseudotime

Diffusion pseudotime are assigned to UGE atlas first, using function *velocity_pseudotime* from scVelo package. Next, UGE of ST data are mapped to UGE atlas the same way as in cell type annotation with neighbor=300. The resulting "connectivities" matrix $C \in \mathbb{R}^{N_{ST} \times N_{sc}}$, where $N_{ST}$ is the number of ST cells mapped to UGE atlas, and $N_{sc}$ is the number of UGE atlas cells used to be mapped. The index of interested cells (cells along trajectory from AT1 to Aberrant Basaloid cells) are used to get a subset connectivities matrix $C_{sub} \in \mathbb{R}^{N_{ST} \times \tilde{N}_{sc}}$. This matrix is then normalized by row and multiply the pseudotime from UGE atlas,

$$t_{sp} = C_{sub} \cdot t_{sc},$$

where $t_{sc} \in \mathbb{R}^{\tilde{N}_{sc}}$ is the pseudotime of interest cells in UGE atlas and $t_{sp}$ is the transferred pseudotime from UGE atlas to ST data.

## Neighboring cell molecular variation along pseud time

Given cells from AT1 to Aberrant Basaloid trajectory, to investigate the neighboring cell type, for example, FB ECM's gene expression variation, cells that contain FB ECM as neighbors are filtered and the neighboring FB ECM's mean expression are calculated. Cells are further divided into 3 groups, early pseudotime, median pseudotime and late pseudotime. The DE analysis is then applied to find DE genes between stages.

## Bulk data deconvolution

Normalized matrices from the Cancer Genome Atlas (TCGA) bulk RNA-seq samples were downloaded from the Genomic Data Commons (GDC). The cohort includes 695 samples in total, comprising 446 IDH-mutant and 249 IDH-wild-type cases. RCTD[2] model is used to deconvolute bulk samples using UGE atlas as reference data. Developing cell states are removed from reference. Specifically, cell from samples in the first, second or third trimester or cell types predominantly originating from immature samples. Tumor cell states in UGE atlas are divided into IDH-wild-type specific or IDH-mutant specific, and two groups are used as reference to deconvolute IDH-wild-type bulk samples or IDH-mutant bulk samples respectively.

## GBM ST data deconvolution

The spots with total counts smaller than 1000 are filtered and slices with median total counts smaller than 2000 are also filtered. With low resolution Visium data, to identify cell type composition in spots, we used Cell2location[1] model to deconvolution the Visium spots to cell state compositions with IDH-wide-type specific UGE atlas as reference data, the same as used for TCGA bulk data. We include the following modification of Cell2location model to get better deconvolution results. In Cell2location model, the elements of spatial expression count matrix $d_{s,g}$ are modeled as negative binomial distribution,

$$d_{s,g} \sim NB(\mu_{s,g}, \alpha_{e,g}),$$

with locations $s$, gene $g$ and multiple datasets $e$. The expression rate $\mu_{s,g}$ is further modeled as a linear function of reference cell type signatures $g_{f,g}$:

$$\mu_{s,g} = \left( m_g \cdot \sum_f w_{s,f} g_{f,g} + s_{e,g} \right) \cdot y_s,$$

where $w_{s,f}$ denotes the abundance of cells expressing cell type $f$ at location $s$; $m_g$ is a gene-specific scaling parameter and $s_{e,g}$ is gene-specific additive shift. Originally, $s_{e,g}$ is modeled as Gamma distribution which parameter are learned from the data. However, we found the Gamma distribution could not adjust for the background expression of genes, which may bias the deconvolution results. To address this, we down weight these highly expressed genes by including the observed background gene expression in $s_{e,g}$ as following,

$$s_{e,g} \sim N(\log(\bar{d}_{\cdot g} + \varepsilon), 1),$$

where $\bar{d}_{\cdot g}$ is the observed log-normalized mean of gene $g$. We adjust Cell2location model parameters N_cells_per_location=20 and the others remain the default setting.

## Spatial distribution graph

With highly irregular distribution of cell types in spatial, we partition the slice into 4 regions according to the enrichment level of certain cell type $C_{center}$. For any other cell type $C$, the non-

zero mean of cell type composition in each region are calculated in each region to get a $4$ dimensional vector $w_C^{(C_{center})}$. Next, we randomly shuffle the position of spots $10$ times and calculate the baseline $w_{C,i}^{(C_{center})}$, $i = 1, 2, \cdots, 10$. We get the z-score of $w$ as following,

$$z_c^{C_{center}} = \frac{w_C^{(C_{center})} - w_{C,mean}^{(C_{center})}}{w_{C,mean}^{(C_{center})} - w_{C,min}^{(C_{center})}}$$

where $w_{C,mean}^{(C_{center})} = \frac{1}{10}\sum_{i=1}^{10} w_{C,i}^{(C_{center})}$, $w_{C,min}^{(C_{center})} = \min_i w_{C,i}^{(C_{center})}$. Then the z-score are binarized to a $4$ dimensional one hot matrix to assign cell type $C$ to one of the regions of $C_{center}$. Concatenate all cell type result in a matrix $Z^{(C_{center})} \in \mathbb{R}^{F \times 4}$, where $F$ is the number of cell type. The rows whose entropy smaller than $0.8$ are set to all $0$, which are considered do not have clear distribution. Concatenate all partition way by different center cell type, result in the final matrix z-score matrix $Z \in \mathbb{R}^{F \times (4F)}$. The correlation between rows of $Z$ are considered colocalization score between two cell types. The linkage algorithm can be further applied to group cell types.

To further investigate the relative spatial distance between groups of cell types, $\mathcal{G}_1$ and $\mathcal{G}_2$, the weight between them is calculated as following. Consider a slice partition with respect to $C_{center}$, $\mathcal{G}_1$ gather in partition $i$ and $\mathcal{G}_2$ gather in partition $j$. Record that partition represent the enrichment level of $C_{center}$. If the enrichment level of partition $i$ and partition $j$ is the same or adjacent, the weight between $\mathcal{G}_1$ and $\mathcal{G}_2$ is increased by 10. In this case we consider $\mathcal{G}_1$ and $\mathcal{G}_2$ to be relatively close in distance. Otherwise, the weight does not increase. Finally, a graph with cell type groups as nodes and beforementioned weight as edge weight, can be built to indicate the spatial distribution of cell types.

## Vascular analysis

We model the gene expression of spot $s$ and gene $g$, $Y_{s,g}$, as a sample drawn from a Gamma Poisson distribution,

$$Y_{s,g} \sim GammaPoisson(l_s \cdot \mu_{s,g}, \alpha_g)$$

where the mean expression is the multiplication between $\mu_{s,g}$ and library size $l_s$. The parameter $\alpha_g$ is the gene-specific dispersion term. We further model the mean term $\mu_{s,g}$ as following,

$$\log \mu_{s,g} = \left( \beta_{s,vas}\, x_{s,g,vas} + \sum_{k \neq vas} \beta_{s,k} x_{s,g,k} \right),$$

where $\beta_k$ is the cell type proportion of cell type $k$ in spot $s$ derived from Cell2location model and $x_{s,g,k}$ is the underline expression of cell type $k$ in spot $s$ of gene $g$ after removing the effect of cell type proportion.

To infer the underline expression of vascular and investigate their variation in disease progression, we selected spots that $\beta_{s,vas} > 0.5$ and genes which $\beta_{s,vas}\, x_{s,g,vas}$ will dominant $\sum_{k \neq vas} \beta_{s,k} x_{k,s,g}$. More specifically, we empirical use the following threshold to select genes that considered expressed by vascular:

$$Q_q \left( \frac{\beta_{s,vas}\, \bar{x}_{g,vas}}{\sum_{k \neq vas} \beta_{s,k} \bar{x}_{g,k}} \right) > 3,$$

where $Q_q$ indicate the $q$-th quantile and we choose $q = 0.7$. The mean expression of cell type $k$ and gene $g$, $\bar{x}_{g,k}$ is the inferred single-cell reference signatures by Cell2location model. By filter spots and genes, we could omit $\sum_{k \neq vas} \beta_{s,k} x_{s,g,k}$ term. To remove batch effect between samples, we further model $x_{s,g,vas}$ as following,

$$x_{s,g,vas} = \tilde{x}_{s,g,vas} + x_g^{batch}$$

where $x_g^{batch}$ is the slice-specific gene-wise batch effect. Approximate variational inference is used to estimate underline vascular expression $\tilde{x}_{s,g,vas}$, implemented by pyro. The inferred $\tilde{x}_{s,g,vas}$ is further used to identify vascular cell states variations in tumor progression.

## Data availability

The HLCA for validation and building PF atlas is publicly available and can be downloaded via cellxgene (https://cellxgene.cziscience.com/collections/6f6d381a-7701-4781-935c-db10d30de293). The Xenium PF dataset can be obtained from GEO under accession code GSE250346. The GBM atlas can be accessed under GEO series GSE199762, GSE217511, GSE182109, GSE199762, https://cells-test.gi.ucsc.edu/?ds=early-brain, GSE204684, GSE168408, GSE162170, GSE131928, GSE163120 and https://cellxgene.cziscience.com/collections/999f2a15-3d7e-440b-96ae-2c806799c08c. GBM Visium data can be obtained from http://(www.gbmspace.org/ and https://zenodo.org/records/12624860.

## Code availability

The Python package for UniGeneX is hosted at https://github.com/xwanaf/UniGeneX/tree/main. The code used for the reproducibility of UniGeneX's downstream analysis can be found in the GitHub repository https://github.com/xwanaf/UniGeneX_reproducibility.


## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (T2422018); the Innovation and Technology Commission (ITC-SKLNSD26SC01); the Research Grants Council of Hong Kong (YCRF C6001-23Y, T12-611/25-N, and JLFS/M-604/24); Schmidt Sciences; Hong Kong Research Grants Council grants (AoE/E-601/24-N, C6040-24G, 16307221, 16307322, 16302823, 16309424, and 16308925); and The Hong Kong University of Science and Technology Startup Grants R9405 and Z0428 from the Big Data Institute. M.H. was supported by the Hong Kong Research Grants Council General Research Fund (17127626). J.X. was supported by the National Natural Science Foundation of China (12401384), the Guangdong Natural Science Foundation General Project (2025A1515011603), and a Sun Yat-sen University Startup Grant.


## Declaration of interests

The authors declare no competing interests.


1. Kleshchevnikov, V. *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.* **40**, 661–671 (2022).
2. Cable, D. M. *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.* **40**, 517–526 (2022).
3. Elmentaite, R., Domínguez Conde, C., Yang, L. & Teichmann, S. A. Single-cell atlases: shared and tissue-specific cell types across human organs. *Nat. Rev. Genet.* **23**, 395–410 (2022).
4. Gavish, A. *et al.* Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. *Nature* **618**, 598–606 (2023).

5. Hanahan, D. Hallmarks of cancer—Then and now, and beyond. *Cell* (2026).
6. Hogan, B. L. *et al.* Repair and regeneration of the respiratory system: complexity, plasticity, and mechanisms of lung stem cell function. *Cell Stem Cell* **15**, 123–138 (2014).
7. Roohani, Y. H. *et al.* Virtual Cell Challenge: Toward a Turing test for the virtual cell. *Cell* **188**, 3370–3374 (2025).
8. Vaswani, A. *et al.* Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, (2017).
9. Szałata, A. *et al.* Transformers in single-cell omics: a review and new perspectives. *Nat. Methods* **21**, 1430–1443 (2024).
10. Cui, H. *et al.* Towards multimodal foundation models in molecular cell biology. *Nature* **640**, 623–633 (2025).
11. Tejada-Lapuerta, A. *et al.* Nicheformer: a foundation model for single-cell and spatial omics. *Nat. Methods* 1–14 (2025).
12. Cui, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).
13. Hao, M. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat. Methods* **21**, 1481–1491 (2024).
14. Sikkema, L. *et al.* An integrated cell atlas of the lung in health and disease. *Nat. Med.* **29**, 1563–1577 (2023).

15. Hogan, B. L. *et al.* Repair and regeneration of the respiratory system: complexity, plasticity, and mechanisms of lung stem cell function. *Cell Stem Cell* **15**, 123–138 (2014).

16. Adams, T. S. *et al.* Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci. Adv.* **6**, eaba1983 (2020).

17. Habermann, A. C. *et al.* Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *Sci. Adv.* **6**, eaba1972 (2020).

18. Mayr, C. H. *et al.* Spatial transcriptomic characterization of pathologic niches in IPF. *Sci. Adv.* **10**, eadl5473 (2024).

19. Vannan, A. *et al.* Spatial transcriptomics identifies molecular niche dysregulation associated with distal lung remodeling in pulmonary fibrosis. *Nat. Genet.* **57**, 647–658 (2025).

20. Franzén, L. *et al.* Mapping spatially resolved transcriptomes in human and mouse pulmonary fibrosis. *Nat. Genet.* **56**, 1725–1736 (2024).

21. Liu, K. *et al.* Tracing the origin of alveolar stem cells in lung repair and regeneration. *Cell* **187**, 2428-2445. e20 (2024).

22. Basil, M. C. *et al.* Human distal airways contain a multipotent secretory cell that can regenerate alveoli. *Nature* **604**, 120–126 (2022).

23. Pezet, M. G. *et al.* Human respiratory airway progenitors derived from pluripotent cells generate alveolar epithelial cells and model pulmonary fibrosis. *Nat. Biotechnol.* **44**, 394–405 (2026).

24. De Man, R. *et al.* Single-cell atlas of human lung aging identifies cell type dyssynchrony and increased transcriptional entropy. *Nat. Commun.* (2026).

25. Ireland, A. S. *et al.* Basal cell of origin resolves neuroendocrine–tuft lineage plasticity in cancer. *Nature* **647**, 257–267 (2025).

26. Kadur Lakshminarasimha Murthy, P. *et al.* Human distal lung maps and lineage hierarchies reveal a bipotent progenitor. *Nature* **604**, 111–119 (2022).

27. Ceccarelli, M. *et al.* Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* **164**, 550–563 (2016).

28. Molinaro, A. M., Taylor, J. W., Wiencke, J. K. & Wrensch, M. R. Genetic and molecular epidemiology of adult diffuse glioma. *Nat. Rev. Neurol.* **15**, 405–417 (2019).

29. Mossi Albiach, A., Janusauskas, J. & Kapustova, I. Glioblastoma is spatially organized by neurodevelopmental programs and a glial-like wound healing response, bioRXiv. *Prepr. Httpsdoi Org101101202309* **1**, (2023).

30. Bhaduri, A. *et al.* Outer radial glia-like cancer stem cells contribute to heterogeneity of glioblastoma. *Cell Stem Cell* **26**, 48-63. e6 (2020).

31. Pollen, A. A. *et al.* Molecular identity of human outer radial glia during cortical development. *Cell* **163**, 55–67 (2015).

32. Neftel, C. *et al.* An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell* **178**, 835-849. e21 (2019).

33. Greenwald, A. C. *et al.* Integrative spatial analysis reveals a multi-layered organization of glioblastoma. *Cell* **187**, 2485-2501. e26 (2024).

34. Saraswat, M. *et al.* Decoding plasticity regulators and transition trajectories in glioblastoma with single-cell multiomics. *BioRxiv* 2025.05. 13.653733 (2025).

35. de Jong, G. *et al.* A spatiotemporal cancer cell trajectory underlies glioblastoma heterogeneity. *biorxiv* 2025.05. 13.653495 (2025).

36. Verhaak, R. G. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98–110 (2010).

37. Venteicher, A. S. *et al.* Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science* **355**, eaai8478 (2017).

38. Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).

39. Guelfi, S., Hodivala-Dilke, K. & Bergers, G. Targeting the tumour vasculature: from vessel destruction to promotion. *Nat. Rev. Cancer* **24**, 655–675 (2024).

40. Rong, Y., Durden, D. L., Van Meir, E. G. & Brat, D. J. 'Pseudopalisading'necrosis in glioblastoma: a familiar morphologic feature that links vascular pathology, hypoxia, and angiogenesis. *J. Neuropathol. Exp. Neurol.* **65**, 529–539 (2006).

41. Puchalski, R. B. *et al.* An anatomic transcriptional atlas of human glioblastoma. *Science* **360**, 660–663 (2018).

42. Wälchli, T. *et al.* Single-cell atlas of the human brain vasculature across development, adulthood and disease. *Nature* **632**, 603–613 (2024).

43. Argelaguet, R., Cuomo, A. S., Stegle, O. & Marioni, J. C. Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.* **39**, 1202–1215 (2021).

44. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *ArXiv Prepr. ArXiv13126114* (2013).

45. Sankarapandian, S. & Kulis, B. $\beta $-Annealed Variational Autoencoder for glitches. *ArXiv Prepr. ArXiv210710667* (2021).

46. Elmentaite, R., Domínguez Conde, C., Yang, L. & Teichmann, S. A. Single-cell atlases: shared and tissue-specific cell types across human organs. *Nat. Rev. Genet.* **23**, 395–410 (2022).

47. He, K. *et al.* Masked autoencoders are scalable vision learners. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 16000–16009 (2022).

## Supplementary Files

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

- UniGeneXfiguressupply.pdf