

COMPUTATIONAL BIOLOGY

Image-based inference of tumor cell trajectories enables large-scale cancer progression analysis

Yang Liu¹, Ling Cai^{1,2}, Ruichen Rong¹, Shidan Wang¹, Liwei Jia³, Peiran Quan¹, Qin Zhou¹, Guanghua Xiao^{1,2,4*}, Yang Xie^{1,2,4*}

Current approaches to estimating cell trajectories, tumor progression dynamics, and cell population diversity of tumor microenvironment often depend on single-cell RNA sequencing, which is costly and resource intensive. To address this limitation, we developed an artificial intelligence (AI) model that leverages cell morphology features and histological spatial organization to classify tumor cell differentiation status, infer cell dynamic trajectories, and quantify tumor progression from hematoxylin and eosin (H&E)-stained whole-slide images. In three independent lung adenocarcinoma cohorts, our AI-based model accurately predicted cell differential status and provided quantifiable measures of tumor progression that were prognostic of patient survival. Spatial transcriptomic integrative analyses revealed cell components and gene signatures enriched in different cell differentiation statuses. Bulk transcriptomic analyses revealed that fast-progressing tumors exhibit up-regulated cell cycle pathways, while slow-progressing tumors retain characteristics of normal lung epithelium. This cost-effective method enables large-scale analysis of tumor progression dynamics using routinely collected pathology slides and provides insights into intratumor heterogeneity.

INTRODUCTION

Lung cancer has been the deadliest cancer type for several decades. In 2024, an estimated 234,580 cases of lung and bronchus cancer are expected to be diagnosed in the United States, with approximately 340 deaths per day attributed to lung cancer. Despite advances in diagnosis and treatment, the prognosis for patients with lung cancer remains poor, with a 5-year survival rate of less than 25% (1). A key challenge in improving outcomes is tumor heterogeneity, composed of spatially and functionally distinct populations of cells that drive drug resistance, metastatic potential, disease progression, and, ultimately, treatment failure (2, 3). Understanding the dynamic progression of tumors is a cornerstone of advancing cancer research and improving patient outcomes.

Estimating how tumors evolve over time is critical for identifying therapeutic targets, predicting disease trajectory, and personalizing treatment strategies. Cutting-edge single-cell RNA sequencing (RNA-seq) offers powerful solutions for modeling the dynamic process of cell differentiation by ordering cells along a developmental or evolutionary trajectory based on their gene expression profiles (4, 5). This approach has been applied to uncover critical insights into tumor progression and evolution, offering a deeper understanding of how cellular differentiation influences tumor behavior (6–9). However, these methods come with notable limitations: They are costly, require specialized equipment and expertise, and are not widely accessible, especially in resource-limited settings. To address this gap, we leverage histopathological images to investigate the spatial and temporal dynamics through the transition of cell differentiation status, which is a critical aspect of histopathological diagnosis and plays

a central role in monitoring tumor progression and guiding treatment strategies in clinical practice (10). While previous studies have explored predicting cell differentiation status from pathology images across various cancer types (11–14), none have addressed whether pathology images can characterize tumor spatial evolution trajectories or how these trajectories are associated with tumor progression and clinical outcomes.

In this work, we developed a computing framework to quantify tumor progression from the hematoxylin and eosin (H&E) stained whole-slide tumor images (WSIs), which have been widely used in lung cancer diagnosis, providing critical insights into tumor grades, cancer subtype, and the tumor microenvironment (15–17). Recent advancements in artificial intelligence (AI) have opened avenues for extracting cell quantitative information from histopathological images (18, 19). Those AI-based computing approaches can now identify patterns and features previously imperceptible to the human eyes, enabling the analysis of tumor architecture, cellular heterogeneity, and spatial relationships between tumor and stromal components, offering a deeper understanding of lung cancer biology (16, 20–22). Powered by AI, we trained a deep learning model to predict cell differentiation status. Then, the image-derived features encoding cell morphological phenotypes were extracted from the trained model to estimate cell pseudotime. Last, we derived an objective tumor progression score by integrating cell differentiation status, cell pathological spatial organization, and pseudotime. We validated the proposed method's capability through patient stratification using three independent lung adenocarcinoma (ADC) cohorts: The Cancer Genome Atlas Lung Adenocarcinoma (TCGA-LUAD), the National Lung Screening Trial (NLST-ADC), and the University of Texas Special Program of Research Excellence (SPORE) in Lung Cancer. The stratification further guided the exploration of related biological pathways and genes associated with tumor progression by integrating matched bulk RNA-seq data with image-derived differentiation metrics in the LUAD and SPORE cohorts, providing plausible explanations for the observed associations between these metrics and clinical outcomes. In addition, we also

Copyright © 2025 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

¹Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA.

²Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA. ³Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA. ⁴Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA.

*Corresponding author. Email: guanghua.xiao@utsouthwestern.edu (G.X.); yang.xie@utsouthwestern.edu (Y.X.)

integrate our methods with spatial transcriptomic data to explore the gene signatures and immune cell components among different cell differentiation statuses (Fig. 1).

RESULTS

Development of the cell differentiation status classification model

Cell differentiation status is a critical aspect of histopathological diagnosis and plays a central role in monitoring tumor progression and guiding treatment strategies in clinical practice (10). Tumor grades reflect the level of cellular differentiation to normal cells and are typically categorized into three levels: low-grade tumors: well-differentiated (G1), intermediate: moderately differentiated (G2), and high-grade tumors: poorly differentiated (G3) and undifferentiated (G4) (23). Low-grade tumors closely resemble native lung tissue, tend to grow and spread slowly, and maintain some degree of structured organization. In contrast, high-grade tumors are more aggressive, exhibiting substantial loss of structured organization and little to no similarity to normal lung tissue (24).

A critical aspect of cell status classification is developing a robust five-class prediction model, encompassing the four established tumor grade categories (G1 to G4) and an additional “other” class that includes normal, necrosis, and other pathological conditions. To achieve this specific goal, a deep learning-based model was constructed using the high-quality, annotated dataset from the NLST (25), which includes 381 WSIs of lung ADC (Fig. 2A). We fine-tuned the Phikon histopathology foundation model (26) using the image patches randomly extracted from the annotated regions of interest

(ROIs) on WSIs (Fig. 2B). Details on the dataset and the model training process are provided in Materials and Methods.

Validation of cell differentiation status prediction accuracy

A total of 440 ROIs were annotated in the NLST dataset and subsequently divided into training ($n = 395$) and testing ($n = 45$) subgroups. The model was trained on 1976 patches randomly extracted from training ROIs, and the model's performance was evaluated using 2250 patches from independent testing ROIs (Fig. 2C). Because of the limited number of G4 tumors in our model, we classify G1 as low grade, G2 as intermediate grade, and G3 and G4 as high-grade tumors following the guidance of the National Cancer Institute. We compared the performance of five pretrained foundation models (Table 1), and the Phikon model outperformed the Phikon-V2 (27) and other convolutional neural network-based models (28). The result shows the patch-level classification weighted accuracy of 0.61 for all classes (Fig. 2D and table S1) of the Phikon model. To evaluate our model's performance on WSIs, we aggregated the grades of all extracted tumor patches and assigned each slide the rounded average grade. On a cohort of 148 WSIs, the model achieved a weighted accuracy of 0.70, with an area under the receiver operating characteristic curve (AUROC) of 0.89 for distinguishing low-grade from intermediate/high-grade tumors and an AUROC of 0.83 for distinguishing low/intermediate-grade from high-grade tumors (Fig. 2, E and F). Most discrepancies arose from the limited number of training samples of undifferentiated tumors (G4) and the misclassification of moderately (G2) and poorly (G3) differentiated tumors. However, the classification of G2 and G3 is inherently challenging due to the subtle boundaries between these two categories. In

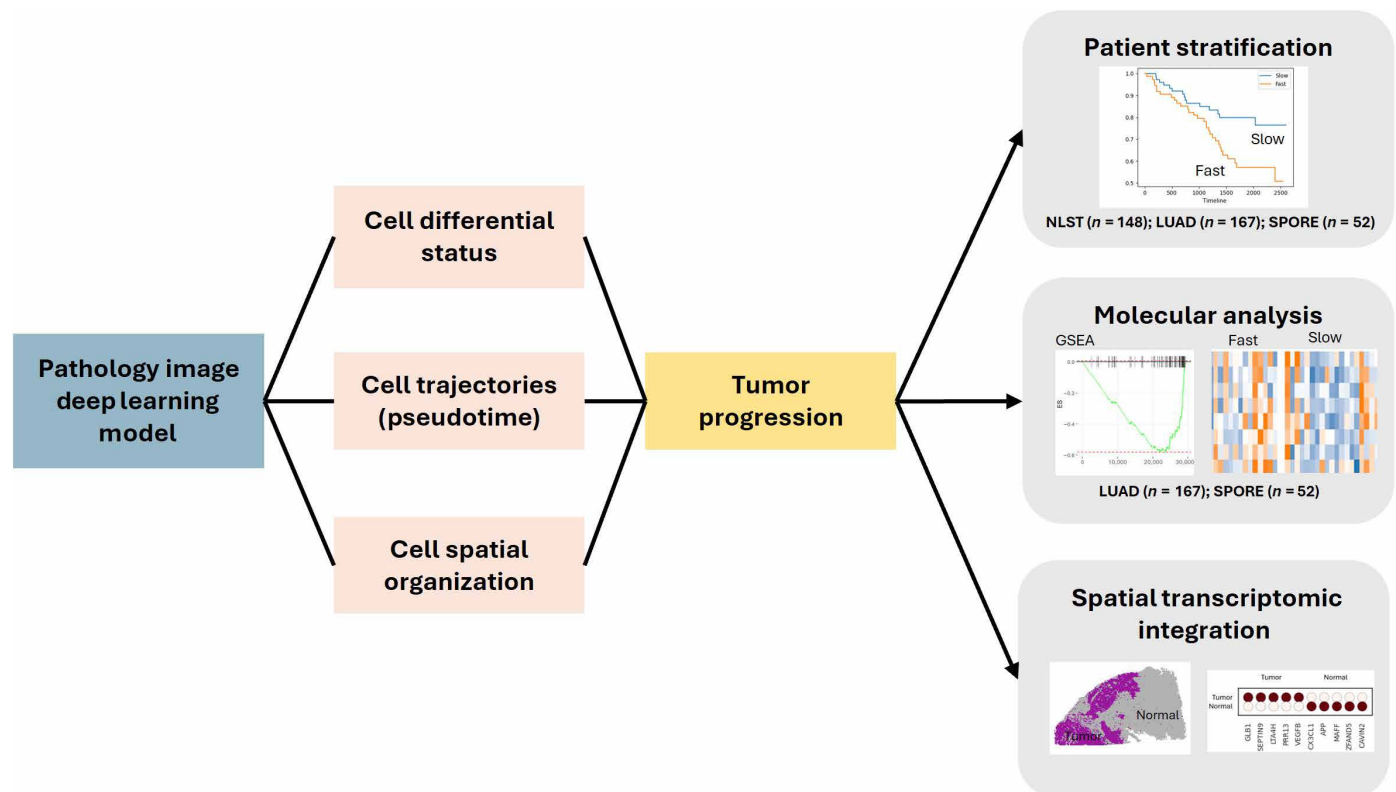


Fig. 1. Overview of tumor cell dynamic framework.

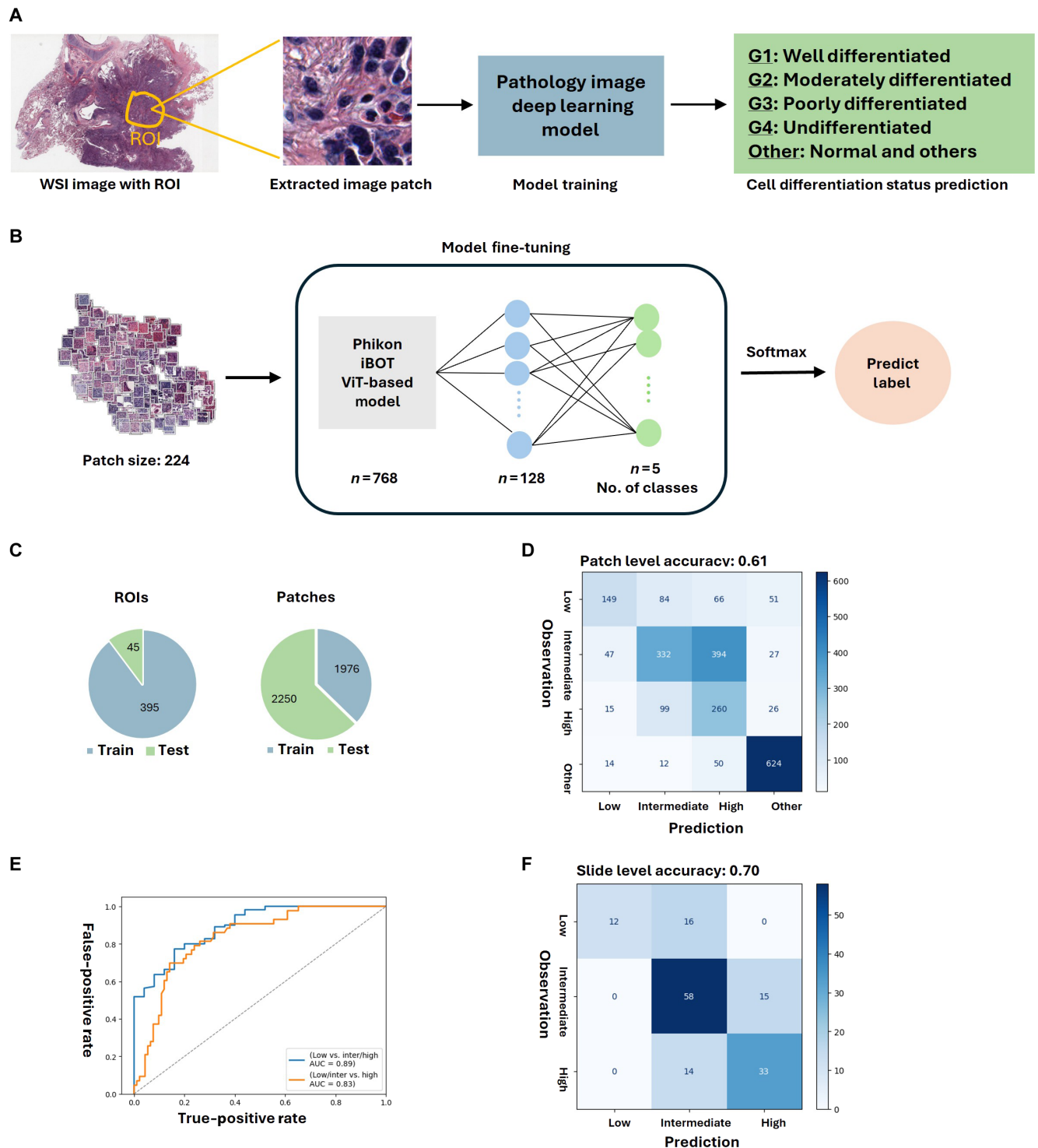


Fig. 2. Illustration of using pathology image deep learning model to predict cell differential status. (A) Flowchart of predicting cell differential status from digital pathology images. Image patches extracted from annotated ROIs of H&E images are used to train a deep learning model for predicting cell differential status. (B) Detailed network structure of the proposed deep learning model. Image patches of size 224×224 serve as the inputs of our fine-tuned model. The model is built on a pretrained Phikon model followed by a feed-forward layer with 128 nodes and a layer of 5 nodes representing the five classes. A softmax function is applied to compute the probabilities across the five classes, determining the predicted class. (C) Summary of the number of training and testing ROIs (left) and patches (right) used in the model. (D) Confusion matrix of the patch-level performance using the independent testing patches of low (G1), intermediate (G2), and high (G3 and G4) grades and other pathological conditions. (E) AUROC of slide-level performance of the NLST dataset. (F) Confusion matrix of the slide-level prediction performance of the NLST dataset.

Table 1. Performance comparison of different deep learning backbone models.				
Model	Patch size	Feature size	Weighted accuracy	Cohen's K
Phikon (26)	224	768	0.61	0.468
Phikon-V2 (27)	224	1024	0.36	0.00
Resnet-50 (28)	500	2048	0.35	0.050
Resnet-101 (28)	500	2048	0.36	0.088
Resnet-152 (28)	500	2048	0.41	0.088

addition, the manual annotation of ROIs may also introduce variability stemming from subjective decisions made by pathologists.

Pseudotime analysis of WSIs using cell differentiation status model

The pretrained patch-level tumor grading model enables application at the WSI level for lung ADC analysis. An example of WSI processing is shown in Fig. 3A. The pretrained model rapidly scans the masked tissue region pixel by pixel, providing spatialized results that display the distribution of differential status across the WSI. During this scanning process, image-derived features were extracted and saved for downstream quantitative analyses. Analogous to gene expression values in single-cell RNA-seq, image-derived features capture cell morphological phenotypes that can be used to estimate pseudotime of image patches. We used the extracted image features for pseudotime analysis to spatially investigate tumor cell differentiation status and their trajectories (Fig. 3A; more details in Materials and Methods). Image-based pseudotime analysis sorts image patches according to cell morphological heterogeneity, enabling the exploration of tumor evolution spatially.

We conducted a detailed analysis of the predicted cell differentiation status and annotated ROIs confirmed by our expert pathologist. Figure 3B illustrates two whole-slide H&E image examples: a well-differentiated (G1) tumor and an undifferentiated (G4) tumor. In both cases, the predicted status accurately matched the annotated ROIs. Pseudotime analysis further revealed a later time point in the tumor regions compared to normal regions, reflecting a more advanced stage of tumor cell differentiation and progression, consistent with the status predictions. In the G4 example, although both the “normal” and “necrosis” regions were classified as other, the pseudotime analysis still effectively distinguished the necrotic region, which exhibited the latest cell stage compared to normal and tumor regions. These findings demonstrate that our model successfully aligns predicted status with annotated ROIs. In addition, the image feature–based pseudotime analysis provides valuable insights into cell trajectories, further enhancing our understanding of tumor progression.

Prognostic image biomarker integrating spatiality and pseudotime of tumor

The fine-tuned cell differentiation status classification model enables objective quantification of tumor progression using an H&E image. To further analyze tumor dynamics, we propose a general fitness score to quantify the level of tumor progression (Fig. 4A). This fitness score is defined as the product of two components: the progression speed (*S*), which reflects the biological aggressiveness of

tumor growth based on pseudotime and the spatial organization of image patches, and the Shannon diversity (*I*), which represents tumor heterogeneity based on cell morphological diversity. Additional details on the mathematical model and calculations can be found in Materials and Methods.

We applied our model to three independent ADC datasets [NLST, University of Texas Southwestern Medical Center (UTSW)-SPORE, and TCGA-LUAD] to examine the association between the derived scores and patient overall survival. Table 2 summarizes the detailed clinical characteristics of patients included in each dataset. Patients were divided into two groups based on the median value of progression speed, diversity index, and fitness score: “slow” ($\leq$ median value) and “fast” ($>$ median value). We performed the log-rank test between two patient stratification groups in the three datasets: NLST (fitness: $P = 0.01$; speed: $P = 0.04$; diversity: $P = 0.05$), LUAD (fitness: $P = 0.02$; speed: $P = 0.02$; diversity: $P = 0.05$), and SPORE (fitness: $P = 0.03$; speed: $P = 0.04$; diversity: $P = 0.08$) (Fig. 4, B to J). The slow progression group shows better survival than the fast progression group in all three datasets. Both fitness and speed scores showed substantial success in stratifying patients in all three datasets, while the tumor Shannon diversity is significant in both the NLST and LUAD datasets.

In the univariate overall survival analysis using the Cox proportional hazard (Cox-PH) regression model, the fitness score is a significant independent predictor of hazard risk in the NLST ($P = 0.05$) and SPORE ($P = 0.01$) datasets, and speed is also substantial for the SPORE ($P = 0.02$) datasets (Fig. 4, B to J). Integrating the image-derived scores with clinical risk factors (e.g., cancer stage, gender, and age) in the multivariate Cox-PH analysis also improved the risk prediction results of higher concordance index (C-index) in both the NLST (from 0.67 to 0.7) and SPORE (from 0.72 to 0.82) datasets. In the LUAD dataset, the C-index remained at 0.62 with no improvement after adding image features. Detailed outputs of the multivariate Cox-PH analysis are provided in tables S2 to S4.

We then investigated the correlation between progression speed and Shannon diversity across three datasets. A weak positive correlation was observed, suggesting that aggressive tumor progression is potentially associated with increased and more evenly distributed tumor clonotypes, as the Shannon index reflects tumor morphological heterogeneity calculated from image features (Fig. 4, K to M). Using only speed and diversity scores, patients were further stratified into the three progression groups of extremely slow ($\leq$ median of both speed and Shannon index), extremely fast ($>$ median of both speed and Shannon index), and moderate (all other cases). This combined stratification approach

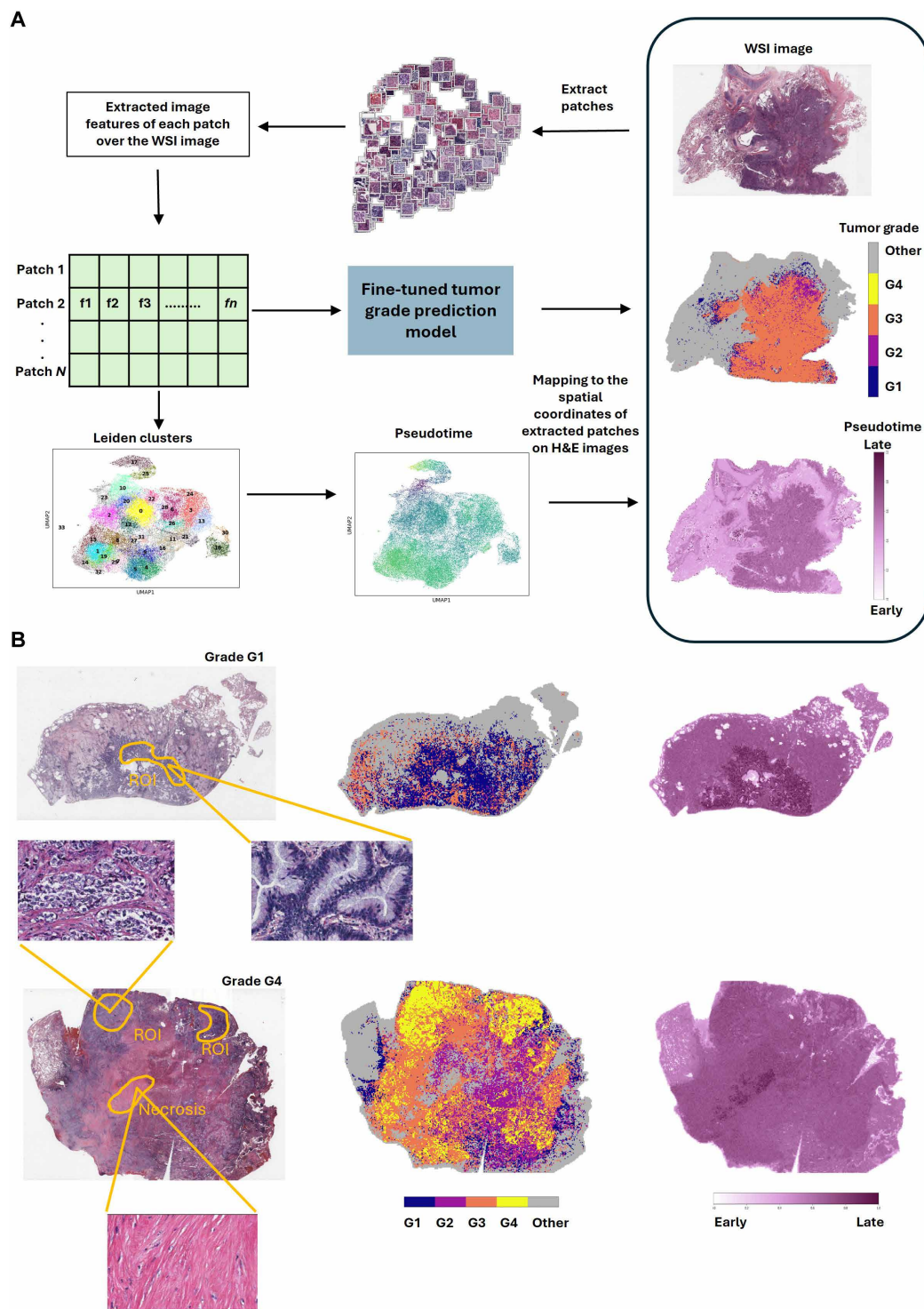


Fig. 3. Cell differential status and pseudotime on whole-slide H&E images. (A) Flowchart of cell differential status prediction and pseudotime inference at the WSI level. To predict WSI-level cell differential, the trained deep learning model scans each image patch extracted across the tissue region. At the same time, patch features used for prediction are extracted for the downstream pseudotime analysis. Patches were clustered by the Leiden algorithm using the extracted features, and pseudo time was estimated according to the Leiden clusters using the diffusion pseudotime algorithm, more details in Materials and Methods. UMAP, uniform manifold approximation and projection. **(B)** Results of G1 tumor (top) and G4 tumor (bottom). Both G1 and G4 identified the tumor ROI region, and the pseudotime results of G4 identified the necrosis region.

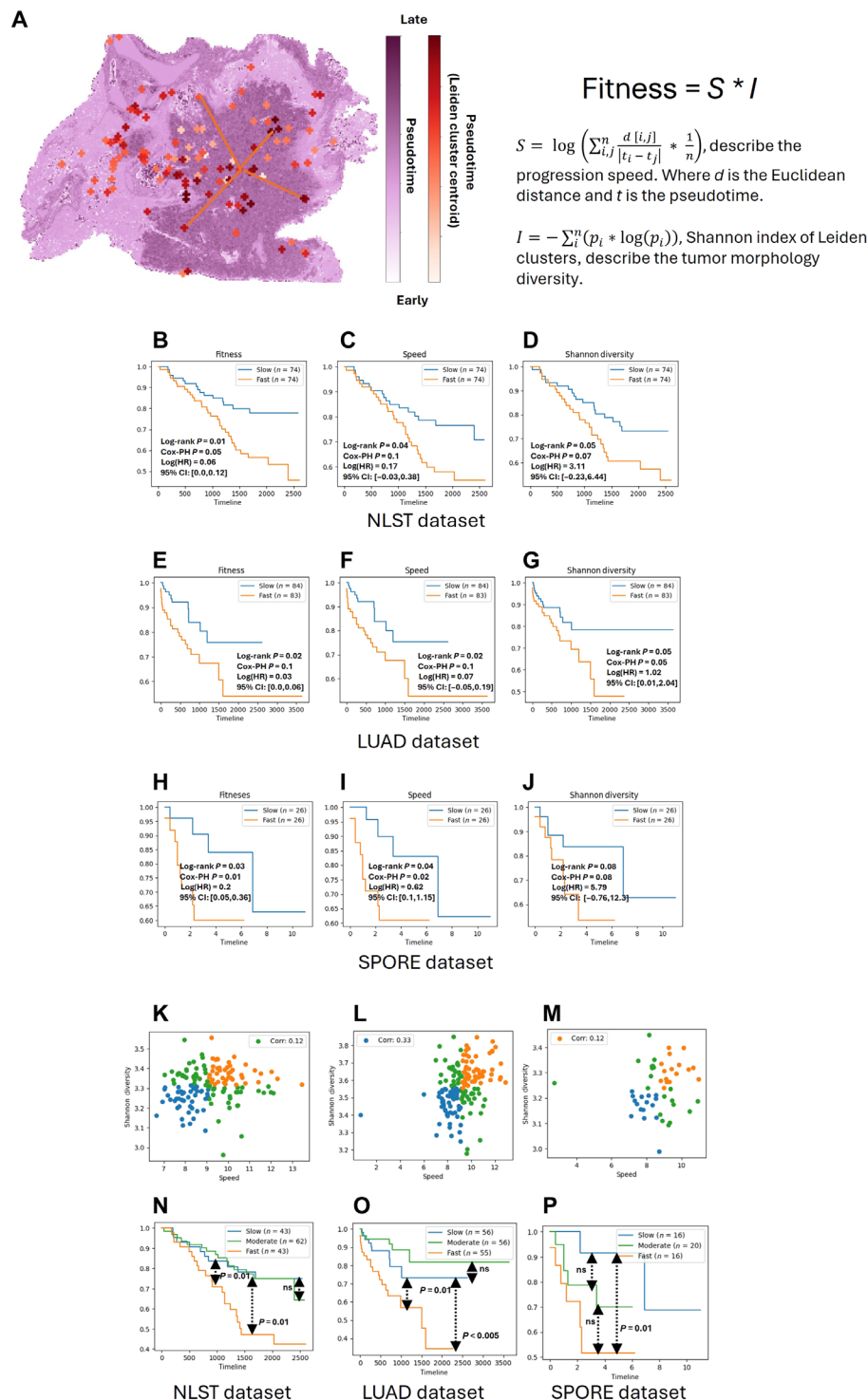


Fig. 4. Tumor progression fitness model on WSI images and survival analysis. (A) Illustration of tumor progression quantification integrating spatial organization on WSI; more details in Materials and Methods. (B to J) Results of Kaplan-Meier log-rank test and univariate Cox-PH regression analysis of NLST, SPORE, and LUAD datasets. Patients were stratified into slow (blue) and fast (orange) groups. HR, hazard ratio; CI, confidence interval. (K to M) Correlation plot between the tumor progression speed and Shannon index of three datasets. Patients were stratified into slow (blue), “moderate” (green), and fast (orange) groups by the median value of both speed and Shannon index. (N to P) Results of Kaplan-Meier curve of three-group patient stratification on three datasets. ns, not significant.

Table 2. Clinical characteristics of patients in NLST, SPORE, and LUAD datasets.			
	NLST	SPORE	LUAD
Total patients	N = 148	N = 52	N = 167
Stage			
I	103 (69.5%)	32 (61.5%)	92 (55.0%)
II	18 (12.1%)	8 (15.3%)	44 (26.3%)
III	20 (13.5%)	11 (21.1%)	21 (12.5%)
IV	7 (4.7%)	1 (1.9%)	10 (5.9%)
Gender			
Male	77 (52%)	20 (38.4%)	75 (45.0%)
Female	71 (48%)	32 (61.5%)	92 (55.0%)
Smoke history			
Yes	80 (54.1%)	47 (90.3%)	–
No	68 (45.9%)	5 (9.4%)	–
Age			
≤59	32 (21.6%)	–	47 (28.1%)
>60	107 (78.4%)	–	120 (71.8%)
Race			
Caucasian	–	44 (84.6%)	134 (80.2%)
African American	–	5 (9.6%)	22 (13.1%)
Asian	–	3 (5.8%)	7 (4.2%)
Not reported	–	–	4 (2.3%)

further improved patient differentiation into slow, moderate, and fast progression groups across the three datasets (Fig. 4, N to P).

Robustness of patient stratification using tumor progression quantification

To further validate our framework, we repeated our patient stratification analysis using Louvain clustering results (fig. S1) (29). In the NLST cohort, all three stratification metrics derived from Louvain clusters robustly distinguished patient groups, and in the SPORE dataset, Shannon diversity similarly stratified clinical outcomes. Overall, greater tumor progression remained consistently associated with poorer survival in the Louvain analysis.

We then evaluated how the number of principal components (PCs) influences pseudotime and Shannon diversity metrics for patient stratification. Across PC counts of 10, 20, 30, 40, and 50, all three metrics exhibited positive correlations. Moreover, slower tumor progression quantification consistently predicted better survival (Fig. 4, table S5, and figs. S2 to S4). Notably, the fitness score proved more robust than either speed or Shannon diversity across all three datasets, showing minimal sensitivity to PC count (see Materials and Methods for details).

Molecular function of genes associated with tumor progression

We downloaded the match transcriptomic data of the SPORE and LUAD datasets and performed gene set enrichment analysis (GSEA) using genes ranked by Spearman correlation between gene expression and Shannon diversity, leveraging pathway annotations from REACTOME pathway database (30), cell type signature (C8) gene sets from the Molecular Signature Database (MSigDB) pathway

database (31). In the fast progression group, up-regulated pathways included those related to cell cycle regulation, such as the synthesis phase, DNA replication (Fig. 5A, SPORE dataset), and G₂M checkpoint (fig. S5, LUAD dataset). These findings indicate increased cell division and proliferation in these tumors, corresponding to the classification of the fast-progress group. In contrast, the slow progression group showed higher expression levels of genes involved in surfactant metabolism, reflecting a more differentiated cell state characteristic of normal lung epithelial cells, which aligns with their classification as slow-progress tumors (Fig. 5A and fig. S5A). Supporting this notion, genes associated with slow progression were also significantly enriched in alveolar type 2 cell (AT2) signatures (Fig. 5B and fig. S5B) (28), a known cell type responsible for maintaining the gas exchange function of the lung, surfactant production, alveolar repair and regeneration, and innate immune regulation (32–34). The GSEA analysis also identified important genes, including members of the surfactant protein family (*SFTPA*, *SFTPB*, and *SFTPC*) and *CD81* expressed on the surface of B cells (35), involved in antitumor activity. Analysis using both SPORE and TCGA datasets revealed that *CXCL17* reported to exhibit anti-inflammatory activity in macrophages (36) and a prognostic marker in ADC (37). Although *CXCL17* is commonly expressed in the lung, it has reported both protumor and antitumor activities (38, 39), its underlying mechanisms remain poorly understood.

To further investigate this relationship, we calculated an AT2 score as the average expression of leading-edge genes from the GSEA results. A negative correlation between the AT2 score and derived speed, diversity, and fitness score was observed in both datasets. Notably, slow progression patients exhibited higher expression of AT2 cell type gene signatures compared to fast progression patients

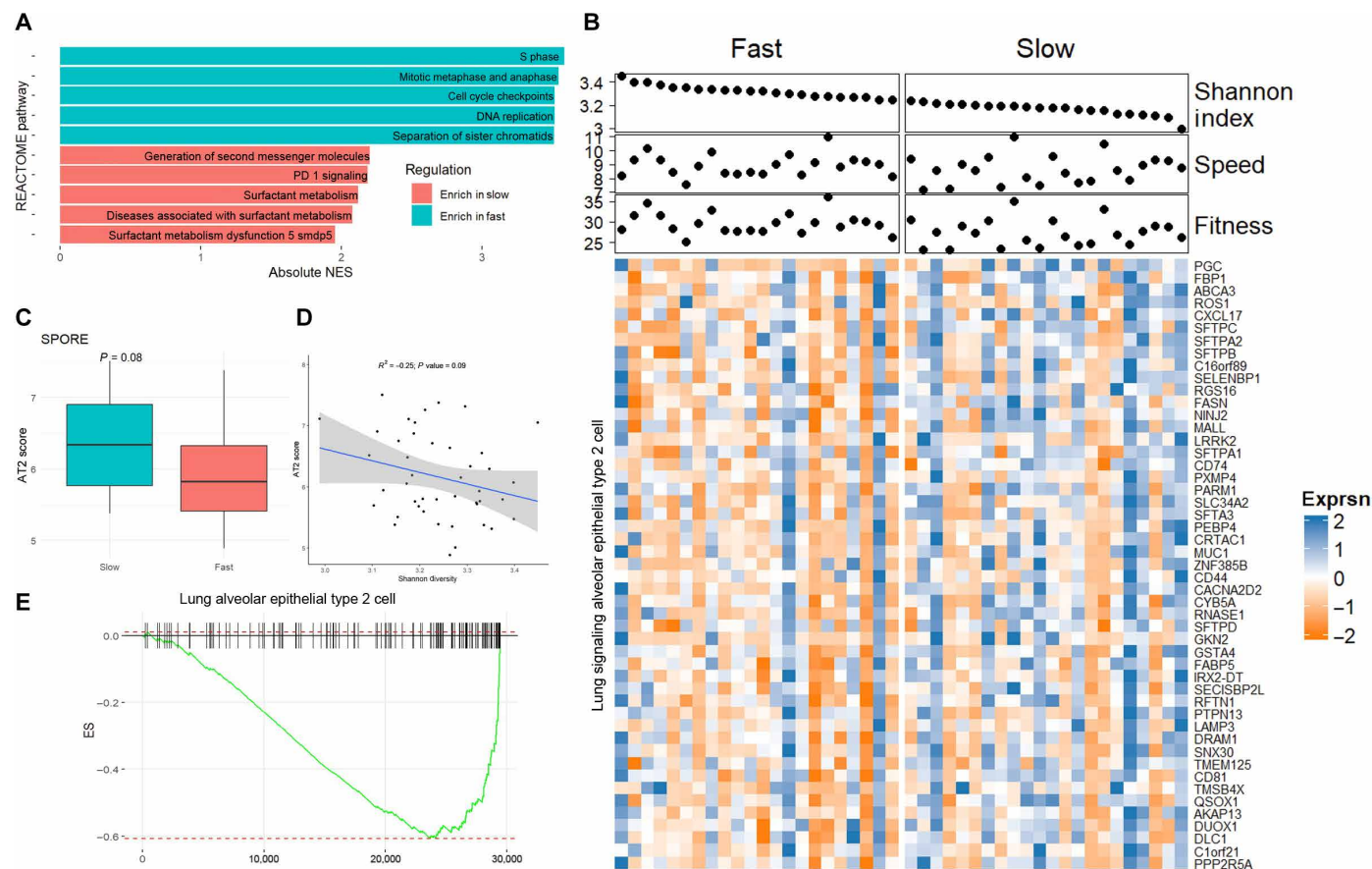


Fig. 5. Transcriptomic analysis of the patient stratification based on the tumor progression quantification. (A) GSEA analysis results of the REACTOME pathway database using genes ranked by the correlation between gene expression and Shannon index from the LUAD dataset. NES, normalized enrichment score. (B) Normalized gene expression profile of genes enriched in the AT2 cell gene signatures between the slow progression and fast progression patients of the LUAD dataset. The AT2 gene signatures are acquired from the MSigDB. (C) Wilcoxon rank sum test results of AT2 scores between the slow and fast groups. (D) Spearman correlation of AT2 score and Shannon index. R^2 , coefficient of determination. (E) Genes negatively correlated with Shannon diversity index were significantly enriched in the AT2 cell type signatures.

(Fig. 5, C and D, and fig. S5, C and D). This suggests an enhanced normal cell function, and it may also indicate that the tumor cells retain features of AT2 cells and are well differentiated (low-grade tumor), which usually has a slower growth rate and less aggressive behavior than poorly differentiated (high-grade tumor).

Integrating cell differentiation model with spatial transcriptomic data

To assess our model on spatial transcriptomic data, we applied it to the Xenium 5K ADC dataset downloaded from 10X Genomics, predicting cell differentiation status across WSI (Fig. 6A). Patch-level gene expression was calculated by averaging spot-level profiles, and highly variable genes were identified between differentiation states. The top five markers for G2 tumor regions (*GLB1*, *SEPTIN9*, *LTA4H*, *PRR13*, and *VEGFB*) and normal regions (*CX3CL1*, *APP*, *MAFF*, *ZFAND5*, and *CAVIN2*) are highlighted in Fig. 6B. Notably, *LTA4H* is up-regulated in tumor versus normal tissue, reflecting its roles in inflammatory signaling and cell proliferation (40, 41). Both *PRR13* and *VEGFB* correlate with poor survival in ADC patients (42, 43), and *VEGFB* further drives metastasis, making it a potential therapeutic target (43–45). In contrast, *CX3CL1*—a chemokine with context-dependent pro- and antitumor effects—may regulate natural killer

cell infiltration (46), while *MAFF* acts as a tumor suppressor by inhibiting ADC cell proliferation and blocking cell cycle progression (47).

We next applied Leiden clustering to the spatial transcriptomic expressions and annotated cell type based on its highly variable genes, enabling us to assess immune-cell composition across cell differential statuses (Fig. 6, C to E). We then quantify the cell components of each slide and compare the compositions between predicted grades (Fig. 6, F and G, and fig. S6). As anticipated, tumor regions were dominated by epithelial tumor cells (Fig. 6F). These regions also recruit an increased infiltration of memory T cells, identified by *CD44* and *CXCR4* expression, relative to adjacent normal tissue (Fig. 6, E to G). Conversely, vascular endothelial cells and alveolar fibroblasts—marked by *DHCR24*, *SDC4*, *TSPAN3*, *ATP1A1*, and *TPPP3*—were significantly enriched in normal regions (Fig. 6, H and I), which tends to correlate with better overall survival (48).

DISCUSSION

In this study, we developed an AI model that leverages the potential of cost-effective H&E images to analyze cell differentiation status, trace cell trajectories, and assess tumor progression. Overcoming the limitation of expensive, bad data quality and limited accessibility

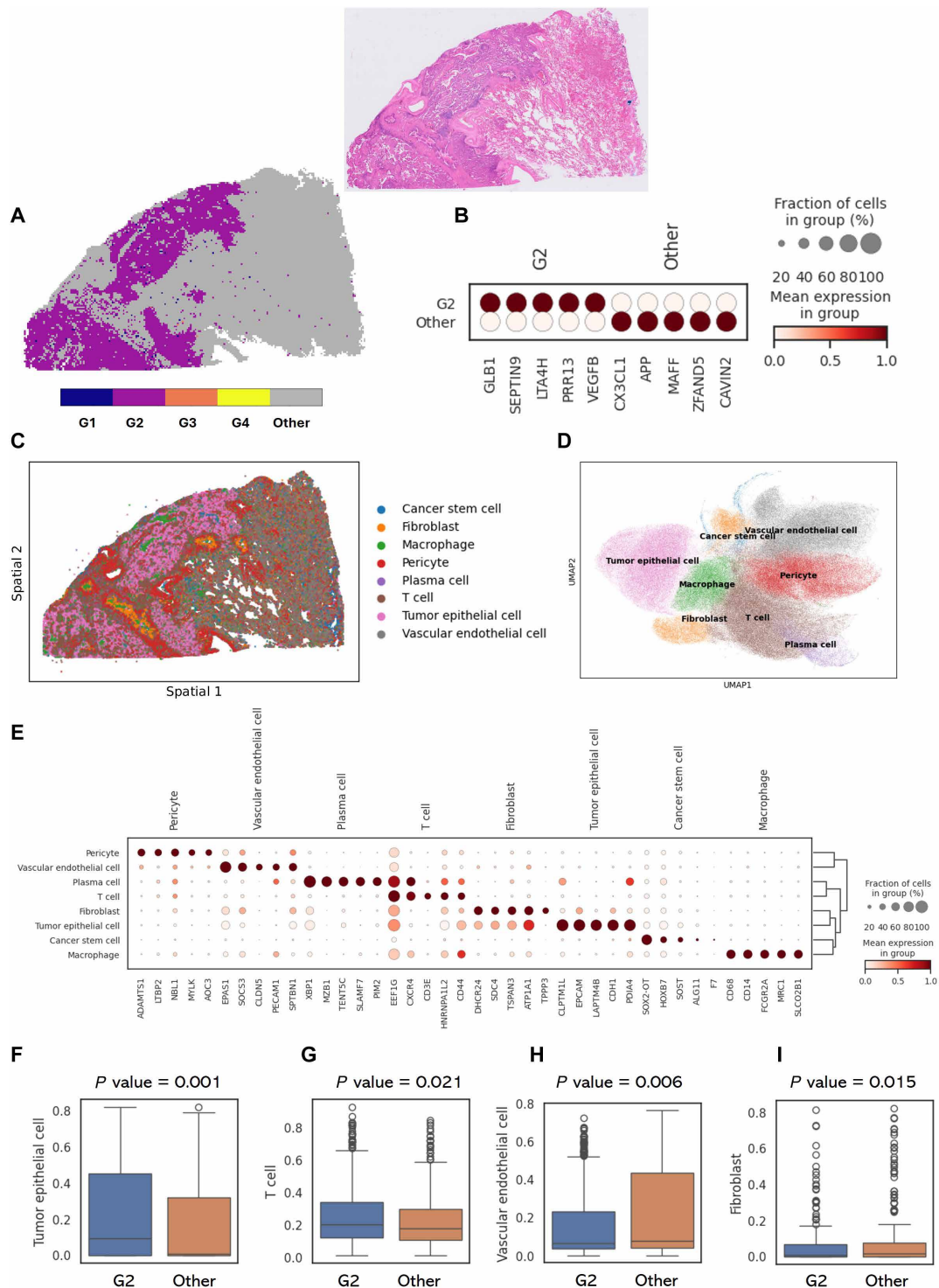


Fig. 6. Integrative analysis with spatial transcriptomic data of ADC. (A) Results of the cell differentiation status prediction from the trained model. (B) Highly variable genes differentiate between the predicted G1 and the other normal region. (C) Cell type annotation of the Xenium 5K ADC dataset. (D) UMAP of the identified cell clusters. (E) Highly variable genes in each annotated cell cluster. (F to I) Comparison of cell components from image patches predicted as G2 tumor or other normal regions.

of single-cell RNA-seq, our proposed approach aids in patient stratification and provides guidance for clinical cancer treatment strategies in lung ADC,

To class the cell differentiation statuses, we fine-tuned the Phikon (26) foundation model that performs better than Resnet models in this task since it was specifically developed for pathological image analysis and was pretrained on H&E images from TCGA, while Resnet (28) models were pretrained from the natural ImageNet (49) dataset. Our framework effectively captures the morphological features of cells, enabling accurate performance at both the patch and WSI levels. When applied to WSIs, our model successfully identified ROIs with validation from expert pathologists. Therefore, our AI model offers an efficient tool to assist pathologists in manually examining H&E slides and ROI annotations in clinical settings.

In addition, the image-derived features demonstrate the ability to infer pseudotime, effectively depicting cell differentiation trajectories across the WSI. Moreover, although all other pathological conditions were labeled as other, image-based pseudotime analysis can differentiate necrotic regions from normal and tumor cells as the trained model can capture the morphological difference among different cell types. This highlights the model's capacity to extract nuanced biological insights beyond standard classifications.

By integrating spatial coordinates and clustering results of image patches, we developed three metrics to capture tumor progression status: (i) speed, which quantifies the spatial tumor expansion velocity within a given area; (ii) the index of tumor diversity, reflecting tumor morphological heterogeneity; and (iii) fitness, a combined measure of speed and diversity index. We validated these metrics in three independent ADC cohorts—NLST, SPORE, and LUAD—with promising results. First, we demonstrated the patient stratification capability of these metrics. Predicted slow progression patients, characterized by low speed, low diversity, and low fitness, exhibited significantly better survival rates than fast progression patients, as confirmed by log-rank tests. Furthermore, incorporating these image-derived features into clinical phenotypes improved survival risk prediction, highlighting their value in prognostic modeling.

To further elucidate the biological basis of tumor progression, we conducted GSEA using transcriptomic data from the SPORE and LUAD cohorts. Fast progression patients displayed up-regulated pathways associated with active cell cycle processes, such as increased proliferation and reduced surfactant metabolism, contributing to more aggressive tumor behavior. In contrast, slow progression patients exhibited high expression of AT2 cell gene signatures, indicating that their tumors retained more normal-like and well-differentiated characteristics. In addition, the integrative analysis with spatial transcriptomics identified both meaningful genes and cell types that spatially vary between tumor and normal regions. Overall, these findings were consistent with our predicted cell differentiation status and stratified patient groups, providing robust support for the biological relevance of the proposed metrics in understanding tumor progression and stratifying patients by their clinical outcomes.

There are limitations to our study. First, our study has a relatively small number of verified ROIs in our training dataset, which may affect the prediction efficacy. However, the impact on downstream analyses is minimal, as pseudotime calculations rely more on the extracted image features than predicted status. Expanding the high-quality dataset to include a larger variety of tumor-grade slides and other pathological conditions would further enhance the model's

performance and generalizability. Second, given that cell differential status reflects the degree of similarity between tumor and normal cells, our proposed tumor progression quantification incorporates image features from both tumor and nontumor regions. Incorporating cell type segmentation could further refine the model by isolating tumor-specific features and focusing on interactions among tumor, stromal, and immune cells.

In conclusion, we developed an innovative AI-assisted model for cell differentiation status classification, tracing cell trajectories, and assessing tumor progression. This study uses only image-derived features to estimate pseudotime and tumor progression from H&E slides. We demonstrated the clinical applicability of our model across three independent lung ADC cohorts, showcasing its robustness and utility. By enabling accurate and efficient processing of histopathological images, our model has the potential to advance cancer research and improve clinical cancer treatment practices.

MATERIALS AND METHODS

Dataset

NLST-ADC dataset

The lung ADC pathology image datasets used to support cell differentiation status classification were obtained from the NLST-ADC. The publicly available dataset from the NLST (<https://biometry.nci.nih.gov/cdas/nlst/>) comprises 318 H&E-stained pathology images ($\times 40$ magnification) along with the clinical data from 148 patients. Multiple pathology slides may correspond to a single patient. To enhance our analysis, a total of 440 ROIs encompassing tumor, normal, necrosis, and other pathological conditions were annotated and confirmed by pathologists using the I-viewer software (50).

TCGA-LUAD dataset

Clinical data are downloaded from the TCGA data portal. Gene-normalized RNA-seq data from TCGA LUAD, processed by RSEM (RNA-seq by expectation maximization) algorithm was downloaded from Firebrowse (<http://firebrowse.org/>) (doi:10.7908/C19P30S6). All the imaging data are available at National Cancer Institute Imaging Data Commons (<https://datacommons.cancer.gov/repository/imaging-data-commons>). More information can be found at the TCGA website (<https://cancer.gov/ccg/research/genome-sequencing/tcga>). All participating sites obtained local institutional review board approval for participation in this study. All samples in TCGA have been collected and used following strict human subject protection guidelines and informed protocols.

UTSW-SPORE dataset

The SPORE in the Lung Cancer dataset includes 111 patients with ADC (51). Tissue samples were obtained via surgical resection from patients who provided written informed consent. All tumor tissue slides were formalin-fixed and paraffin-embedded and scanned at either $\times 20$ or $\times 40$ magnifications. Microarray data for this dataset were retrieved from the Gene Expression Omnibus (GSE42127) (52). The Illumina probe IDs from this dataset were converted to gene symbols using the “illuminaHumanv3.db (1.26.0)” R packages.

Cell differentiation status classification from histopathology images

We integrated the Phikon model (26), followed by sequential feed-forward layers with 128 and 5 nodes (number of classes), to predict patch-level cell differentiation statuses (Fig. 2B). The Phikon model

is a foundation model pretrained on 40 million TCGA histopathology patches. To tailor it for our specific task, we fine-tuned the model using tumor grade-specific patches extracted from annotated ROIs. The dataset was split into 395 ROIs for training and 45 ROIs for testing. In addition to the four traditional grade (G1 to G4) classifications, we included other cell statuses as a fifth class that covers normal and necrosis pathological conditions. Patches were randomly extracted from ROIs at a resolution of 224×224 pixels to match the input size requirements of the Phikon and Phikon-V2 models. For the final dataset, 5 patches were extracted from each training ROI, and 50 patches were extracted from each testing ROI (Fig. 2C). When testing the performance of Resnet models (28), patches were extracted at 500×500 . The final model was trained for 100 epochs with a batch size of 20. Cross-entropy loss was applied as the objective function, which was minimized by the Adam optimizer with a step-wise learning rate of 0.001. To mitigate overfitting from the imbalanced ROI distribution across cell differential status, we applied cross-entropy weights derived from the inverse class frequencies in our training set: 0.0607, 0.0184, 0.0293, 0.8695, and 0.0221 for G1, G2, G3, G4, and other, respectively. We selected this weighted model after comparison with unweighted versions, which failed to predict any G4 tumors and showed lower overall accuracy (fig. S7). The “scikit-image” library (v0.22.0) in Python was used for patch extraction, while “torch” (v2.2.1) and “torchvision” (v0.17.1) were used for model training. The Phikon model was implemented using the “ViTModel” from the “transformers” package (v4.38.2) in Python.

Extracting WSI image features from histopathology images

After training the patch-level prediction model, we applied it to scan WSIs, processing all individual patches extracted at 400-pixel intervals. For each patch, 768 Phikon features were extracted from the fine-tuned model to predict the cell differentiation status of the specific patch region. These features were subsequently used to infer pseudotime at the WSI level (Fig. 3A). Details and scripts of Phikon feature extraction are available at <https://huggingface.co/owkin/phikon>.

Pseudotime inference from extracted WSI histopathology image features

Similar to pseudotime inference from gene expression of single-cell RNA-seq, we inferred histopathology pseudotime using image patch features extracted from our fine-tuned model. First, the feature dimensions were reduced using principal components analysis. We computed the nearest-neighbor distance matrix using four neighbors and the first 20 PCs, capturing $\geq 63\%$ of the variance across all H&E images. Patch clustering was then performed at a resolution of 1 using the Leiden algorithm, a computationally efficient method that ensures all patches are well-connected for optimal downstream trajectory inference (29). To construct a topological graph based on patch partitions, we used partition-based graph abstraction (53). A randomly selected normal patch was designated as the trajectory root node. Subsequently, the diffusion pseudotime (54) method was applied to estimate pseudotime values, capturing the dynamic tumor cell progression in histological space. All analyses were conducted using the “Scanpy” Python package (v1.9.8) (55).

Integrating spatial data and pseudotime for tumor progression quantification

The quantification of tumor progression describes the tumor expansion velocity within an area, integrating pseudotime and spatial

organization over the WSI. The tumor progression fitness score is calculated as

$$\text{Fitness} = S * I \quad (1)$$

$$S = \log \left(\sum_{i,j}^n \frac{d[i,j]}{|t_i - t_j|} * \frac{1}{n} \right) \quad (2)$$

$$I = - \sum_i^n [p_i * \log(p_i)] \quad (3)$$

where S represents the average pairwise tumor progression speed among Leiden clusters, t denotes the mean pseudotime of each cluster, and d is the Euclidean distance between the centroids of clusters i and j . Hereby, a smaller S value indicates slower tumor progression, characterized by a limited spatial tumor expansion area on H&E images within a unit of pseudotime. In addition, a larger S value represents more aggressive cell growth and expansion activity. I is the Shannon index calculated on the basis of the Leiden clusters, reflecting the tumor clonal diversity inferred from the cell morphology heterogeneity. A low Shannon index indicates a high dominance of certain pathological cell clonotypes, corresponding to a relatively purer tumor microenvironment, suggesting a slow progression with relatively less active cell status transition or cell growth activity. In contrast, a higher Shannon index indicates more unique clonotypes and active cell activities. Together, the fitness score integrates tumor progression speed (S) and clonal diversity (I) to provide a comprehensive measure of tumor dynamics in H&E images, capturing both diversity and the presence of homogeneous cell components.

Patient selection and survival analysis

We conducted survival analysis on the three lung ADC cohorts to investigate the association between our tumor progression quantifications and clinical outcomes. H&E images with less than 10,000 patches extracted from the TCGA and NLST datasets and less than 3500 patches extracted from the SPOR dataset were excluded from the cohort analysis. Patients were stratified into two groups based on the median values of the fitness score, speed (S), or Shannon diversity (I). Patients with values less than or equal to the median were classified into the slow tumor progression group. In contrast, those with values above the median were placed in the fast progression group. The Kaplan-Meier (KM) curve and log-rank test were used to evaluate the survival difference between slow and fast groups. In addition, the Cox-PH model was applied to perform univariate and multivariate survival regression analyses. The C-index is used to measure the prognostic ability of survival models. All analysis was conducted using the Python package “lifelines” (v0.28.0).

Gene set enrichment analysis

The hallmark gene set (H) and cell type signature gene sets (C8) from the REACTOME pathway database (30) and MSigDB (31) were used for GSEA analysis. Genes were ranked on the basis of their Spearman correlation between RNA expression and the Shannon index. Correlation is calculated using the “spicy.stats” Python packages. GSEA was performed using the “fgsea” R package (v1.28.0).

Spatial transcriptomic analysis

The “Squidpy” (v1.6.5) (56) and Scanpy (v19.8) (55) Python packages were used to process the Xenium ADC 5K dataset. First, we computed a nearest-neighbor distance matrix using the four closest neighbors in the space defined by the top 20 PCs. Cells were then clustered via the Leiden algorithm at a resolution of 0.1, and highly variable genes were identified by Wilcoxon rank sum testing across clusters. Last, cell type annotation was performed by applying the GPTCelltype prompt to the marker genes discovered for each cluster (57). The exact prompt used is given below.

“Identify cell types of human lung tumor tissue using the following markers. Identify one cell type for each row. Only provide the cell type name. \nMarkerGeneList”

Statistical analysis

The gene signature score of AT2 cell types (AT2 score) is the average expression of leading-edge genes from the GSEA results. The Kruskal-Wallis test is used to compare gene signature scores between patient stratification groups. All statistical analyses are implemented using the “ggplot2 (3.5.0)” R package.

Supplementary Materials

This PDF file includes:

Figs. S1 to S7

Tables S1 to S5

REFERENCES AND NOTES

- R. L. Siegel, A. N. Giaquinto, A. Jemal, Cancer statistics, 2024. *CA Cancer J. Clin.* **74**, 12–49 (2024).
- I. Dagogo-Jack, A. T. Shaw, Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* **15**, 81–94 (2018).
- C. E. Meacham, S. J. Morrison, Tumour heterogeneity and cancer cell plasticity. *Nature* **501**, 328–337 (2013).
- C. Trapnell, Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
- W. Saelens, R. Cannoodt, H. Todorov, Y. Saeys, A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
- S. C. Bendall, K. L. Davis, E. A. D. Amir, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan, D. Pe'er, Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725 (2014).
- C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, J. L. Rinn, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
- A. Crinier, P. Y. Dumas, B. Escalière, C. Piperoglou, L. Gil, A. Villacreses, F. Vély, Z. Ivanovic, P. Milpied, É. Narni-Mancinelli, É. Vivier, Single-cell profiling reveals the trajectories of natural killer cell differentiation in bone marrow and a stress signature induced by acute myeloid leukemia. *Cell. Mol. Immunol.* **18**, 1290–1304 (2021).
- Z. Liu, H. Lou, K. Xie, H. Wang, N. Chen, O. M. Aparicio, M. Q. Zhang, R. Jiang, T. Chen, Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. *Nat. Commun.* **8**, 22 (2017).
- A. Pérez-González, K. Bévart, C. Blanpain, Cancer cell plasticity during tumor progression, metastasis and response to therapy. *Nat. Cancer* **4**, 1063–1082 (2023).
- R. Jaroensri, E. Wulczyn, N. Hegde, T. Brown, I. Flament-Auvigne, F. Tan, Y. Cai, K. Nagpal, E. A. Rakha, D. J. Dabbs, N. Olson, J. H. Wren, E. E. Thompson, E. Seetao, C. Robinson, M. Miao, V. Beckers, G. S. Corrado, L. H. Peng, C. H. Mermel, Y. Liu, D. F. Steiner, P. H. C. Chen, Deep learning models for histologic grading of breast cancer and association with disease prognosis. *NPJ Breast Cancer* **8**, 113 (2022).
- G. Campanella, M. G. Hanna, L. Geneslaw, A. Mirafior, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, L. F. Fuchs, Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
- S. C. Wetstein, V. M. de Jong, N. Stathonikos, M. Opdam, G. M. H. E. Dackus, J. P. W. Pluim, P. J. van Diest, M. Veta, Deep learning-based breast cancer grading and survival analysis on whole-slide histopathology images. *Sci. Rep.* **12**, 15102 (2022).
- X. Pan, K. AbdulJabbar, J. Coelho-Lima, A. I. Grapa, H. Zhang, A. H. Cheung, J. Baena, T. Karasaki, C. R. Wilson, M. Sereno, S. Veeriah, S. J. Aitken, A. Hackshaw, A. G. Nicholson, M. Jamal-Hanjani, TRACERx Consortium, J. le Quesne, S. M. Janes, A. M. Hacker, A. Sharp, S. Smith, H. K. Dhanda, K. Chan, C. Pilotti, R. Leslie, D. Chuter, M. MacKenzie, S. Chee, A. Alzetani, E. Lim, P. de Sousa, S. Jordan, A. Rice, H. Raubenheimer, H. Bhayani, L. Ambrose, A. Devaraj, H. Chavan, S. Begum, S. I. Buder, D. Kaniy, M. Malima, S. Booth, N. Fernandes, P. Shah, C. Proli, M. Hewish, S. Danson, M. J. Shackcloth, L. Robinson, P. Russell, K. G. Blyth, A. Kidd, A. Kirk, M. Asif, R. Bilancia, N. Kostoulas, M. Thomas, C. Dick, J. F. Lester, A. Bajaj, A. Nakas, A. Sodha-Ramdeen, M. Tufail, M. Scotland, R. Boyles, S. Rathinam, D. A. Fennell, C. Wilson, D. Marrone, S. Dulloo, G. Matharu, J. A. Shaw, J. Riley, L. Primrose, E. Boleti, H. Cheyne, M. Khalil, S. Richardson, T. Cruickshank, G. Price, K. M. Kerr, S. Benafif, D. Papadatos-Pastos, J. Wilson, T. Ahmad, J. French, K. Gilbert, B. Naidu, A. J. Patel, A. Osman, C. Lacson, G. Langman, H. Shackleford, M. Djearaman, S. Kadiri, G. Middleton, A. Leek, J. D. Hodgkinson, N. Totten, A. Montero, E. Smith, E. Fontaine, F. Granato, J. Novasio, K. Rammoohan, L. Joseph, P. Bishop, R. Shah, S. Moss, V. Joshi, P. Crosbie, A. Paiva-Correia, A. Chaturvedi, L. Priest, P. Oliveira, F. Gomes, K. Brown, M. Carter, C. R. Lindsay, F. H. Blackhall, M. G. Krebs, Y. Summers, A. Clipson, J. Tugwood, A. Kerr, D. G. Rothwell, C. Dive, H. J. W. L. Aerts, R. F. Schwarz, T. L. Kaufmann, P. van Loo, G. A. Wilson, R. Rosenthal, A. Rowan, C. Bailey, C. Lee, E. Colliver, K. S. S. Enfield, M. S. Hill, M. Angelova, O. Pich, M. Leung, A. M. Frankell, C. T. Hiley, E. L. Lim, H. Zhai, M. A. Bakir, N. J. Birkbak, O. Lucas, A. Huebner, C. Puttick, K. Grigoriadis, M. Dietzen, D. Biswas, F. Athanasopoulou, S. Ward, J. Demeulemeester, C. Castignani, E. L. Cadieux, J. Kisistok, M. Sokac, Z. Szallasi, M. Dossy, R. Salgado, A. Stewart, A. Magness, C. E. Weeden, D. Levi, E. Grönroos, I. Noorani, J. Goldman, M. Escudero, P. Hobson, R. Vendramin, S. Boeing, T. Denner, V. Barbé, W. T. Lu, W. Hill, Y. Naito, Z. Ramsden, G. Kassiotis, A. Dwornik, A. Karamani, B. Chain, D. R. Pearce, D. Karagianni, F. Gálvez-Cancino, G. Stavrou, G. Mastrokalos, H. L. Lowe, I. G. Matos, J. L. Reading, J. A. Hartley, K. Selvaraju, K. Chen, L. Ensell, M. Shah, M. Litovchenko, O. Chervova, P. Pawlik, R. E. Hynds, S. Gamble, S. K. A. Ung, S. K. Bola, V. Spanswick, Y. Wu, O. al-Sawaf, T. P. Jones, S. Beck, M. Tanic, T. Marafioti, E. Borg, M. Falzon, R. Khirya, A. Toncheva, C. Abbosh, C. Richard, C. Naceur-Lombardelli, F. Gimeno-Valiente, K. Thakkar, M. W. Sunderland, M. Sivakumar, N. Kanu, P. Prymas, S. Saghafia, S. Vanloo, J. M. Lam, W. K. Liu, A. Bunkum, S. Hessey, S. Zaccaria, C. Martínez-Ruiz, J. R. M. Black, K. Thol, R. Bentham, K. Litchfield, N. McGranahan, S. A. Quezada, M. D. Forster, S. M. Lee, J. Herrero, E. Nye, R. K. Stone, J. Nicod, J. K. Rane, K. S. Peggs, K. W. Ng, K. Dijkstra, M. R. Huska, E. M. Hoogenboom, F. Monk, J. W. Holding, J. Choudhary, K. Bhakhri, M. Scarci, P. Gorman, R. C. M. Stephens, Y. N. S. Wong, Z. Kaplar, S. Bandula, T. B. K. Watkins, C. Veiga, G. Royle, C. A. Collins-Fekete, F. Fraioli, P. Ashford, A. J. Procter, A. Ahmed, M. N. Taylor, A. Nair, D. Lawrence, D. Patrini, N. Navani, R. M. Thakrar, C. Swanton, Y. Yuan, J. le Quesne, D. A. Moore, The artificial intelligence-based model ANORAK improves histopathological grading of lung adenocarcinoma. *Nat. Cancer* **5**, 347–363 (2024).
- N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, A. Tsigos, Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
- Y. Wang, X. Ju, R. Hua, J. Chen, X. Dai, L. Liu, G. Wang, Y. Bai, H. Hu, X. Li, Deep learning analysis of histopathological images predicts immunotherapy prognosis and reveals tumour microenvironment features in non-small cell lung cancer. *Br. J. Cancer* **131**, 1833–1845 (2024).
- R. S. Vanguri, J. Luo, A. T. Aukerman, J. V. Egger, C. J. Fong, N. Horvat, A. Pagano, J. D. Araujo-Filho, L. Geneslaw, H. Rizvi, R. Sosa, K. M. Boehm, S.-R. Yang, F. M. Bodd, K. Ventura, T. J. Hollmann, M. S. Ginsberg, J. Gao, MSK MIND Consortium, R. Vanguri, M. D. Hellmann, J. L. Sauter, S. P. Shah, Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L) 1 blockade in patients with non-small cell lung cancer. *Nat. Cancer* **3**, 1151–1164 (2022).
- A. H. Song, G. Jaume, D. F. Williamson, M. Y. Lu, A. Vaidya, T. R. Miller, F. Mahmood, Artificial intelligence for digital and computational pathology. *Nat. Rev. Bioeng.* **1**, 930–949 (2023).
- Z. Huang, E. Yang, J. Shen, D. Gratzinger, F. Eyerer, B. Liang, J. Nirschl, D. Bingham, A. M. Dussa, C. Kunder, R. Rojansky, A. Gilbert, A. L. Chang-Graham, B. E. Howitt, Y. Liu, E. E. Ryan, T. B. Tenney, X. Zhang, A. Folkins, E. J. Fox, K. S. Montine, T. J. Montine, J. Zou, A pathologist–AI collaboration framework for enhancing diagnostic accuracies and efficiencies. *Nat. Biomed. Eng.* **9**, 455–470 (2025).
- Y. Fu, A. W. Jung, R. V. Torne, S. Gonzalez, H. Vöhringer, A. Shmatko, L. R. Yates, M. Jimenez-Linan, L. Moore, M. Gerstung, Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat. Cancer* **1**, 800–810 (2020).
- J. M. Wang, R. Hong, E. G. Demicco, J. Tan, R. Lazcano, A. L. Moreira, Y. Li, A. Calinawan, N. Razavian, T. Schraink, M. A. Gillette, G. S. Omenn, E. An, H. Rodriguez, A. Tsigos, K. V. Ruggles, L. Ding, A. I. Robles, D. R. Mani, K. D. Rodland, A. J. Lazar, W. Liu, D. Fenyö, F. Aguet, Y. Akiyama, S. Anand, M. Anurag, Ö. Babur, J. Bavarva, C. Birger, M. J. Birrer, L. C. Cantley, S. Cao, S. A. Carr, M. Ceccarelli, D. W. Chan, A. M. Chinnaiyan, H. Cho, S. Chowdhury, M. P. Cieslik, K. R. Clauser, A. Colaprico, D. C. Zhou, F. da Veiga Leprevost, C. Day, S. M. Dhanasekaran, M. J. Domagalski, Y. Dou, B. J. Druker, N. Edwards, M. J. Ellis, M. E. Selvan, S. M. Foltz, A. Francis, Y. Geffen, G. Getz, T. J. Gonzalez Robles, S. J. C. Gosline,

- Z. H. Gümüş, D. I. Heiman, T. Hiltke, G. Hostetter, Y. Hu, C. Huang, E. Huntsman, A. Iavarone, E. J. Jaehnig, S. D. Jewell, J. Ji, W. Jiang, J. L. Johnson, L. Katsnelson, K. A. Ketchum, I. Kolodziejczak, K. Krug, C. Kumar-Sinha, J. T. Lei, W. W. Liang, Y. Liao, C. M. Lindgren, T. Liu, W. Ma, F. M. Rodrigues, W. McKerrrow, M. Mesri, A. I. Nesvizhskii, C. J. Newton, R. Oldroyd, A. G. Paulovich, S. H. Payne, F. Petralia, P. Pugliese, B. Reva, D. Rykunov, S. Satpathy, S. R. Savage, E. E. Schadt, M. Schnaubelt, S. Schürer, Z. Shi, R. D. Smith, X. Song, Y. Song, V. Stathias, E. P. Storris, N. V. Terekhanova, R. R. Thangudu, M. Thiagarajan, N. Tignor, L. B. Wang, P. Wang, Y. Wang, B. Wen, M. Wiznerowicz, Y. Wu, M. A. Wyczalkowski, L. Yao, T. M. Yaron, X. Yi, B. Zhang, H. Zhang, Q. Zhang, X. Zhang, Z. Zhang, Deep learning integrates histopathology and proteogenomics at a pan-cancer level. *Cell Rep. Med.* **4**, 101173 (2023).
- V. M. L. de Sousa, L. Carvalho, Heterogeneity in lung cancer. *Pathobiology* **85**, 96–107 (2018).
- S. M. Telli, Tumor staging and grading: A primer. *Methods Mol. Biol.* **1606**, 1–17 (2017).
- A. Jögi, M. Vaapil, M. Johansson, S. Pålman, Cancer cell differentiation heterogeneity and aggressive behavior in solid tumors. *Ups. J. Med. Sci.* **117**, 217–224 (2012).
- National Lung Screening Trial Research Team, The National Lung Screening Trial: Overview and study design. *Radiology* **258**, 243–253 (2011).
- A. Filiot, R. Ghemi, A. Olivier, P. Jacob, L. Fidon, A. Mac Kain, C. Saillard, J. B. Schiratti, Scaling self-supervised learning for histopathology with masked image modeling. medRxiv 23292757 [Preprint]. <https://doi.org/10.1101/2023.07.21.23292757>.
- A. Filiot, P. Jacob, A. Mac Kain, C. Saillard, Phikon-v2, A large and public feature extractor for biomarker prediction. arXiv:2409.09173 [eess.IV] (2024).
- K. He, X. Zhang, S. Ren, J. Sun, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778.
- V. A. Traag, L. Waltman, N. J. Van Eck, From Louvain to Leiden: Guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
- A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, M. Milacic, C. D. Roca, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, G. Viteri, J. Weiser, G. Wu, L. Stein, H. Hermjakob, P. D'Eustachio, The reactome pathway knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).
- A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, J. P. Mesirov, Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
- C. E. Barkauskas, M. J. Cronce, C. R. Rackley, E. J. Bowie, D. R. Keene, B. R. Stripp, S. H. Randell, P. W. Noble, B. L. M. Hogan, Type 2 alveolar cells are stem cells in adult lung. *J. Clin. Invest.* **123**, 3025–3036 (2013).
- R. J. Mason, M. C. Williams, Type II alveolar cell: Defender of the alveolus. *Am. Rev. Respir. Dis.* **115**, 81–91 (1977).
- R. J. Mason, Biology of alveolar type II cells. *Respirology* **11**, S12–S15 (2006).
- S. Levy, S. C. Todd, H. T. Maack, CD81 (TAPA-1): A molecule involved in signal transduction and cell adhesion in the immune system. *Annu. Rev. Immunol.* **16**, 89–109 (1998).
- A. M. Burkhardt, J. L. Maravillas-Montero, C. D. Carnevale, N. Vilches-Cisneros, J. P. Flores, P. A. Hevez, A. Zlotnik, CXCL17 is a major chemotactic factor for lung macrophages. *J. Immunol.* **193**, 1468–1474 (2014).
- K. Wang, R. Li, Y. Zhang, W. Qi, T. Fang, W. Yue, H. Tian, Prognostic significance and therapeutic target of CXCL17 in the microenvironment of lung adenocarcinoma. *Int. J. Gen. Med.* **15**, 2283–2300 (2022).
- J. A. Choreño-Parra, S. Thirunavukkarasu, J. Zúñiga, S. A. Khader, The protective and pathogenic roles of CXCL17 in human health and disease: Potential in respiratory medicine. *Cytokine Growth Factor Rev.* **53**, 53–62 (2020).
- S. F. Hashemi, H. Khorramdelazad, The cryptic role of CXCL17/CXCR8 axis in the pathogenesis of cancers: A review of the latest evidence. *J. Cell Commun. Signal.* **17**, 409–422 (2023).
- T. T. L. Vo, W. J. Jang, C. H. Jeong, Leukotriene A4 hydrolase: An emerging target of natural products for cancer chemoprevention and chemotherapy. *Ann. N. Y. Acad. Sci.* **1431**, 3–13 (2018).
- J. Z. Haeggström, Leukotriene biosynthetic enzymes as therapeutic targets. *J. Clin. Invest.* **128**, 2680–2690 (2018).
- C. Papadaki, D. Mavroudis, M. Trypaki, A. Koutsopoulos, E. Stathopoulos, D. Hatzidaki, E. Tsakalaki, V. Georgoulas, J. Souglakos, Tumoral expression of TXR1 and TSP1 predicts overall survival of patients with lung adenocarcinoma treated with first-line docetaxel-gemcitabine regimen. *Clin. Cancer Res.* **15**, 3827–3833 (2009).
- X. Yang, Y. Zhang, K. Hosaka, P. Andersson, J. Wang, F. Tholander, Z. Cao, H. Morikawa, J. Tegner, Y. Yang, H. Iwamoto, VEGF-B promotes cancer metastasis through a VEGF-A-independent mechanism and serves as a marker of poor prognosis for cancer patients. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E2900–E2909 (2015).
- G. Liu, S. Xu, F. Jiao, T. Ren, Q. Li, Vascular endothelial growth factor B coordinates metastasis of non-small cell lung cancer. *Tumor Biol.* **36**, 2185–2191 (2015).
- D. Frezzetti, M. Gallo, M. R. Maiello, A. D'Alessio, C. Esposito, N. Chicchinelli, N. Normanno, A. De Luca, VEGF as a potential target in lung cancer. *Expert Opin. Ther. Targets* **21**, 959–966 (2017).
- M. J. Conroy, J. Lysaght, CX3CL1 signaling in the tumor microenvironment. *Adv. Exp. Med. Biol.* **1231**, 1–12 (2020).
- J. Liang, G. Bi, Y. Huang, G. Zhao, Q. Sui, H. Zhang, Y. Bian, J. Yin, Q. Wang, Z. Chen, C. Zhan, MAFF confers vulnerability to cisplatin-based and ionizing radiation treatments by modulating ferroptosis and cell cycle progression in lung adenocarcinoma. *Drug Resist. Updat.* **73**, 101057 (2024).
- C. J. Hanley, S. Waise, M. J. Ellis, M. A. Lopez, W. Y. Pun, J. Taylor, R. Parker, L. M. Kimbley, S. J. Chee, E. C. Shaw, J. West, A. Alzetani, E. Woo, C. H. Ottensmeier, M. J. J. Rose-Zerilli, G. J. Thomas, Single-cell analysis reveals prognostic fibroblast subpopulations linked to molecular and immunological subtypes of lung cancer. *Nat. Commun.* **14**, 387 (2023).
- J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, Fei-Fei L. "Imagenet: A large-scale hierarchical image database" in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp.248–255.
- R. Rong, D. Luo, Z. Gu, P. Quan, I. Villanueva-Miranda, J. Wang, S. Yang, Z. Chi, P. Leavey, D. M. Yang, Y. Xie. I-Viewer: An online digital pathology analysis platform with Agentic-RAG AI copilot v1 (2024); <https://www.researchsquare.com/article/rs-5404747/>.
- X. Luo, S. Yin, L. Yang, J. Fujimoto, Y. Yang, C. Moran, N. Kalhor, A. Weissferdt, Y. Xie, A. Gazdar, J. Minna, I. I. Wistuba, Y. Mao, G. Xiao, Development and validation of a pathology image analysis-based predictive model for lung adenocarcinoma prognosis—a multi-cohort study. *Sci. Rep.* **9**, 6886 (2019).
- H. Tang, G. Xiao, C. Behrens, J. Schiller, J. Allen, C. W. Chow, M. Suraokar, A. Corvalan, J. Mao, M. A. White, I. I. Wistuba, J. D. Minna, Y. Xie, A 12-gene set predicts survival benefits from adjuvant chemotherapy in non-small cell lung cancer patients. *Clin. Cancer Res.* **19**, 1577–1586 (2013).
- F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, F. J. Theis, PAGA: Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
- L. Haghverdi, M. Büttner, F. A. Wolf, F. Büttner, F. J. Theis, Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
- F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 19 (2018).
- G. Palla, H. Spitzer, M. Klein, D. Fischer, A. C. Schaar, L. B. Kuemmerle, S. Rybakov, I. L. Ibarra, O. Holmberg, I. Virshup, M. Lotfollahi, S. Richter, F. J. Theis, Squidpy: A scalable framework for spatial omics analysis. *Nat. Methods* **19**, 171–178 (2022).
- W. Hou, Z. Ji, Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nat. Methods* **21**, 1462–1465 (2024).

Acknowledgments

Funding: This work was supported by the National Institutes of Health grant P50CA070907 (Y.X.), the National Institutes of Health grant R01GM140012 (G.X.), the National Institutes of Health grant R01DE030656 (G.X.), the National Institutes of Health grant R01CA285336 (L.C.), the National Institutes of Health grant R01GM115473 (G.X.), the National Institutes of Health grant 1U01CA249245 (G.X.), the National Institutes of Health grant 1U01AI169298 (Y.X.), the National Institutes of Health grant R35GM136375 (Y.X.), the Cancer Prevention and Research Institute of Texas grant CPRIT RP240521 (Y.X.), and the Cancer Prevention and Research Institute of Texas grant CPRIT RP230330 (G.X.). **Author contributions:** Conceptualization: Y.L., S.W., G.X., and Y.X. Methodology: Y.L., S.W., R.R., and P.Q. Software: Y.L., S.W., and P.Q. Investigation: Y.L., R.R., S.W., L.C., L.J., and Q.Z. Visualization: Y.L. and S.W. Formal analysis: Y.L., S.W., and L.C. Data curation: Y.L. and S.W. Supervision: L.C., G.X., and Y.X. Resource: Y.L., P.Q., R.R., S.W., and L.C. Writing—original draft: Y.L. Writing—review and editing: Y.L., Q.Z., L.C., G.X., and Y.X. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The datasets can be found in the following: NLST: <https://biometry.nci.nih.gov/cdas/nlst/>; LUAD: images (<https://wiki.cancerimagingarchive.net/pages/viewpage.action?pagelid=6881474>) and expression (doi:10.7908/C19P30S6); SPORE: expression GSE42127 (<https://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42127>); the SPORE image data are controlled access for qualified researchers who have been approved to access and use the data for legitimate research purposes. Xenium ADC 5K: <https://10xgenomics.com/cn/datasets/xenium-human-lung-cancer-post-xenium-technote>. All codes used for the analysis are available at <https://github.com/kateyliu/ImagePseudo>. The pretrained model and dataset used for validating the model are also available at <https://zenodo.org/records/15659784> (10.5281/zenodo.15659783). All other data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 21 January 2025

Accepted 17 June 2025

Published 18 July 2025

10.1126/sciadv.adv9466