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.

Identification of a Conserved Dedifferentiated Cancer Cell State Stratifying Poor Survival in Lung Adenocarcinoma

Bhavesh Soni

bsoni08@gmail.com

University of Oxford

Ines Grazina de Matos

F. Hoffmann-La Roche AG

Christopher Yau

University of Oxford

Stephan Gasser

F. Hoffmann-La Roche AG

Article

Keywords:

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11609499/v1

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

Additional Declarations: No competing interests reported.

Identification of a Conserved Dedifferentiated Cancer Cell State Stratifying Poor Survival in Lung Adenocarcinoma

Authors: Bhavesh Soni[1,2], Ines Grazina de Matos[2] , Christopher Yau[1], Stephan Gasser[2]

Affiliations: [1] Nuffield Department for Women's & Reproductive Health, University of Oxford, UK. [2] Roche Pharma Research and Early Development, Roche Innovation Center Zurich, Schlieren, Switzerland

Correspondence: bsoni08@gmail.com, christopher.yau@wrh.ox.ac.uk and ines.grazina_de_matos@roche.com

Abstract

Cancer cell plasticity underlies therapeutic resistance, yet its determinants remain poorly understood. While genetic mutations drive carcinogenesis, the non-malignant cell-of-origin state also dictates tumor behavior. Using single-cell transcriptomics, we characterized cancer cell subpopulations in lung adenocarcinoma by integrating non-malignant alveolar type 2 and cancer cell transcriptional states. In clinical trial bulk transcriptomes, gene signatures of four malignant subpopulations (Groups 0–3) revealed that Group 1 is strongly associated with significantly worse overall survival and exhibits a highly aggressive phenotype with stemness, hypoxia, epithelial-mesenchymal transition and intermicrovillar adhesion complex dysregulation. An ElasticNet Cox model trained on pooled cohorts (IMpower150 + OAK, N = 1,572, events = 1,205) yielded a 35-gene risk signature. In locked-model validation across held-out cohorts (IMpower110, IMpower130, BIRCH, and POPLAR), high Group 1 risk scores demonstrated strong prognostic value for elevated mortality (pooled random-effects hazard ratio [HR] = 1.326 per 1 SD increase, 95% CI: 1.218 to 1.443; $p < 0.001$), and tertile contrast confirmed pronounced survival separation (HR = 1.971, 95% CI: 1.486 to 2.614). In multivariable Cox regression, the signature was independent of age, sex, treatment regimen, smoking status and geographic location. High-risk patients across immunotherapy and chemotherapy exhibited strong activation of pro-survival and oncogenic networks (e.g., EGFR, PI3K and MAPK), whereas low-risk patients showed higher immune-related pathway activity (e.g., JAK-STAT). These findings suggest that cancer cells co-opt conserved developmental programs associated with poor outcomes regardless of treatment. The 35-gene Group 1 signature is a robust prognostic biomarker across immunotherapy and chemotherapy and provides a foundation for rational combination strategies.

Plain language summary

People with lung cancer who receive the same treatment can have very different outcomes. One reason may be that cancer cells can exist in different states, which behave differently. We studied lung adenocarcinoma, the most common type of lung cancer, by measuring gene activity in thousands of cancer cells. We used gene activity patterns from cancer cells as well as the healthy lung cells where these cancers often start to sort the tumors into groups. We found four main cancer cell states and in one of the states, the cancer cells had lost many features of normal lung cells and looked more like flexible, immature cells. Patients whose tumors had more of this state tended to survive for a shorter time. We turned this state into a score based on 35 genes and tested it in six clinical trials. Patients with high scores had shorter survival, whether they received immunotherapy or chemotherapy. This score could help identify patients at higher risk, who may need new treatment approaches.

Introduction

Tumor progression is characterized by cancer cells actively reshaping the tumor microenvironment (TME), a process that includes alterations in its cellular makeup [1]. Also, a key consequence of this is intra-tumoral heterogeneity (ITH), which enables the evolution of diverse cancer cell states within a single tumor. Although genetic mutations are commonly associated with ITH, emerging research underscores that non-genetic factors, especially the state of the cell-of-origin, play an equally crucial role in giving rise to tumor adaptability and resistance to treatment [2–5]. The unique functional traits inherent to the cells from which a tumor arises contribute to the diversity of cancer cell behaviors and influence therapeutic responses throughout disease progression.

Alveolar type II (AT2) cells are crucial for maintaining the lung alveolar lining and producing surfactant in the alveoli. Lung adenocarcinoma (LUAD) frequently originates from the transformation of these AT2 cells, often triggered by oncogenic drivers such as KRAS [6]. For squamous cell carcinoma (LSCC), the cellular origins are less certain but include basal epithelial cells and club cells [7–9]. Additionally, neuroendocrine cells or common airway epithelial progenitors are considered candidates for small cell lung cancer (SCLC) and other subtypes [10,11]. Recent single-cell studies have revealed that non-malignant AT2 cells constitute a heterogeneous population, comprising stable and transient subpopulations with diverse transcriptional profiles and distinct roles in lung repair and regeneration [12]. However, the complex interplay between this inherent heterogeneity of non-malignant AT2 cells, the acquisition of oncogenic drivers and tumor progression remains poorly understood in humans. Even in a specific lung cancer subtype like LUAD, a critical knowledge gap persists regarding how distinct cancer cell states within the TME influence a patient's response to immune checkpoint inhibitors (ICI) and other standard therapies.

To overcome this challenge, we systematically profile and utilize the transcriptional programs of non-malignant AT2 cells alongside cancer cell transcriptional programs to pinpoint and define distinct cancer cell subpopulations in human LUAD. Our objective is to identify cancer cell driven functional distinctions between tumor samples, elucidate their molecular mechanisms, and

evaluate the prognostic significance of these subpopulations within clinical cohorts. With this study we aim to enhance patient stratification and refine therapeutic targeting strategies.

Methods

Non-malignant lung single cell data processing

Single-cell RNA sequencing data were collected and processed from non-malignant human lung samples, utilizing a lung atlas developed by Sikkema et al [13]. To ensure statistical robustness and sampling depth, the dataset was filtered to include only donor samples with a minimum of 100 cells. Subsequently, potential doublets were identified using the *scanpy.pp.scrublet* function from Scanpy [14]. For this procedure, donor information was provided as the batch key to account for potential batch effects. 15 principal components were utilized in the analysis to capture the variability necessary for accurate doublet detection while avoiding overfitting. Cells identified as doublets were excluded, and only those marked as non-doublets were retained for further analysis.

The quality of the sequencing data was assessed by calculating the quality control metrics using *scanpy.pp.calculate_qc_metrics* function. For this, mitochondrial genes, ribosomal genes, and hemoglobin genes were considered to identify any potential technical artifacts or biological variance that could affect the analysis. Further, the data quality was ensured by filtering out cells with fewer than 200 genes and genes which were expressed in less than 3 cells. To adjust for differences in the sequencing depth, normalization and log-transformation was applied to raw count data using *scanpy.pp.normalize_total* and *scanpy.pp.log1p* functions. The top 2,000 highly variable genes were selected using the *scanpy.pp.highly_variable_genes* function, incorporating sample identifiers as a batch key.

Author-defined AT2 cells were isolated from the dataset, and samples with less than 50 AT2 cells were removed. PCA dimensionality reduction was followed by batch correction using *scanpy.tl.pca* and *scanpy.pp.harmony_integrate* functions, respectively. Batch correction was applied to the PCA embeddings of the data, integrating over the dataset and suspension type covariates, and the corrected embeddings were saved. Nearest neighbor graphs using the corrected PCA were calculated for downstream analysis and visualization.

Next, cell state signatures of AT2 cells were derived by identifying consensus metaprograms. For this, the AT2 single-cell data was segmented into subsets, with each subset corresponding to an individual sample. Non-negative matrix factorization was performed across these subsets using the *SplitObject* and *multiNMF* functions from the *geneNMF* package [15]. The *multiNMF* function explored factorization rank '*k*' ranging from 4 to 9, focusing on the top 1000 features to uncover distinct gene expression patterns within AT2 cells. Following this, *getMetaPrograms* function was utilized to extract 50 metaprograms, each consisting of up to 50 genes, to summarize identified patterns into higher level biological programs. Only the metaprograms with more or equal to 5 genes were kept.

Malignant samples analysis

For the integration and preprocessing of malignant single-cell RNA sequencing (scRNA-seq) datasets, multiple publicly available lung cancer datasets were utilized, including data from Xing et al., Bischoff et al., Chan et al., Qian et al., and Lung cancer atlas from Salcher et al. [16–20]. Data was filtered to include only LUAD samples from primary tumor tissues sequenced on 10x Genomics platforms. Samples labeled as normal, normal-adjacent, or undefined were excluded. Studies with pre-sorted cells were also removed. Similar to healthy lung data preprocessing, doublet detection was performed batch-wise on each dataset using the Scrublet algorithm implemented in Scanpy, with sample ids as the batch key. Doublets were excluded to minimize potential artifacts.

To prepare the data for downstream analyses, mitochondrial, ribosomal, and hemoglobin genes were annotated within the dataset. Quality control metrics for each cell, including the number of genes detected, total counts, and the percentage of mitochondrial gene expression, were calculated. Cells with fewer than 200 detected genes and genes expressed in fewer than 3 cells were removed, ensuring that only high-quality cells and informative features were retained. Normalization to median total counts was performed to account for differences in sequencing depth across cells, followed by log-transformation to stabilize variance. Highly variable genes were identified in a batch aware manner by selecting the top 2,000 genes per sample. Since the analysis focused exclusively on cancer cells, batch correction methods were not applied. This decision was based on the understanding that cancer cells generally maintain unique, sample-specific profiles, and batch correction could potentially obscure these distinct characteristics. However, soft batch effects were accounted for by incorporating sample ids during the identification of highly variable genes. Samples with fewer than 50 tumor cells were excluded to ensure robustness for statistical analyses. Further, given the higher percentage of T cells in the samples, cells expressing markers such as CD3E, CD8A, and CD4, were also removed as a precaution to make sure there are no T cell- malignant cell conjugates.

Simultaneously, to validate malignancy, we implemented a computational framework using *InferCNVpy* and *Scanpy* to analyze copy number variations (CNV) in malignant cells from single-cell RNA sequencing (scRNA-seq) data [21]. The preprocessed data containing malignant and non-malignant cells was loaded and genomic positional data for each gene was extracted from the GTF file of the Gencode v38 annotation and linked to the dataset. Normalization to a target sum of 10,000 counts per cell was performed, followed by log-transformation to stabilize variance. First, CNV profiles were inferred using the *infercnvpy.tl.infercnv* function. Genes were ordered by their genomic positions, and each cell's gene expression was compared to a global reference, calculated as the average expression of each gene across all cells, to identify large-scale copy number variations without relying on predefined reference cell type annotations. The resulting deviations were denoised and smoothed along chromosomes to generate per-gene and per-cell CNV profiles for downstream analyses. Additionally, a CNV score was computed for each cell using the *infercnvpy.tl.cnv_score* function, quantifying the extent of CNV alterations.

A key aspect of our work involved identifying subpopulations based on similar states of malignant cells, which were characterized by cancer cells and non-malignant AT2 cells driven metaprograms. For each metaprogram, we computed a gene signature score for individual cells using *scanpy.score_genes* function. Subsequently, we calculated the average gene signature score for each sample. These composite scores were subsequently standardized and used to partition samples via Leiden clustering with the Scanpy implementation (*scanpy.tl.leiden*). To identify differentially expressed genes across identified groups of tumor samples, the Wilcoxon rank-sum test was employed using the *sc.tl.rank_genes_groups* function in Scanpy. Genes were filtered to retain those with an adjusted p-value ≤ 0.01 and a log2 fold change ≥ 1. The top 50 genes per cluster, sorted by log fold change, were selected to be used for further analysis. This analysis enabled the identification of group-specific marker genes, aiding in the characterization of transcriptional signatures within the tumor microenvironment.

To further characterize the obtained groups, the *decoupleR* package was utilized to run models such as the Multivariate Linear Model (MLM) on the dataset, leveraging curated gene sets like *PROGENy and CollecTRI* [22]. These models enabled the inference of pathway activity and transcription factor activity scores across clusters, providing insights into underlying biological processes and cellular functions.

Trajectory analysis, employing the *cellrank* package, was conducted on cancer cells to understand their developmental and transitional relationships [23–25]. Initially, a diffusion map was constructed using principal components and a nearest-neighbor graph via the *sc.tl.diffmap* function. The diffusion pseudotime (DPT) algorithm then assigned a measure of progression along the inferred trajectory.

To establish the trajectory's origin, non-malignant AT2 cells from malignant samples were designated as root cells, based on the premise that LUAD malignant cells originate from AT2-like states. The root cell was computationally identified as the cell closest to the median centroid of AT2 cells in PCA space, ensuring a representative starting point. DPT was subsequently calculated from this designated root cell using the top 5 diffusion components.

For robust probabilistic inference of cell fate and lineage, a pseudotime kernel was initialized with the DPT using *cr.kernels.PseudotimeKernel*. This computed a transition matrix, modeling the probability of transitions between cells. *cr.estimators.GPCCA* (Generalized Potential of Cellular Differentiation and Commitment Analysis) then estimated the partitioning of trajectories into six macrostates.

Terminal states, representing putative stable endpoints, were predicted using the *predict_terminal_states* function, and fate probabilities were computed with *compute_fate_probabilities*. Finally, gene expression trends for specific genes of interest and signatures were modeled over the DPT axis and visualized using the *cr.pl.gene_trends* function.

To analyze the specific communication strategies employed by the identified malignant cell subpopulations, we utilized the CellPhoneDB framework [26]. The original single-cell data was

partitioned into four distinct datasets based on the identified malignant subpopulations (corresponding to groups 0, 1, 2, and 3) to ensure the analysis captured the malignant cell driven heterogeneity of the tumor microenvironment within each subset. CellPhoneDB analysis was run independently on each resulting subset to identify statistically significant Ligand-Receptor (L-R) interactions between malignant cells and key stromal/immune compartments (Fibroblasts and T cells). For each subset, the L-R data was analyzed to prioritize the most biologically relevant interactions. Only interactions with a strong statistical significance (p-value <= 0.01) were kept. These interactions were then ranked by their calculated mean expression magnitude in descending order to highlight the most physically abundant pathways. From these, only the top five highest-ranked interactions were selected for each comparison.

In Silico Characterization of Clinical Trial Cohorts

Pre-processing and Sample Stratification Across Clinical Trials

Bulk RNA-sequencing transcripts per million (TPM) expression matrices and annotated clinical metadata were collected across six Phase II and Phase III lung adenocarcinoma (LUAD) clinical trial cohorts: IMpower110, IMpower130, IMpower150, OAK, POPLAR, and BIRCH. Patient metadata were filtered prior to analysis to restrict the study population exclusively to non-squamous lung adenocarcinoma histological subtypes.

To assign a dominant malignant subpopulation identity (Groups 0 through 3) to each bulk patient sample, candidate gene signatures corresponding to the four derived malignant subpopulations were evaluated. Gene expression values were z-score normalized across samples within each cohort independently. For each subpopulation across Groups 0 to 3, a continuous activity score was calculated as the unweighted mean z-score of its constituent signature genes. Each patient sample was then assigned a dominant subpopulation label corresponding to whichever subpopulation yielded the highest signature score.

Inferred Pathway, Transcription Factor, and Microenvironmental Cell-Type Activities

Next, to dissect the functional, regulatory, and microenvironmental states across the identified subpopulations, the decoupleR package was utilized. Signaling pathway activities, including oncogenic pathways, were inferred using the human *PROGENy* network resource (*decoupler.get_progeny*, selecting the top 500 footprint genes per pathway). Additionally, transcriptional regulon activity was quantified using the human *CollecTRI* network resource (*decoupler.get_collectri*), which contains curated transcription factor-target gene interactions. Furthermore, the tumor microenvironment (TME) composition was determined by constructing cell-type marker networks from *PanglaoDB* via *decoupleR.get_resource*, encompassing primary immune and stromal lineages including T cells, B cells, NK cells, macrophages, monocytes, dendritic cells, mast cells, fibroblasts, myofibroblasts, endothelial cells, basal cells, naive T cells, cytotoxic T cells, and regulatory T cells. Finally, activity scores for these pathways, transcription

factors, and TME cell types were estimated using Multivariate Linear Models as implemented in *decoupleR.run_mlm*.

Cohort-Level Statistical Testing and Stouffer's Meta-Analysis

Within each clinical trial cohort, differential activity for each inferred feature (pathway, transcription factor, or cell type) was evaluated by comparing sample activity scores of the malignant subpopulation of interest against all remaining samples using a one-sided Mann-Whitney U test via *scipy.stats.mannwhitneyu*. Mean activity differences between in-group and out-group samples were calculated. Raw p-values within each group were adjusted for multiple hypothesis testing using the Benjamini-Hochberg False Discovery Rate (FDR) procedure via *statsmodels.stats.multitest.multipletests*.

To establish consensus signatures across all six clinical trials, a meta-analytic approach was applied. The mean activity difference was computed as the unweighted arithmetic mean of activity differences across cohorts. For meta-analyzed p-values, cohort-level raw p-values were combined using Stouffer's Z-score method via *scipy.stats.combine_pvalues*, which converts individual p-values to standard normal quantiles, sums them across cohorts, and scales by the square root of the number of cohorts. Combined Stouffer p-values were subsequently corrected across features using *statsmodels.stats.multitest.multipletests*.

Features meeting an FDR threshold of less than 0.05 were prioritized for visualizations.

Clinical Trial Cohort Processing and Signature Optimization

Discovery Model Training and Regularization

To evaluate and validate the clinical prognostic utility of the four malignant cell subpopulation signatures, bulk RNA-sequencing and clinical annotation data were compiled across clinical trial cohorts. Overall survival (OS) time was defined as the duration in days from treatment initiation (TRTSDT) to the last known alive date (LSTALVDT) or date of death (DTHDT). Censoring status was coded binary (1 = death, 0 = censored). Samples lacking duration or event indicators were removed.

To remove cohort-specific batch effects and sequencing depth variations while preserving true biological signal, gene expression matrices were z-score standardized independently within each cohort prior to statistical modeling. Signature discovery was performed on a pooled training set combining the IMpower150 and OAK cohorts (n = 1,572, total events = 1,205).

We used Elastic-Net Cox regression (via *scikit-survival*) to select the most relevant genes and build the survival model. Elastic-Net helps eliminate uninformative genes while handling genes that are closely related to one another. To find the best settings (*l1_ratio and alpha*) for the model, we ran a grid search using 3-fold cross-validation on the training data. In this process, the model was repeatedly trained on two-thirds of the data and tested on the remaining one-third ("out-of-fold" data) to identify the settings that gave the highest C-index score.

To evaluate the generalization performance of the feature selection procedure without data leakage, we performed a repeated nested cross-validation on the training set. We split the data into 3 parts, tuned the model parameters strictly on 2 of the parts, and measured performance on the remaining part. We repeated this entire 3-fold process 5 times using different random splits, resulting in 15 test evaluations. Finally, we trained a single model on the full pooled training dataset (IMpower150 and OAK) using the best settings found during cross-validation, which produced a fixed 35-gene signature. We then tested this locked 35-gene model, without making any further changes, individually on four separate test cohorts (BIRCH, POPLAR, IMpower110, and IMpower130) to evaluate its performance on unseen patient groups.

To assess how stable individual gene contributions were under sample variations, we performed bootstrap stability selection using 100 random resamples with replacement (drawing 1,572 patients per sample). On each iteration, we ran Elastic-Net Cox regression and tracked how often each candidate gene was selected. Genes selected in more than 90% of the resamples, such as *FLNC, NTSR1, NTS, DLX5, CREG2, DSG3, ZNF365*, and *FGD5*, were identified as core drivers of the signature.

Locked Signature Validation Across Independent Cohorts

The frozen 35-gene signature was evaluated on four held-out clinical trial cohorts (IMpower110, IMpower130, BIRCH, and POPLAR) that were never accessed during model training or parameter selection. For each patient in a validation cohort, a continuous risk score was computed as the linear combination of the patient's intra-cohort z-scored gene expression values and the frozen regression weights.

Prognostic association with overall survival was evaluated using two complementary frameworks. First, to allow standardized cross-trial comparison via a continuous risk metric, risk scores were continuous-standardized to unit variance within each cohort, after which univariate Cox proportional hazards models fitted via *lifelines.CoxPHFitter.fit* estimated the hazard ratio per 1 standard deviation increase in risk score along with 95% confidence intervals and Wald p-values. Second, for categorical stratification, patients were stratified into risk tiers using quantile thresholds calculated via *pandas.Series.quantile* (top vs. bottom tertiles), Kaplan-Meier survival curves were generated using *lifelines.KaplanMeierFitter*, and survival curve separation between high- and low-risk groups was tested via two-sided log-rank tests using *lifelines.statistics.logrank_test*.

Meta-Analysis and Multivariable Independence

To synthesize continuous and categorical risk estimates across validation cohorts, a DerSimonian-Laird random-effects meta-analysis was performed using statsmodels.stats.meta_analysis.combine_effects (with method_re = 'dl'). Log hazard ratios and their corresponding standard errors were weighted according to inverse variance plus between-study variance. In-sample training cohort estimates were excluded from pooled meta-analysis metrics to prevent optimism bias.

To establish whether the prognostic capacity of the 35-gene Group 1 signature is independent of established clinico-pathological variables, multivariable Cox proportional hazards regression models were fitted per cohort using *lifelines.CoxPHFitter*. Covariates adjusted for included age, sex, treatment arm, smoking status, and country of enrollment. A slight *L2* ridge penalization (via the penalizer parameter in *lifelines.CoxPHFitter*) was applied to stabilize parameter estimates in cohorts with lower events-per-variable ratios.

Characterization of Signature Risk Tiers Across Treatment Regimens

Signature Score Calculation and Risk Stratification

To evaluate functional, regulatory, and microenvironmental variations driven by the optimized Group 1 signature across clinical trial cohorts, bulk RNA-sequencing matrices (TPM) and clinical metadata were processed. For each cohort independently, gene expression values across patient samples were standardized via z-score normalization. A continuous risk score was assigned to each patient sample by matrix-multiplying the standardized expression values of the 35 signature genes by their locked model weights.

Within each clinical trial cohort, patient samples were stratified into risk tertiles based on quantile cutoffs (bottom tertile for low risk and top tertile for high risk, excluding the intermediate tertile from extreme-contrast comparisons). To evaluate treatment-specific resistance mechanisms, sample tertile classifications were intersected with assigned therapeutic classes (e.g., atezolizumab monotherapy vs. platinum-based chemotherapy), defining distinct study arms combining tertile rank and treatment type.

Activity Inference for Pathways, Transcription Factors, and Microenvironmental Lineages

To infer functional pathway activity, transcription factor dynamics, and microenvironmental cell-type enrichment across risk tiers and treatment groups, log-transformed expression matrices were processed using decoupleR with Multivariate Linear Models (*decoupler.run_mlm*). Signaling pathway activity was quantified using the *PROGENy* network resource via *decoupler.get_progeny*, while transcription factor activity was assessed using the *CollecTRI* human regulon network resource via *decoupler.get_collectri*. Additionally, tumor microenvironment analysis was performed using marker gene regulatory networks constructed from *PanglaoDB* via *decoupler.get_resource*, encompassing primary stromal and immune cell lineages.

Statistical Testing and Meta-Analysis Across Risk-Treatment Tiers

Within each clinical trial cohort, differential activity for each inferred feature (pathway, transcription factor, or cell type) across tertile-treatment groups was evaluated using a one-sided Mann-Whitney U test via *scipy.stats.mannwhitneyu*, comparing sample activity scores in a target group against all other samples. For each comparison, the mean activity difference and raw p-value

were recorded. Multiple testing correction was performed within each group using the Benjamini-Hochberg False Discovery Rate procedure via *statsmodels.stats.multitest.multipletests*.

Consensus differences across all clinical trial cohorts were established using meta-analytic pooling. The pooled mean activity difference was calculated as the unweighted arithmetic mean of activity differences across trials using *pandas.DataFrame.mean*. Cohort-level raw p-values were combined using Stouffer's Z-score method via *scipy.stats.combine_pvalues*, and these combined p-values were adjusted for multiple testing across features using *statsmodels.stats.multitest.multipletests*. Statistically significant meta-analyzed features (FDR < 0.05) were considered.

Results

Healthy AT2 Heterogeneity Shapes Malignant States

State-of-the-art single-cell mapping has established that defining disease states requires a high-resolution reference of healthy tissue architecture [27,28]. First, to accurately characterize the baseline heterogeneity of non-malignant AT2 cells, we analysed scRNAseq data from the Human Lung Cell Atlas (HLCA) [13] (Figure 1A) , which includes 115666 AT2 cells from normal, tumour-adjacent, and distal lung regions across 148 samples (112 donors) after preprocessing and filtering (Figure 1B). Figure 1A and 1C illustrate the successful identification and clustering of diverse epithelial, immune, and stromal cell types, providing a comprehensive foundation for downstream AT2-specific analyses. This baseline transcriptomic profile is further characterized by the robust expression of core pulmonary surfactant machinery genes, including SFTPC, SFTPB, SFTPA1, and ABCA3 (Figure 1D), alongside an aggregated AT2-specific signature score (Figure 1E), together establishing the functional benchmarks of homeostatic AT2 cells in healthy tissue.

To map the primary biological processes underlying transcriptional heterogeneity in AT2 cells, we extracted recurrent gene expression modules, Metaprograms (MPs), by adapting the NMF approach previously described by Gavish et al. (2023) [29] for cancer cells (see *Methods* for details). Figure 1F depicts the pairwise similarity among the generated NMF gene programs, revealing structured relationships between distinct transcriptional states. Several gene programs showed consistent activity across donors, indicating conserved AT2 functions and following a consensus analysis of generated NMF programs, 37 MPs were identified. Functional enrichment analysis of these MPs against the MSigDB Hallmark gene sets (Figure 1G) enabled biological annotation, highlighting various pathways as contributors to non-malignant AT2 cell-state diversity. Evaluating MP activities in non-malignant samples revealed a broad functional diversity in AT2 cells and these findings underscore the remarkable transcriptional plasticity of AT2 cells and demonstrate how their functional state can be modulated by their microenvironmental cues.

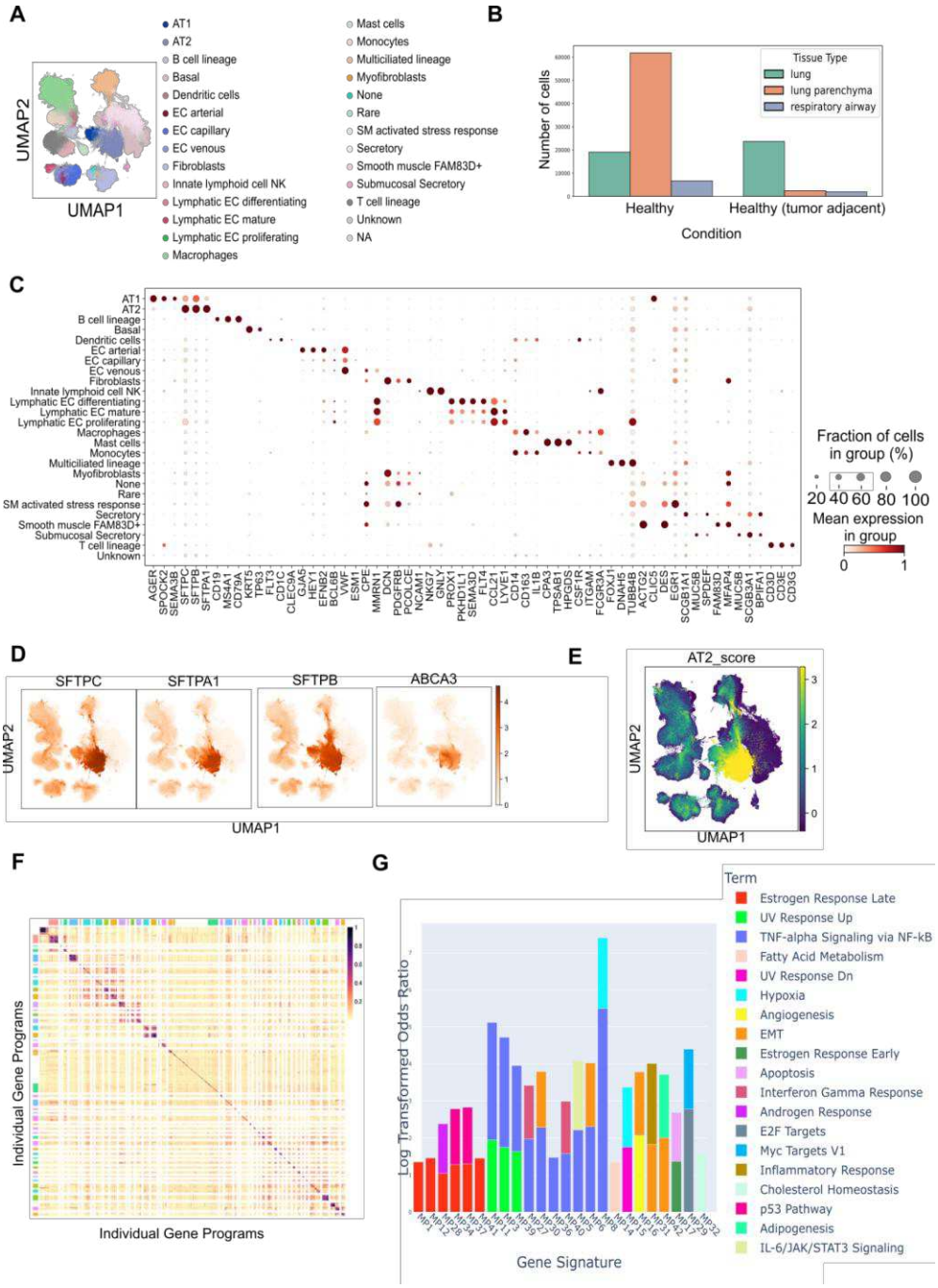


Figure 1: Characterization of Healthy AT2 Cell Heterogeneity and Functional States. (A) UMAP visualization of the human lung cell atlas colored by annotated cell types. **(B)** AT2 Cell counts across anatomical tissue types under healthy and tumor-adjacent conditions. **(C)** Dot plot of canonical cell-type marker expression across identified cell populations; dot size represents the percentage of expressing cells, and color intensity indicates mean expression level. **(D)** UMAP feature plots displaying expression levels of core surfactant-related markers **(E)** Transcriptional

AT2 signature score projected on the UMAP space. **(F)** Heatmap displaying pairwise similarity between individual gene programs derived from NMF. Color blocks represent metaprograms **(G)** Stacked bar chart showing gene set enrichment of identified MPs against MSigDB Hallmark pathways, where bar height reflects the log-transformed odds ratio of enrichment.

Non-Malignant AT2 and Malignant Metaprograms Define LUAD Subpopulations

Current understanding assumes that the transcriptional landscape of LUAD is shaped by a complex interplay between oncogenic driver mutations and inherited transcriptional programs of the cell of origin. To deconstruct these convergent influences, we analysed single-cell RNA sequencing data from the LUAD Cell Atlas [16] (Figure 2A), encompassing 38025 malignant cells derived from 63 patient samples (Figure 2B). To capture both oncogenic and lineage-associated transcriptional diversity, we considered 41 pre-existing, established malignant metaprograms [29] along with 37 programs derived from non-malignant AT2 cells (Table s1). For each LUAD sample, we calculated the average gene signature scores across cancer cells and comparative analysis of metaprogram activity profiles enabled stratification of tumours into four molecular phenotypes (Groups 0–3), each representing a distinct cancer cell transcriptional state (Figure 2C).

Differential expression analysis was performed across the four malignant cell subpopulations to identify marker genes for subsequent functional characterization (Figure 2D, Table s2). Specifically, analysis within these malignant subpopulations showed that Group 0 is characterized by genes crucial for mucosal integrity, endosomal trafficking, and homeostatic signaling. These included a mucosal tumor suppressor (*GKN2*), a plasma zymogen regulating fibrinolysis (*CPB2*), an endosomal sorting receptor (*SORCS1*), an inhibitory GABA receptor subunit (*GABRB1*), and a structural microtubule regulator (*CCDC141*). The retention of these baseline structural, protective, and homeostatic regulatory pathways reflects a well-differentiated phenotype resembling normal lung architecture, directly underpinning the favorable overall prognosis observed in Group 0 tumors. Group 1 was defined by an aggressive phenotype characterized by the expression of the neurotensin receptor *NTSR1*, the developmental transcription factor *ONECUT3*, the G-protein coupled receptor *MRGPRF*, the neutrophil/tumor marker *CD177*, and the Wnt modulator *APCDD1L*, collectively driving lineage plasticity, tumor cell adhesion, migration, and altered growth signaling. Notably, Group 1 malignant cells also showed a synchronized upregulation of genes associated with the Intermicrovillar Adhesion Complex (IMAC), including *USH1C, CDHR2, CDHR5, and EPS8L3*, which signals a breakdown of epithelial integrity. Group 2 highlighted malignant cells' reliance on a high-proliferation and high-metabolic phenotype, driven by extensive chromatin/histone replication (*H4C11, H3-3A, H2AC19, PWWP3B*), nucleoside/translational machinery (*NME2, MRPS24, EEF1G*), homeobox transcription factors (*RHOXF1*), lipid metabolic adaptation (*APOC2*), and developmental signaling (*WNT6, BMPR1B-DT, JAKMIP2*). Finally, Group 3 also represented the signature of a highly proliferative and oncogenically active malignant subtype, defined by the fatty acid binding protein *FABP7*, the cancer/testis-associated marker *TEX101*, the long non-coding RNA *DSCR8*, the branched-chain amino acid transporter *SLC6A15*, and the oncogenic antisense transcript *ST8SIA6-AS1*. This expression profile reflects a lineage program that confers metabolic adaptation, and rapid cell-cycle progression.

Further characterization of Group 0 shows that this group is enriched for AT2-specific metaprograms (Table s3) and expresses cytokine- and growth factor-inducible transcription factors (*STAT6*) along with key lineage-defining transcription factors such as *SALL2* and *NKX2-1* (Figure 2E), closely resembling non-malignant progenitors and suggesting a well-differentiated phenotype strongly reliant on lung-identity developmental transcription networks. In contrast, Group 1 is driven by *BACH1* (metastasis and oxidative stress adaptation), E4F1 (cell cycle control and metabolic reprogramming), *MEIS2* (stem-like plasticity and therapeutic resistance), *BHLHA15* (Basic Helix-Loop-Helix Family Member A15) and *GTF2IRD1* (General Transcription Factor II-I Repeat Domain-Containing Protein 1) (Figure 2E). Together, these transcription factors coordinate a regulatory network that upregulates pro-survival, invasive, and stem-like programs while suppressing differentiated cell states. More specifically, Group 1 displayed a transcriptional profile combining hypoxia and epithelial-to-mesenchymal transition (EMT) metaprograms derived from both normal and malignant cells (Table s3). This phenotype potentially reflects a preserved hypoxic signature from the cell-of-origin, coupled with the acquisition of malignant epithelial-mesenchymal plasticity. Driven by *HOXA7* and *ISL1* (driving intermediate developmental pathways and lineage specification), *NR3C2* (regulating homeostatic steroid signaling), *PBRM1* (mediating SWI/SNF-dependent chromatin remodeling), and *ZNF362* (modulating transcriptional repression and stress responses), Group 2 likely reflects a transitional phenotype. Group 2 possesses only a few significant metaprograms, therefore, rather than being defined by a unique transcriptional signature, it shares key metaprograms with other subpopulations, including those related to high cell-cycle and proliferative activity. Finally, Group 3 exhibits a highly proliferative phenotype dominated by cell-cycle and DNA replication programs (Table s3), characterized by elevated expression of the epigenetic and transcriptional regulators *ASXL1*, *LYL1*, *NEUROD6*, *TBXT*, and *ZBED1* (Figure 2E).

Additionally, we assessed the mean expression profiles of key LUAD driver genes and tumor suppressors (including *KRAS, EGFR, TP53, ERBB2, PTEN, MDM2, ARNT, FGFR1, PIK3CA, STK11, CDKN2A, CDKN2B, PDGFRA, BRF2, TERT, SOX2*, and *NKX2-1*) across the four subpopulations (Supplementary Figure S1), revealing distinct, group-specific expression patterns. The CNV profiles also varied markedly across groups. Groups 1 and 3 had the highest CNV burden, followed by Group 2 at an intermediate level, and Group 0 with the lowest level. Collectively, these observations underscore the extensive transcriptomic and genomic diversity within malignant LUAD cells, illustrating that each subpopulation is regulated by a distinct combination of lineage-derived programs and specific transcription factor networks.

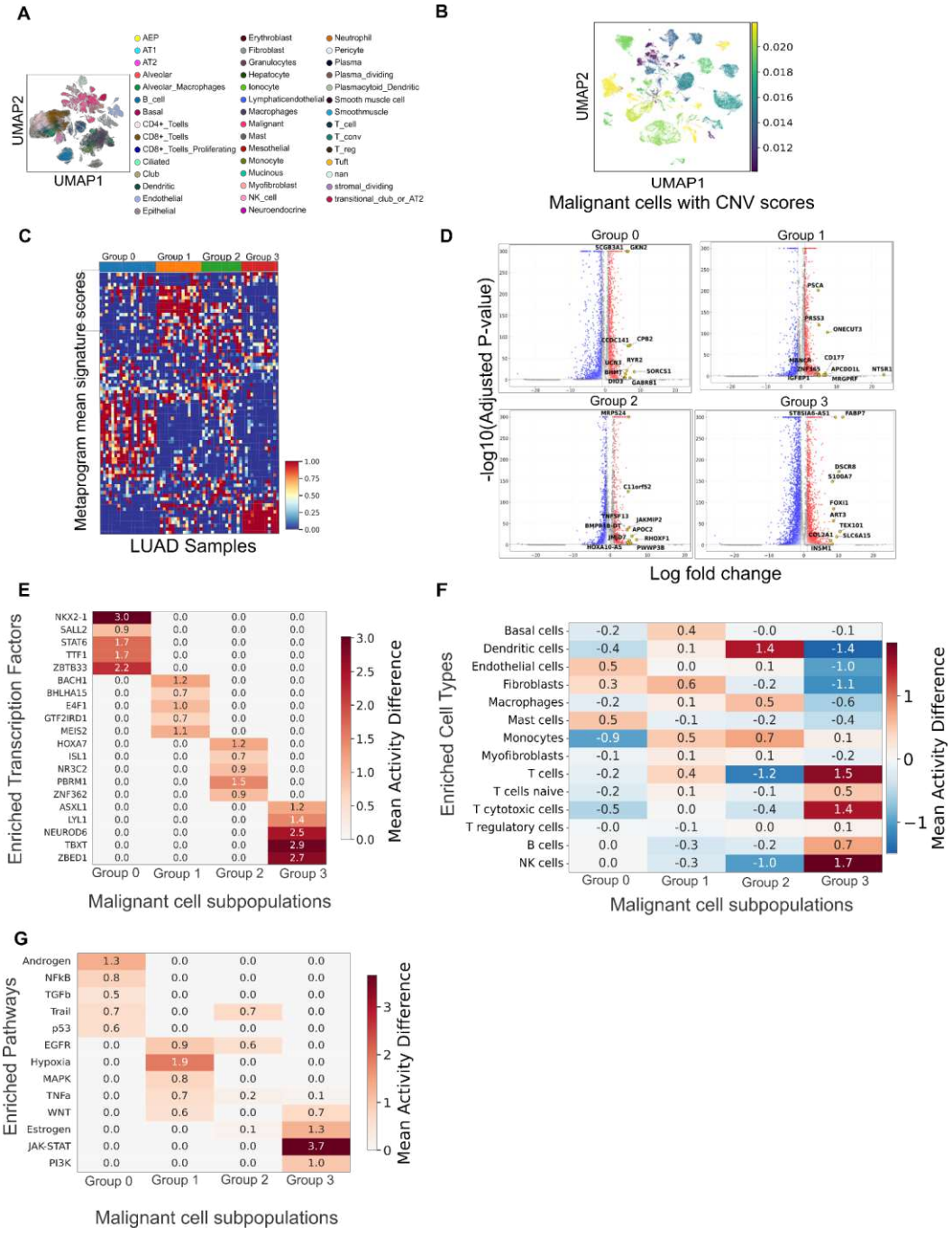


Figure 2: Identification and Characterization of Four Distinct Malignant Transcriptional States in LUAD. (A) UMAP visualization of LUAD atlas cell type annotations. **(B)** UMAP plot of isolated malignant cells colored by inferred CNV scores. **(C)** Heatmap of average MPs signature scores across LUAD samples, stratifying tumors into four functional molecular groups (Groups 0–3). **(D)** Expression of top significantly upregulated differentially expressed (DE) marker genes for each subpopulation. **(E)** Heatmap of transcription factor activity scores across the four cancer cell subpopulations. **(F)** Heatmap of cell type enrichment across the four malignant subpopulations, illustrating tumor microenvironment composition. **(G)** Heatmap of PROGENy signaling pathway

activity scores across the four cancer cell subpopulations.

Next, we examined the composition of the immune and stromal tumor microenvironment (TME) associated with each of the four malignant phenotypes (Figure 2F). Consistent with our earlier observations, each subpopulation exhibited a distinct cell-type signature reflecting specialized capacities to interact with its surrounding niche. Group 0 tumors displayed enrichment for endothelial cells, and mast cells, alongside a marked depletion of monocytes, reflecting a vascularized and tissue-maintaining niche. In contrast, Group 1 showed enrichment for fibroblasts, basal cells, and monocytes, consistent with a stromal-activated and remodeling microenvironment. Group 2 was characterized by high expression across myeloid lineages, specifically dendritic cells, monocytes, and macrophages, indicating a strongly myeloid-dominant immune context, combined with significant depletion of lymphoid lineages including T cells and NK cells. In contrast, Group 3 exhibited a prominent lymphoid-enriched TME defined by elevated NK cells, total T cells, cytotoxic T cells, and B cells but significant depletion of dendritic cells and fibroblasts. Together, these results demonstrate that the intrinsic transcriptional programs of malignant LUAD subpopulations are tightly linked to the composition and functional state of their local TME. This coupling between tumor-intrinsic states and microenvironmental architecture likely drives differences in immune evasion, stromal remodeling, and therapeutic response across LUAD subtypes.

To understand the link between TME composition and the intrinsic state of LUAD cancer cells, pathway activity inference was performed using the PROGENy framework (Figure 2G). Specifically, Group 0 exhibited selective enrichment of Androgen, NFkB, Trail, p53, and TGFb signaling pathways. Group 1 demonstrated a hyper-activated signaling flux characterized by pronounced Hypoxia activity along with coordinated EGFR, MAPK, TNFa, and WNT signaling. Group 2 displayed signaling marked primarily by Trail and EGFR activity. Conversely, Group 3 was characterized by hyper-activation of the JAK-STAT axis alongside strong Estrogen, PI3K, and WNT signaling.

Transitional States Define Metastatic Competence

In mouse models of lung cancer, Marjanovic et al. 2020 pinpointed a highly plastic transitional cell state that drives metastatic competence [30]. Key hallmarks of this aggressive state include a pronounced lineage shift marked by the loss of tissue identity through *Nkx2-1* downregulation, alongside the concurrent gain of stem-like regenerative traits via *Hmga2* upregulation. To determine whether this reciprocal gene expression pattern translates to human disease, we mapped cell-state fate probabilities to evaluate the functional plasticity of malignant cells across the four designated groups, mapping potential developmental pathways and dynamic interactions (Figure 3A). Our analysis revealed six distinct terminal populations (0 to 3_2), each defined by specific phenotypic and transcriptional programs, with terminal population 0 predominantly comprising cells from malignant subpopulation Group 0, terminal populations 1_1, 1_2, and 1_3 consisting primarily of cells from malignant subpopulation Group 1, and terminal populations 3_1 and 3_2 deriving from malignant subpopulation Group 3. The observation that most cells from

malignant group 2 remain at the center during the calculation of fate probabilities reinforces our previous finding that these cells remain uncommitted.

Cells exhibiting elevated self-transition probabilities were defined as terminal states, signifying stable, differentiated outcomes. However, most malignant cells remained uncommitted and localized centrally.

The terminal state population 0 displayed a higher AT2 cell score (Figure S2) and higher *NKX2-1* expression (Figure 3B). In contrast, other terminal state populations demonstrate a marked reduction in *NKX2-1* expression as a function of pseudotime. Across pseudotime, distinct gastric lineage-specific behaviors are evident among terminal state populations (Figures 3C and 3D). *HMGA2* expression is prominently upregulated in terminal state 1_1, which exhibits a steep increase, whereas population 0 shows a delayed, gradual peak and decline. 1_2, 1_3, 3_1 and 3_2 display only transient peaks before returning to baseline (Figure 3C). Consistent with this trend, lineage 1_1 drives a pronounced, continuous elevation in the gastric differentiation signature score (*HNF4A, CDX2, MUC5AC* and *AGR2)* , contrasting with the sustained intermediate plateau of lineage 0 and the early, transient upregulation followed by a sharp decline observed in populations 3_1, 3_2, 1_2, and 1_3 (Figure 3D).

To elucidate the dynamic changes in cellular phenotypes along pseudotime, we tracked various biological signature scores (Table s4) across terminal populations (Figure S2). While most lineages follow a conserved trajectory during early pseudotime, significant divergences emerge thereafter. Specifically, the AT2 score declines sharply from an initial high of approximately 0.8 across all groups, yet populations 0 and 1_2 uniquely reverse this trend with a secondary increase late in pseudotime. Population 1_1 shows a steep increase in stemness and population 0 displays a delayed increase, whereas populations 3_1 and 3_2 show a moderate increase and populations 1_2 and 1_3 exhibit a decline in stemness. All populations except population 0 show an increase in the glycolysis score. Populations 1_2 and 1_1 show a sharp early peak in the EMT score before declining in the terminal cells, while the EMT score continuously decreases in population 0. Furthermore, population 1_2 shows a sharp increase in the angiogenesis score, whereas it decreases across the rest of the populations. Population 1_2 also displays a sharp peak in the hypoxia score, while the remaining populations, except for population 0, show a moderate increase in hypoxia scores. Population 1_2 exhibits a sharp decline in immune evasion, whereas the rest of the populations show score modulation while maintaining higher scores for immune evasion. Finally, populations 3_1 and 3_2 show a sharp increase in the proliferation score along pseudotime.

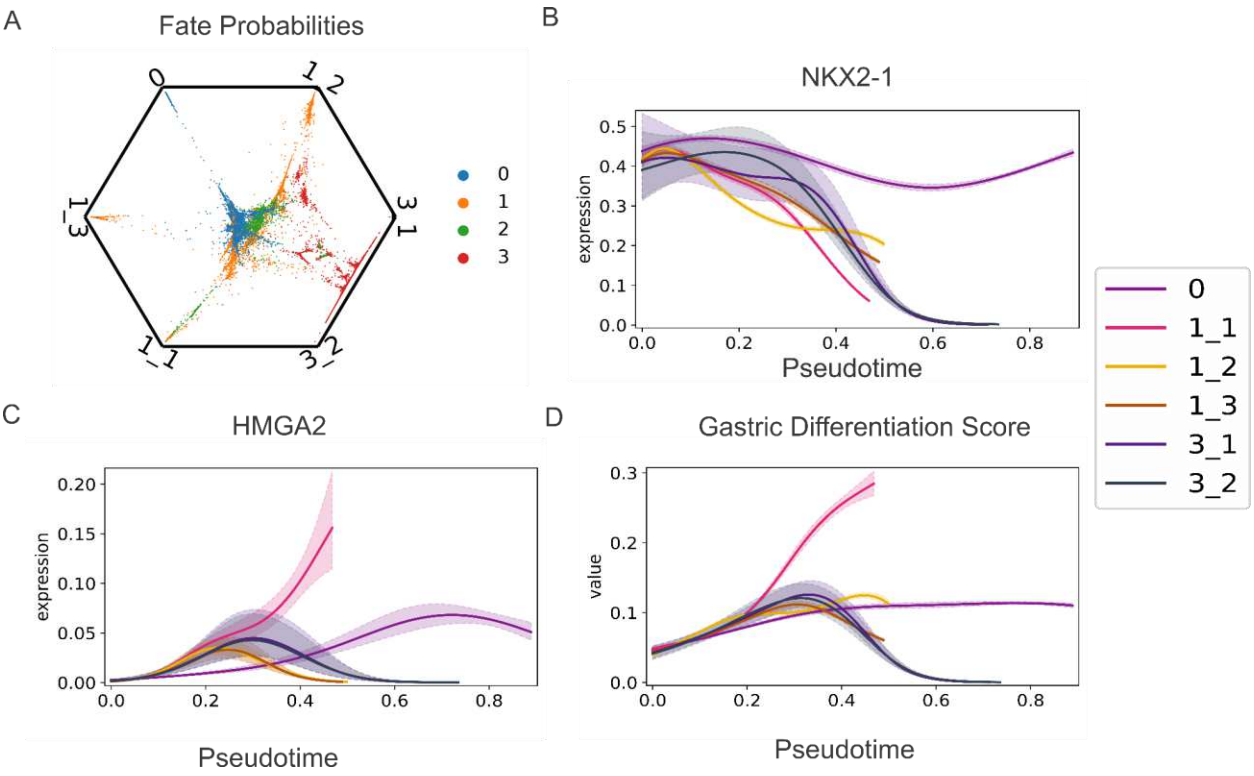


Figure 3: Malignant Cell-State Transition and Lineage Plasticity. (A) Hexagonal fate probabilities plot showing the probabilistic commitment of malignant cells toward six identified terminal states (0, 1_2, 3_1, 3_2, 1_1, and 1_3), with cells colored by their malignant subpopulation group (Groups 0–3). **(B–C)** Smoothed expression trends of canonical lineage markers (B) NKX2-1 and (C) HMGA2 across diffusion pseudotime across terminal trajectories (0, 1_1, 1_2, 1_3, 3_1, and 3_2), demonstrating reciprocal lineage marker dynamics and Group 1 dedifferentiation. **(D)** Dynamic activation of the gastric differentiation signature score plotted along diffusion pseudotime across terminal states, highlighting a sharp elevation specifically along the 1_1 lineage.

Malignant Subpopulations Drive Heterogeneous Cell-Cell Communication Networks

To functionally define how the four malignant subpopulations (Groups 0-3) remodel their microenvironment, we analyzed the top most abundant and statistically significant Ligand-Receptor (L-R) interactions from Malignant cells to two critical TME components : Fibroblasts and T cells (Table s5). The analysis of the Malignant-to-Fibroblast axis showed a highly conserved core strategy for stromal activation across all four subpopulations. This core was defined by the most abundant interaction detected across all four groups, TIMP1 signaling to CD63, a pro-tumorigenic pathway supporting fibroblast proliferation, survival, and ECM remodeling [31]. The presence of GSTP1 signaling to EGFR further indicated a shared reliance on the EGFR pathway to promote fibroblast survival. Beyond this core, each subpopulation employed unique pathways. Group 0 leverages regulatory axes such as SLPI signaling to *PLSCR1/4* and *SCGB3A1* signaling to *NOTCH3*, which likely tune local proteolytic and differentiation pathways in the stroma. Group

1 employed *LGALS3BP* and *LGALS1* signaling to *ITGB1* and *MDK* signaling to *LRP1* and *NCL*, focusing on fibroblast Activation, ECM Remodeling, adhesion, and growth factor signaling. Group 2 uses *MIF* signaling to *EGFR/TNFRSF14* to potentially drive aggressive inflammatory signaling and inhibit fibroblast apoptosis. Group 3 showed higher HMGB1 signaling to the SDC1 axis, suggesting this subpopulation is under cellular stress or activating a wound-healing response in the stroma [32,33].

Malignant cell Interactions with T cells primarily revealed a focus on immune evasion, with conserved signaling by the CXCR4 receptor being dominant. The most widely preserved regulatory connection involved HMGB1-mediated signaling through the CXCR4 receptor. Similarly, TFF3 signaling to CXCR4 indicated a common mechanism contributing to immune evasion [34]. Specific subpopulations, however, utilized specialized cancer cell - T cell interactions. Group 0 shows TGFB1 signaling to CXCR4, possibly employing the immunosuppressive cytokine to inhibit cytotoxic T cell function and steer T cells toward regulatory phenotypes [35]. Group 1 utilized LGALS1 (Galectin-1) signaling to PTPRC (CD45), a potent inhibitory mechanism where Galectin-1 interferes with the CD45 phosphatase, suppressing T cell activation [36]. Group 2 is characterized by an aggressive immunosuppressive phenotype through robust inflammatory signaling via MIF, particularly through the MIF signaling to TNFRSF14, MIF signaling to CD44_CD74 receptor complex, and MIF signaling to CD74_CXCR4 receptor complex, alongside antigen-dependent engagement of peptide–HLA. The pleiotropic effects of MIF in Group 2 are likely to functionally impair T and NK cell responses, both directly, through CD74/CD44- and CD74/CXCR4- mediated signaling that alters lymphocyte activation and migration, and indirectly, by sustaining an immunosuppressive myeloid compartment [37–40]. This is consistent with the myeloid enrichment and lymphoid depletion observed in Group 2 tumors (Figure 2F). On the other hand, Group 3 displayed a dedicated strategy, where the complement regulatory protein CD59 on cancer cells interacts with CD2 on T cells, supporting the role of CD59 in T-cell costimulation [41].

Identified malignant cell subpopulations exhibit highly diversified and specialized strategies when communicating with both the stromal and immune components of the TME likely due to different selection pressure the groups experience during tumorigenesis. The diversity underscores the necessity of pursuing therapeutic approaches that target diverse immune evasion pathways.

Clinical Relevance of Malignant Cell Subpopulation Signatures

To evaluate the clinical relevance of the four identified malignant cell states, we derived a 50-gene transcriptional signature for each group based on their top DE genes. As an initial step, we conducted a meta-analysis encompassing several major clinical trials (Table s6), including IMpower110, IMpower150, IMpower130, POPLAR, BIRCH, and OAK [42–47]. The mean expression of each malignant subpopulation group signature was correlated with patient survival in each cohort individually, directly linking single-cell observations to clinical outcomes. Initial group comparisons and random-effects meta-analysis (Figure S3 and Figure 4A, B) identified Group 1 as the signature associated with the poorest survival, while Group 0 corresponded to the most favorable prognosis.

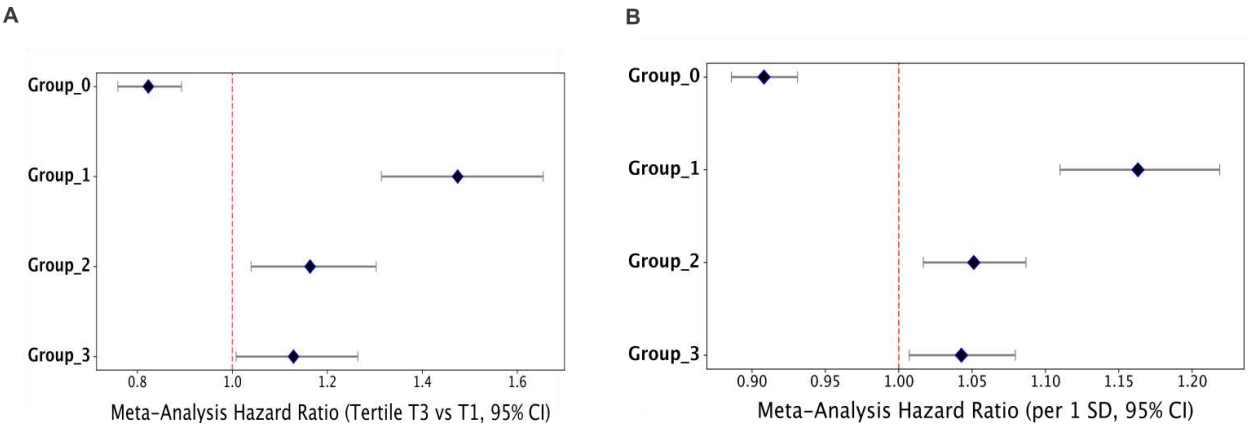


Figure 4: Prognostic Relevance of Malignant Subpopulation Signatures Across Clinical Trials. (A) Forest plot showing meta-analyzed hazard ratios (HR) and 95% confidence intervals (CI) for overall survival comparing extreme signature tertiles (Top tertile T3 vs. Bottom tertile T1) across 6 clinical trial cohorts for Groups 0–3 **(B)** Forest plot showing meta-analyzed hazard ratios per 1 standard deviation (SD) increase in continuous signature score across 6 clinical trial cohorts, identifying Group 1 as the strongest prognostic indicator of poor survival and Group 0 as protective.

To further investigate these malignant phenotypes in bulk RNA-seq clinical cohorts, we profiled their representative transcription factor networks, microenvironmental cell-type signatures, and signaling pathway activities across patient samples (Figure 5). Rather than isolating pure cancer cell-intrinsic drivers, bulk profiles reflect composite signatures arising from both cancer cells and their surrounding microenvironment. We observed that Group 0 was defined by preserved lung lineage factors (*NKX2-1*, *CEBPA*, *HOXC6*, *DLX1*, *PLAGL2*), enriched for stromal and immune constituents including endothelial cells, mast cells, dendritic cells, and fibroblasts, alongside active p53, Androgen, and Trail signaling. Group 1 showed prominent enrichment for *ELF3* and *CDX2*, coupled with a highly activated stromal microenvironment dominated by basal cells, fibroblasts, myofibroblasts, monocytes, and macrophages, alongside signaling flux through TGFb, EGFR, TNFa, Hypoxia, WNT, MAPK, VEGF, and JAK-STAT pathways. Group 2 was marked by cell-cycle and proliferative regulators (*MYC*, *REST*, *E2F4*, *HOXA2*, *MYCN*) together with regulatory T-cell infiltration, displaying prominent MAPK, VEGF, and JAK-STAT pathway enrichment. Finally, Group 3 displayed enrichment of *FOXE1*, *PTF1A*, *MBD1*, *TCF4*, and *ZNF804A*, capturing a bulk transcriptional signature defined by cytotoxic and naive T-cell signals alongside hyper-activation of PI3K, Estrogen, and NFkB signaling pathways.

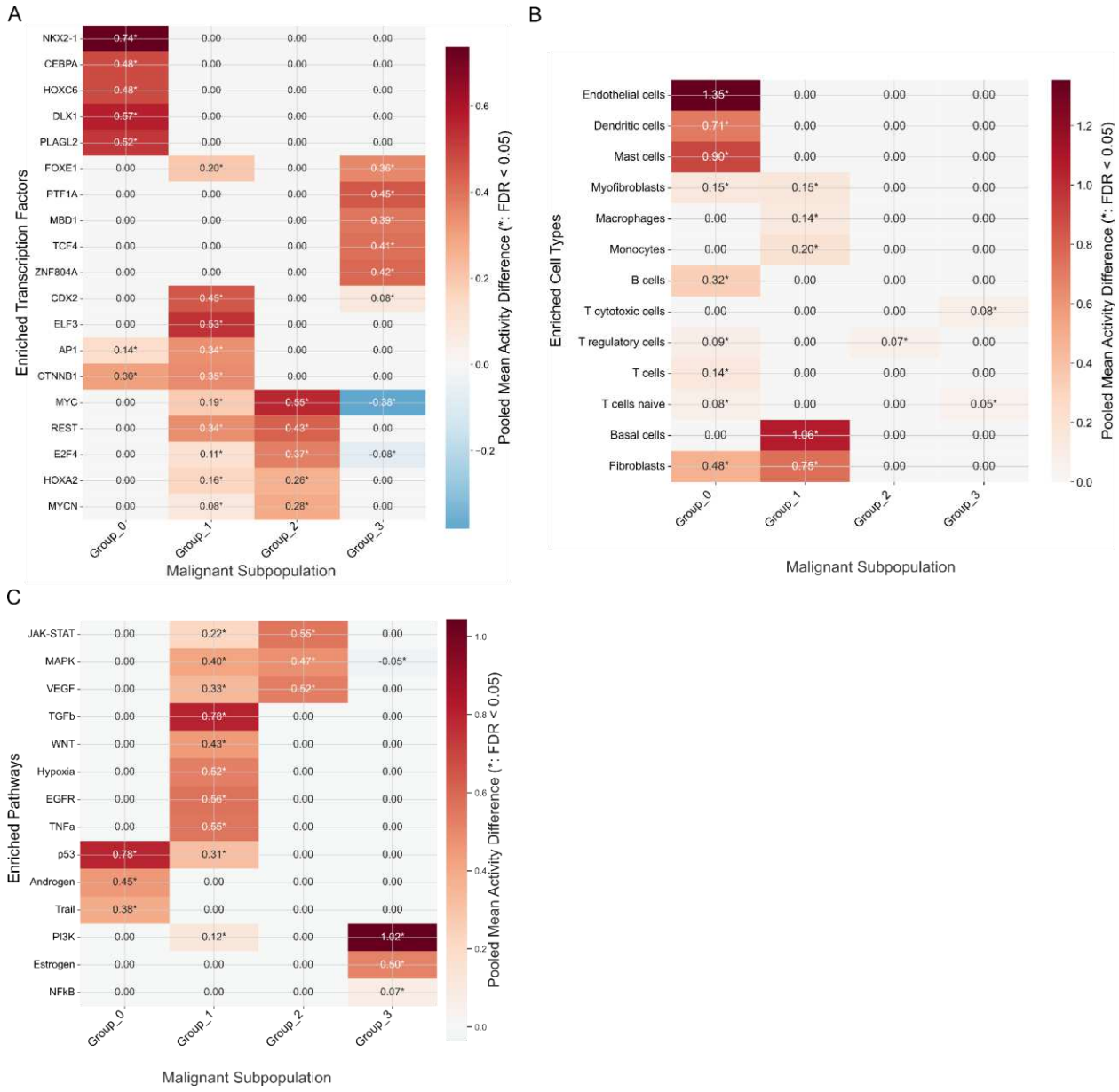


Figure 5: Molecular Characterization and Microenvironmental Profiling of Malignant Subpopulations Across 6 Clinical Cohorts. (A) Heatmap of pooled mean activity differences for key enriched transcription factors across the four malignant subpopulations (Groups 0–3). **(B)** Heatmap displaying microenvironmental cell-type enrichment (TME composition) across the four malignant groups. **(C)** Heatmap showing pooled mean activity differences for key oncogenic and signaling pathways across the four malignant groups. Asterisks indicate statistical significance (FDR < 0.05).

Optimisation and Biological Interpretation of the Group 1 Signature

To refine the prognostic capacity of the Group 1 phenotype in bulk RNA-seq cohorts, we leveraged a Cox-regularized machine learning strategy across patient samples (n = 1,572, 1,205 events). This approach isolated a 35-gene core signature, comprising *FLNC, NTSR1, NTS, FGD5, ZNF365, DLX5, DSG3, CREG2, GJB4, ANXA8, IL1RL1, TIMP4, PAEP, GRM5, KLK6, PRSS2, PSCA, CALB2, APCDD1L, FOXB1, PRSS3, SNCG, CDHR5, KREMEN2, PTH1R, ITGBL1, TNNC2, CDH23, IL20RB, PADI1, EPS8L3, CCL20, B3GALT5, IGFBP3,* and *RFLNA,* that captures the essential drivers of Group 1 malignancy. Bootstrap resampling confirmed the exceptional stability of key signature genes, which were consistently selected in over 90% of iterations.

Biologically, this 35-gene panel highlights an aggressive, stromal-activated tumor state driven by cytoskeletal remodeling, altered cell-cell adhesion, and abnormal signaling. Prominent genes included *FLNC* (filamin C) and *DSG3* (desmoglein 3), reflecting enhanced structural flexibility and junctional dynamics necessary for tissue invasion. The signature was further defined by the neuroendocrine/autocrine signaling axis *NTS/NTSR1* (neurotensin and its receptor), alongside developmental regulators like *DLX5* and *CREG2*, underscoring a high degree of lineage plasticity and stem-like identity.

We next evaluated the frozen 35-gene signature across six clinical trial cohorts (Figure 6). Within the pooled training set (IMpower150 and OAK; n = 1,572, 1,205 events), repeated nested cross-validation (5 × 3-fold) yielded a mean out-of-fold C-index of 0.581 ± 0.012, providing an unbiased estimate of the selection procedure's performance. As expected, the final model showed strong in-sample associations with overall survival in both training cohorts (IMpower150: HR per 1 SD = 1.50, 95% CI: 1.40–1.61, $p = 1.25 \times 10^{-30}$; OAK: HR = 1.41, 95% CI: 1.28–1.55, $p = 2.22 \times 10^{-12}$). Because these estimates are subject to optimism bias, they were excluded from all pooled analyses.

In the four held-out validation cohorts, which were never accessed during model training or parameter selection, higher signature scores were consistently associated with shorter overall survival: IMpower110 (HR per 1 SD = 1.45, 95% CI: 1.26–1.67, $p = 3.72 \times 10^{-7}$), IMpower130 (HR = 1.31, 95% CI: 1.19–1.44, $p = 7.93 \times 10^{-8}$), POPLAR (HR = 1.46, 95% CI: 1.18–1.81, $p = 5.08 \times 10^{-4}$) and BIRCH (HR = 1.21, 95% CI: 1.08–1.35, $p = 1.14 \times 10^{-3}$). A DerSimonian–Laird random-effects meta-analysis of the validation cohorts confirmed a significant increase in mortality risk per 1 SD of signature score (pooled HR = 1.33, 95% CI: 1.22–1.44, $p = 7.3 \times 10^{-11}$), with moderate, non-significant between-cohort heterogeneity ($I^2 = 39\%$, Cochran's Q p = 0.18). Categorical stratification gave a consistent result where patients in the top tertile had nearly twice the hazard of death of those in the bottom tertile (pooled HR = 1.97, 95% CI: 1.49–2.61).

To establish whether the Group 1 optimized signature provides prognostic value independent of established clinico-pathological variables, we performed multivariable Cox proportional hazards regression adjusting simultaneously for patient age, sex, treatment regimen, smoking status, and enrollment country across all six clinical trial cohorts (Table s7). The continuous signature risk score maintained a statistically significant, independent association with elevated mortality across

all validation and discovery trials ($p < 0.01$). Specifically, each one-standard-deviation (SD) increase in signature risk score was associated with a 20% to 54% increase in the hazard of death across individual cohorts, ranging from an adjusted HR of 1.20 (95% CI: 1.08-1.34, p = 7.67 x 10^-4) in BIRCH to 1.54 (95% CI: 1.31-1.81, p = 2.14 x 10^-7) in IMpower110. While select indicator variables for some enrollment countries reached baseline statistical significance, reflecting regional variations in sample size and healthcare context, the hazard ratio associated with the signature score remained unchanged. Furthermore, subgroup stratification confirmed that the prognostic capacity of the signature is consistent across clinical and geographic domains (Figure S4). Subgroup forest plots across age categories, sex, smoking history, and baseline treatment class demonstrated stable hazard ratios with no evidence of significant effect modification. Similarly, country-level stratification demonstrated that point estimates of risk remained consistently elevated and centered near the overall trial effect across diverse geographic enrollment sites (Figure S5).

To elucidate the biological mechanisms underlying the divergent clinical trajectories of patients, we analyzed transcription factor activity, microenvironmental cell-type enrichment, and signaling pathway flux across Group 1 signature tertiles and treatment arm groups (Figure 7A-C). Low-risk tumors (Bottom tertile) exhibited a transcriptional landscape defined by preserved alveolar/lineage maintenance and immune regulatory programs (Figure 7A). These tumors were enriched for transcription factors including *NFKB, AR, STAT1, NKX2-1, SPI1, SSRP1, HDAC1, KLF12, USF2, NR2F2, SALL4, CEBPD, MAFK, DBP,* and *HOXC6*. In contrast, high-risk tumors (Top tertile) demonstrated a complete shift toward transcription factors governing lineage transdifferentiation, intestinal/mucous metaplasia, stress response, and proliferative signaling, characterized by the enrichment of *MYC, NFE2L2, SPDEF, VENTX, CREB1, NR2F6, FOXP1, HNF1A, NFE2, CDX2, FOXE1, ATOH1,* and *NR1I3*.

Analysis of the tumor microenvironment revealed distinct cellular architectures between risk tiers (Figure 7B). Low-risk tumors across different treatment arms displayed strong enrichment across immune cell types, including B cells, dendritic cells, mast cells, monocytes, macrophages, T cells, and T regulatory cells. Conversely, high-risk tumors were predominantly depleted of these immune populations and instead showed enrichment for basal cells and fibroblasts, aligning with an immune-excluded, stromal-dense microenvironment.

Pathway activity profiling further highlighted risk-associated signaling programs (Figure 7C). Low-risk tumors showed consistent activation of NF-κB across all treatment regimens, together with Androgen signaling in most regimens and TRAIL activity in three of five regimens (atezolizumab monotherapy, atezolizumab plus bevacizumab and chemotherapy, and chemotherapy plus bevacizumab); JAK-STAT activation was most pronounced in the atezolizumab-containing arms. In contrast, high-risk tumors were defined by elevated activity of oncogenic and growth factor cascades, with Estrogen and WNT signaling enriched across all regimens and TNFa, EGFR, PI3K, and MAPK signaling enriched in a regimen-dependent manner. Together, these findings indicate that high Group 1 signature scores capture a convergence of lineage plasticity, immune evasion, stromal activation, and oncogenic signaling associated with therapeutic resistance and poor survival outcomes in LUAD.

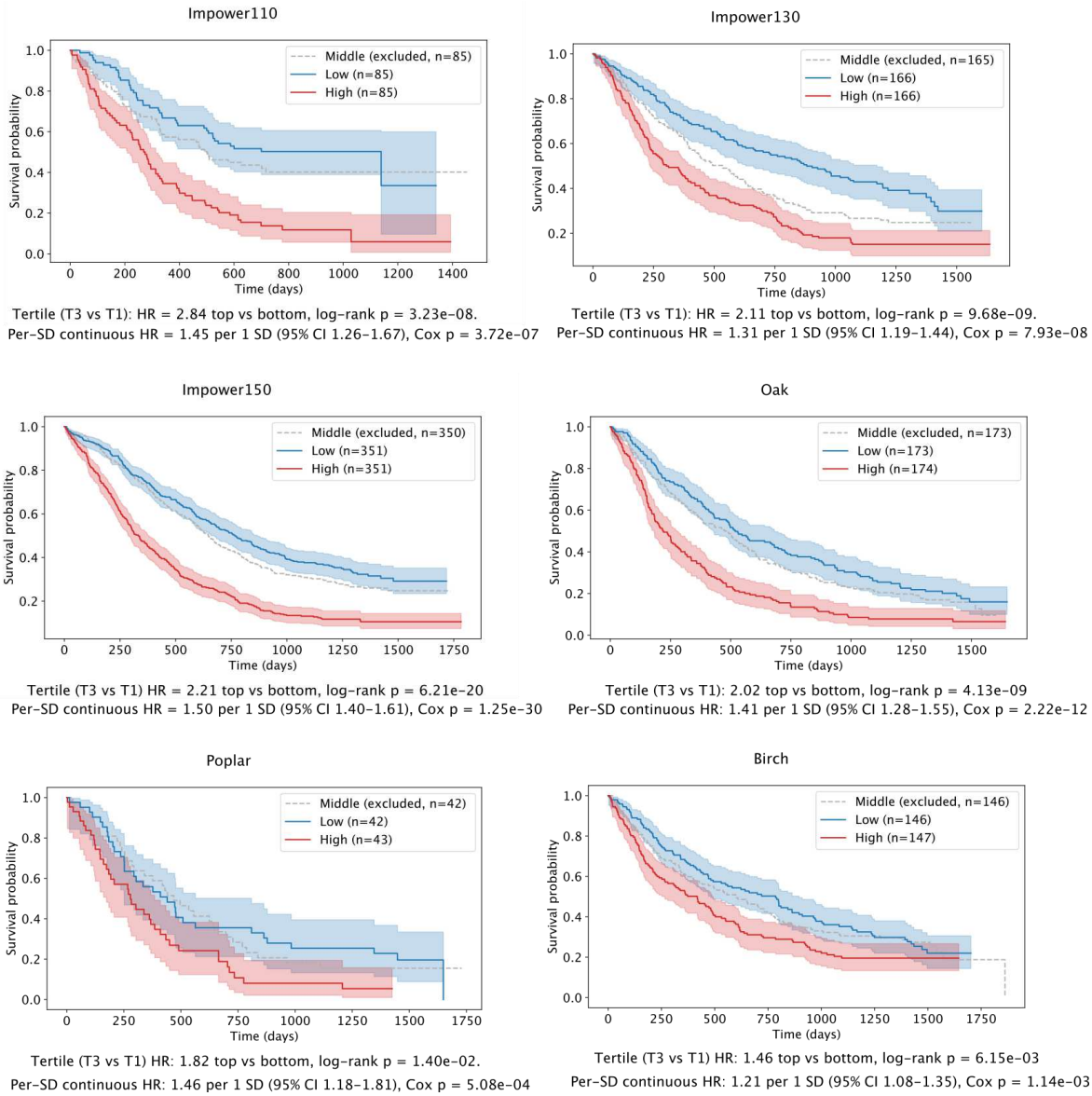


Figure 6: Validation of the Optimized 35-Gene Group 1 Risk Signature Across Clinical Trial Cohorts. Kaplan–Meier overall survival curves comparing high-risk (top tertile, red) versus low-risk (bottom tertile, blue) patient groups stratified by the 35-gene Group 1 signature score across six clinical trial cohorts: the training cohorts (IMpower150 and OAK) and independent validation test cohorts (IMpower110, IMpower130, POPLAR, and BIRCH). Shaded areas denote 95% confidence intervals, and the excluded intermediate tertile is indicated by the dashed grey line. Corresponding tertile contrast hazard ratios (HR with log-rank p-values) and continuous per-1 SD risk hazard ratios (HR with Cox p-values) are reported beneath each cohort panel.

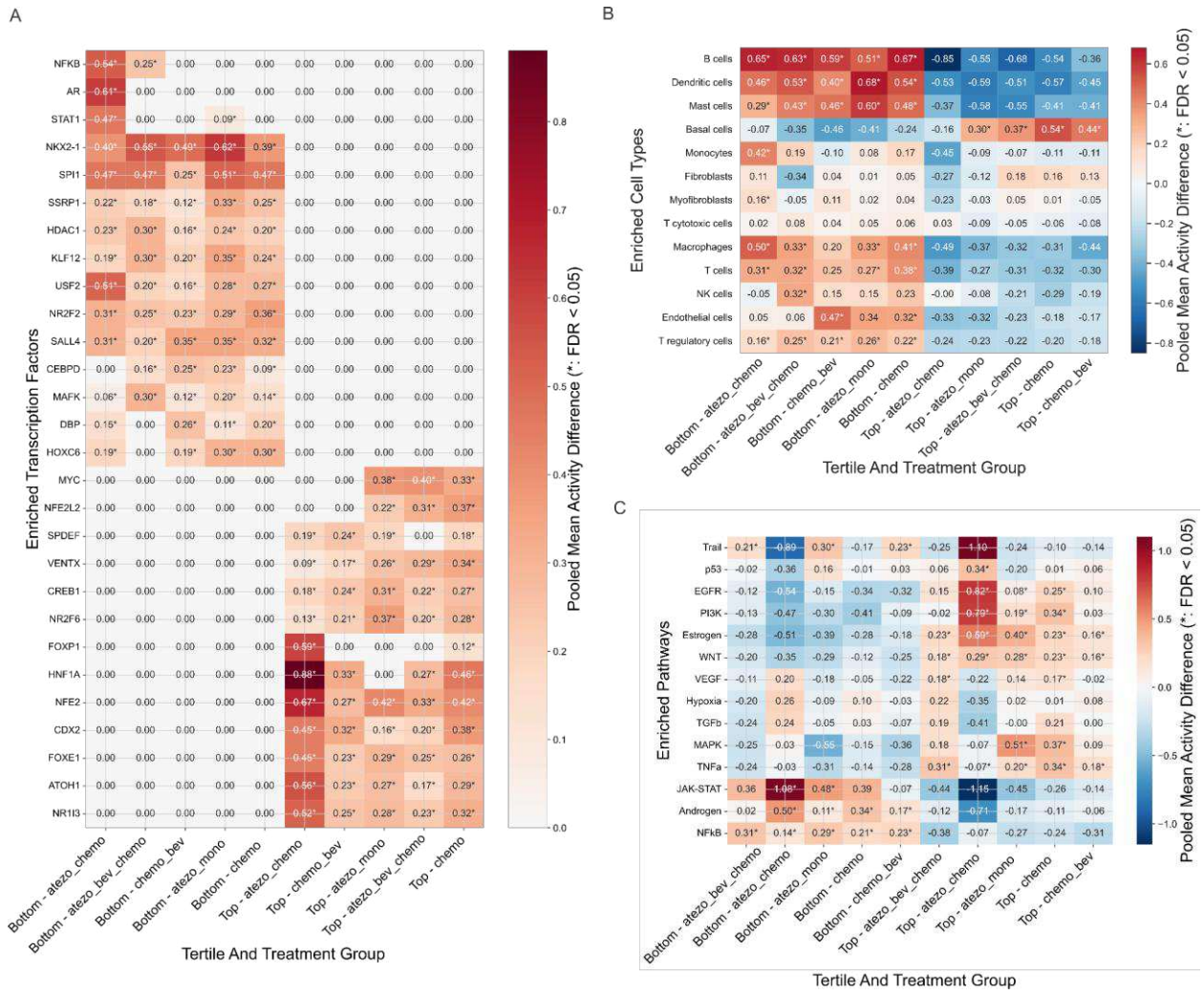


Figure 7: Functional, Microenvironmental, and Signaling Programs Dissected Across Risk Tiers and Treatment Regimens. (A) Heatmap of pooled mean activity differences for key enriched transcription factors, contrasting low-risk (bottom tertile) versus high-risk (top tertile) patient groups across diverse treatment arms (atezolizumab monotherapy, chemotherapy, atezolizumab combined with chemotherapy, and bevacizumab regimens). **(B)** Heatmap illustrating tumor microenvironment cell-type enrichment across risk tertiles and treatment groups. **(C)** Heatmap displaying pooled mean activity differences for key oncogenic and signaling pathways across risk tertiles and treatment classes. Asterisks denote statistical significance (FDR < 0.05).

Discussion

Through integrative transcriptomic profiling, this study characterizes the complex cellular heterogeneity of LUAD by outlining four transcriptionally and clinically distinct cancer cell subpopulations (Groups 0 to 3). While intra-tumoral heterogeneity has frequently been characterized as an unstructured consequence of genomic instability, our computational analyses suggest that LUAD tumors show discrete, reproducible cancer cell states associated with distinct transcriptional programs, metabolic configurations, microenvironmental compositions, and clinical trajectories. Distinctly across this spectrum, Group 0 tumors maintain differentiated epithelial features and correlate with favorable patient outcomes, whereas Group 1 represents an aggressive, dedifferentiated transcriptional state that consistently correlates with poor overall survival and elevated risk across diverse therapeutic regimens.

Biological Interpretations and State-Specific Heterogeneity

First, evaluating the computational profiles of Groups 0, 2, and 3 reveals diverse cellular and metabolic states in LUAD. Group 0 is marked by the retention of markers associated with mucosal integrity (*GKN2*), fibrinolytic regulation (*CPB2*), endosomal trafficking (*SORCS1*), an inhibitory GABA receptor subunit (*GABRB1*), a structural microtubule regulator (*CCDC141*), and canonical pulmonary lineage transcription factors (*NKX2-1, ZBTB33, STAT6, SALL2*). Resembling well-differentiated lung epithelium with a comparatively low copy-number variation burden, Group 0 tumors correlate with an endothelial and mast-cell-enriched microenvironment alongside elevated predicted activity of TRAIL, p53, and androgen tumor-suppressive pathways. In contrast, Group 2 and Group 3 represent distinct proliferative phenotypes. Group 2 is characterized by elevated expression of machinery involved in chromatin and histone replication (*H4C11, H3-3A, H2AC19, PWWP3B*), nucleoside turnover and translational machinery (*NME2, MRPS24, EEF1G*), homeobox transcription factors (*RHOXF1*), lipid metabolism (*APOC2*), and developmental signaling (*WNT6, BMPR1B-DT, JAKMIP2*). In silico prediction suggests that Group 2 tumors reside within a myeloid-dominated microenvironment enriched for monocytes, macrophages, and dendritic cells, with cell-cell communication modeling prioritizing macrophage migration inhibitory factor (*MIF*) interactions (*MIF-CD74/CD44, MIF-TNFRSF14*), indicating a state characterized by coordinated metabolic and inflammatory signatures. Meanwhile, Group 3 exhibits an elevated CNV burden, specialized nutrient transporter expression (*SLC6A15, FABP7*), a cancer/testis-associated marker (*TEX101*), a long non-coding RNA (*DSCR8*), an oncogenic antisense transcript (*ST8SIA6-AS1*), and regulatory transcription factor signals (*ASXL1, NEUROD6, TBXT, ZBED1*). While Group 3 tumors display T-cell infiltration, cell-cell communication analyses also highlight cancer cell to T-cell interactions through the *CD59/CD2* signaling axis.

Interestingly, unlike the differentiated or proliferative features observed in the other groups, Group 1 embodies a dedifferentiated transcriptional phenotype characterized by marked stemness and structural remodeling signatures. Defined computationally by a transcription factor regulon including *BACH1, MEIS2, E4F1, BHLHA15, and GTF2IRD1*, Group 1 exhibits elevated hypoxia and epithelial-to-mesenchymal transition (EMT) scores alongside substantial CNV alterations. Single-cell pseudotime and fate-mapping models place Group 1 cells at multiple terminal states

as well as uncommitted states showing transition probabilities toward a stemness-enriched trajectory (1_1), an angiogenesis and hypoxia-enriched trajectory (1_2), and an additional terminal trajectory (1_3). Dedifferentiated trajectory 1_1 is characterized by an inverse expression pattern marked by the down-regulation of *NKX2-1* and reciprocal elevation of *HMGA2*, accompanied by the expression of gastric lineage markers (*HNF4A, CDX2, MUC5AC, AGR2*). A transient surge in EMT score observed in both the 1_1 and 1_2 trajectories suggests a potential enrichment of EMT within uncommitted Group 1 cells. This transcriptomic profile is consistent with established features of adaptive plasticity in epithelial tumors. In parallel, Group 1 cells also display coordinated upregulation of intermicrovillar adhesion complex (IMAC) components (*USH1C, CDHR2, CDHR5, EPS8L3*), suggesting that cytoskeletal and brush-border remodeling accompanies a loss of apical domain polarity in alveolar cells, pointing to a mechanobiological link to EMT, tissue invasion, and stemness.

The Group 1 subpopulation further correlates with distinct cell-cell communication patterns inferred from tumor, stromal, and immune transcriptomes. While all subpopulations display baseline predicted interactions for stromal activation via *TIMP1-CD63* and *GSTP1-EGFR*, Group 1 preferentially enriches for interaction pairs linked to cancer-associated fibroblasts, such as *MDK-LRP1, MDK-NCL, LGALS3BP-ITGB1, and LGALS1-ITGB1*.

Furthermore, ligand-receptor analysis suggests candidate mechanisms of immune modulation, highlighting Galectin-1 (*LGALS1*) pairing with PTPRC (*CD45*) on T cells, an interaction associated with impaired lymphocyte activation. This stromal-dense, immune-excluded microenvironment contrasts with the vascularized profile of Group 0, the myeloid-predominant signature of Group 2, and the lymphoid infiltration of Group 3, illustrating how tumor-intrinsic transcriptomic states mirror distinct surrounding immunological landscapes.

Clinical Translation and Prognostic Framework

To translate these findings into an objective biomarker framework, an Elastic-Net regularized Cox proportional hazards regression model was applied to derive a 35-gene signature reflecting the Group 1 transcriptomic phenotype. Across six Phase II and Phase III clinical trial cohorts, this locked continuous risk score demonstrated consistent prognostic performance across training cohorts (IMpower150, OAK) and independent test cohorts (IMpower110, IMpower130, POPLAR, BIRCH). Multivariable Cox regression and subgroup meta-analyses confirmed that the prognostic association of the signature is independent of established clinico-pathological variables, including age, sex, treatment arm, smoking history, and geographic location (adjusted HR = 1.20 to 1.54 per SD, $p < 0.01$ across cohorts). Key components of this signature, such as the cytoskeletal gene *FLNC*, the desmosomal cadherin *DSG3*, the homeobox factor *DLX5*, and the neuroendocrine axis *NTS/NTSR1*, represent stable correlates of this aggressive tumor phenotype across clinically diverse cohorts.

Pathway activity enrichment across risk tiers and treatment classes provides correlative clues regarding therapeutic susceptibility and resistance. While low-risk Group 1 tumors display immune infiltration signatures and preserved pulmonary lineage factors, high-risk Group 1 tumors are associated with regimen-specific signaling programs. Under atezolizumab monotherapy, low-

risk tumors are characterized by elevated JAK-STAT, TRAIL, and NF-kB signaling, whereas high-risk tumors demonstrate activation of oncogenic and stress response pathways, including MAPK, Estrogen, WNT, PI3K, TNFa, and EGFR. Under chemotherapy, high-risk tumors similarly exhibit persistent enrichment across pro-survival cascades, including MAPK, PI3K, TNFa, EGFR, and WNT.

Crucially, intersecting signature risk tiers with multi-arm clinical trial regimens reveals deeper therapeutic insights regarding treatment resistance dynamics. While both low-risk monotherapy and combination cohorts benefit from an immune-inflamed microenvironment, high-risk tumors consistently suppress immune infiltration, showing marked depletion of cytotoxic T cells and natural killer cells regardless of whether the regimen incorporates an immune checkpoint inhibitor or standard cytotoxic chemotherapy. This suggests that the cell-intrinsic Group 1 state likely exerts dominant immune exclusion that supersedes the reactivation intended by checkpoint blockade.

Limitations and Future Directions

Despite these robust findings, several limitations warrant consideration. First, the analyses rely on retrospective evaluations of archival clinical trial cohorts, which are inherently susceptible to selection biases and unmeasured confounding factors. Second, while our in silico pathway deconvolution and trajectory models offer valuable mechanistic hypotheses regarding therapeutic resistance, these computational predictions require prospective validation using spatial profiling to map microenvironmental cellular neighborhoods in situ, alongside targeted functional CRISPR screens to establish direct causality before clinical translation. Overcoming the poor outcomes associated with this high-risk state will likely require rational combination strategies that target both cell-intrinsic growth pathways and the immunosuppressive microenvironment.

In conclusion, this study presents a comprehensive computational framework resolving malignant cell state diversity in LUAD, demonstrating that discrete cancer cell subpopulations correlate with specialized transcriptional, metabolic, and microenvironmental programs. The validated 35-gene Group 1 signature provides a robust prognostic tool for clinical risk stratification and offers a foundation for future experimental studies to functionally dissect these state-associated resistance programs.

Code Availability

All custom code developed for this analysis, including scripts for single-cell data processing, metaprogram scoring, sample clustering, pseudotime analysis, and clinical trial analysis, is deposited on (https://github.com/bsoni08/luad_stratification)

Data Availability

All single-cell RNA sequencing (scRNA-seq) data utilized in this study are publicly available. This includes the non-malignant single-cell data from the Human Lung Cell Atlas and the malignant LUAD scRNA-seq datasets.

Qualified researchers may request access to individual patient-level data through the clinical study data request platform (https://vivli.org/). Further details on Roche's criteria for eligible studies are available here:

(https://vivli.org/members/ourmembers/)

For further details on Roche's Global Policy on the Sharing of Clinical Information and how to request access to related clinical study documents, see here:

(https://www.roche.com/research_and_development/who_we_are_how_we_work/clinical_trials/our_commitment_to_data_sharing.htm).

Conflicts of Interests

BS and IGM are affiliated with Roche Glycart AG. SG was an employee of Roche Glycart AG during the conduct of this study. CY is a recipient of funding as part of a Roche DPhil Studentship to BS.

Ethics & Approvals

The raw data utilized in this study were obtained from previously published human studies that received ethical clearance and patient consent from their original institutions. All data were de-identified and analyzed in accordance with the established ethical guidelines and institutional approvals of the authors' affiliated institutions. No new patient samples were collected for this study.

All six clinical trials (IMpower110, NCT02409342; IMpower130, NCT02367781; IMpower150, NCT02366143; OAK, NCT02008227; POPLAR, NCT01903993; BIRCH, NCT02031458) were conducted in accordance with the Declaration of Helsinki and Good Clinical Practice guidelines, were approved by the ethics committee or institutional review board at each site, and all patients provided written informed consent.

Author Contributions

BS: Data Curation, Conceptualization, Methodology, Validation, Software, Visualization, Writing.
CY: Supervision, Conceptualization, Methodology, Writing, Reviewing and Editing.
IGM: Supervision, Conceptualization, Writing, Reviewing and Editing.
SG: Supervision, Conceptualization, Writing, Reviewing and Editing.

Acknowledgements

We thank Rachael Bashford-Rogers, David Church, Ahmed Ahmed, David R Westhead and Christine Trumpfheller for support, advice and comments on the work leading to this paper.

Funding

CY is supported by a UKRI Turing AI Acceleration Fellowship (Ref: EP/V023233/1).

References

1. Hanahan, D. & Coussens, L. M. Accessories to the Crime: Functions of Cells Recruited to the Tumor Microenvironment. *Cancer Cell* **21**, 309–322 (2012).
2. Visvader, J. E. Cells of origin in cancer. *Nature* **469**, 314–322 (2011).
3. Feng, J., Zhao, D., Lv, F. & Yuan, Z. Epigenetic Inheritance From Normal Origin Cells Can Determine the Aggressive Biology of Tumor-Initiating Cells and Tumor Heterogeneity. *Cancer Control* **29**, 10732748221078160 (2022).
4. Krivtsov, A. V. *et al.* Cell of origin determines clinically relevant subtypes of MLL-rearranged AML. *Leukemia* **27**, 852–860 (2013).
5. Bell, C. C. & Gilan, O. Principles and mechanisms of non-genetic resistance in cancer. *Br. J. Cancer* **122**, 465–472 (2020).
6. Xu, X. *et al.* Evidence for type II cells as cells of origin of K-Ras–induced distal lung adenocarcinoma. *Proc. Natl. Acad. Sci.* **109**, 4910–4915 (2012).
7. Gómez-López, S. *et al.* Aberrant basal cell clonal dynamics shape early lung carcinogenesis. *Science* **388**, eads9145 (2025).
8. Ferone, G. *et al.* SOX2 Is the Determining Oncogenic Switch in Promoting Lung Squamous Cell Carcinoma from Different Cells of Origin. *Cancer Cell* **30**, 519–532 (2016).
9. Ruiz, E. J. *et al.* LUBAC determines chemotherapy resistance in squamous cell lung cancer. *J. Exp. Med.* **216**, 450–465 (2019).
10. Sutherland, K. D. *et al.* Cell of Origin of Small Cell Lung Cancer: Inactivation of Trp53 and Rb1 in Distinct Cell Types of Adult Mouse Lung. *Cancer Cell* **19**, 754–764 (2011).
11. Ireland, A. S. *et al.* Basal cell of origin resolves neuroendocrine–tuft lineage plasticity in cancer. *Nature* **647**, 257–267 (2025).
12. Chen, Q. & Liu, Y. Heterogeneous groups of alveolar type II cells in lung homeostasis and repair. *Am. J. Physiol.-Cell Physiol.* **319**, C991–C996 (2020).
13. Sikkema, L. *et al.* An integrated cell atlas of the lung in health and disease. *Nat. Med.* **29**, 1563–1577 (2023).
14. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

15. Yerly, L. *et al.* Wounding triggers invasive progression in human basal cell carcinoma. *bioRxiv* 2024.05.31.596823 (2025) doi:10.1101/2024.05.31.596823.
16. Salcher, S. *et al.* High-resolution single-cell atlas reveals diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer. *Cancer Cell* **40**, 1503-1520.e8 (2022).
17. Xing, X. *et al.* Decoding the multicellular ecosystem of lung adenocarcinoma manifested as pulmonary subsolid nodules by single-cell RNA sequencing. *Sci. Adv.* **7**, eabd9738 (2021).
18. Bischoff, P. *et al.* Single-cell RNA sequencing reveals distinct tumor microenvironmental patterns in lung adenocarcinoma. *Oncogene* **40**, 6748–6758 (2021).
19. Chan, J. M. *et al.* Signatures of plasticity, metastasis, and immunosuppression in an atlas of human small cell lung cancer. *Cancer Cell* **39**, 1479-1496.e18 (2021).
20. Qian, J. *et al.* A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. *Cell Res.* **30**, 745–762 (2020).
21. GitHub - icbi-lab/infercnvpy: Infer copy number variation (CNV) from scRNA-seq data. https://github.com/icbi-lab/infercnvpy.
22. Badia-i-Mompel, P. *et al.* decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinform. Adv.* **2**, vbac016 (2022).
23. Weiler, P., Lange, M., Klein, M., Pe'er, D. & Theis, F. CellRank 2: unified fate mapping in multiview single-cell data. *Nat. Methods* **21**, 1196–1205 (2024).
24. Lange, M. *et al.* CellRank for directed single-cell fate mapping. *Nat. Methods* **19**, 159–170 (2022).
25. Reuter, B., Fackeldey, K. & Weber, M. Generalized Markov modeling of nonreversible molecular kinetics. *J. Chem. Phys.* **150**, 174103 (2019).
26. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).
27. Dann, E. *et al.* Precise identification of cell states altered in disease using healthy single-cell references. *Nat. Genet.* **55**, 1998–2008 (2023).
28. Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. & Teichmann, S. A. The Human Cell Atlas: from vision to reality. *Nature* **550**, 451–453 (2017).
29. Gavish, A. *et al.* Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. *Nature* **618**, 598–606 (2023).
30. Marjanovic, N. D. *et al.* Emergence of a High-Plasticity Cell State during Lung Cancer Evolution. *Cancer Cell* **38**, 229-246.e13 (2020).
31. Gong, Y. *et al.* TIMP-1 Promotes Accumulation of Cancer Associated Fibroblasts and Cancer Progression. *PLoS ONE* **8**, e77366 (2013).
32. Yang, Z., Chen, S., Ying, H. & Yao, W. Targeting syndecan-1: new opportunities in cancer therapy. *Am. J. Physiol.-Cell Physiol.* **323**, C29–C45 (2022).
33. Abe, A., Kuwata, T., Yamauchi, C., Higuchi, Y. & Ochiai, A. High Mobility Group Box1 (HMGB1) released from cancer cells induces the expression of pro-inflammatory cytokines in peritoneal fibroblasts. *Pathol. Int.* **64**, 267–275 (2014).
34. Pan, X. *et al.* Inhibition of TFF3 improves the infiltration and function of CD8+ T cells by downregulating the expression of PD-L1 in colorectal cancer. *Int. J. Color. Dis.* **40**, 170 (2025).
35. Yang, Y. *et al.* CXCL12-CXCR4/CXCR7 Axis in Cancer: from Mechanisms to Clinical Applications. *Int. J. Biol. Sci.* **19**, 3341–3359 (2023).

36. Xu, K. *et al.* Single-cell RNA sequencing reveals cell heterogeneity and transcriptome profile of breast cancer lymph node metastasis. *Oncogenesis* **10**, 66 (2021).
37. Tsou, L. K. *et al.* Harnessing CXCR4 antagonists in stem cell mobilization, HIV infection, ischemic diseases, and oncology. *Med. Res. Rev.* **38**, 1188–1234 (2018).
38. Kang, I. & Bucala, R. The immunobiology of MIF: function, genetics and prospects for precision medicine. *Nat. Rev. Rheumatol.* **15**, 427–437 (2019).
39. Shi, X. *et al.* CD44 Is the Signaling Component of the Macrophage Migration Inhibitory Factor-CD74 Receptor Complex. *Immunity* **25**, 595–606 (2006).
40. Bernhagen, J. *et al.* MIF is a noncognate ligand of CXC chemokine receptors in inflammatory and atherogenic cell recruitment. *Nat. Med.* **13**, 587–596 (2007).
41. Deckert, M. *et al.* CD59 molecule: A second ligand for CD2 in T cell adhesion. *Eur. J. Immunol.* **22**, 2943–2947 (1992).
42. Herbst, R. S. *et al.* Atezolizumab for First-Line Treatment of PD-L1–Selected Patients with NSCLC. *N. Engl. J. Med.* **383**, 1328–1339 (2020).
43. Socinski, M. A. *et al.* Atezolizumab for First-Line Treatment of Metastatic Nonsquamous NSCLC. *N. Engl. J. Med.* **378**, 2288–2301 (2018).
44. West, H. *et al.* Atezolizumab in combination with carboplatin plus nab-paclitaxel chemotherapy compared with chemotherapy alone as first-line treatment for metastatic non-squamous non-small-cell lung cancer (IMpower130): a multicentre, randomised, open-label, phase 3 trial. *Lancet Oncol.* **20**, 924–937 (2019).
45. Rittmeyer, A. *et al.* Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. *Lancet* **389**, 255–265 (2017).
46. Fehrenbacher, L. *et al.* Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial. *Lancet* **387**, 1837–1846 (2016).
47. Peters, S. *et al.* Phase II Trial of Atezolizumab As First-Line or Subsequent Therapy for Patients With Programmed Death-Ligand 1–Selected Advanced Non–Small-Cell Lung Cancer (BIRCH). *J. Clin. Oncol.* **35**, JCO.2016.71.947 (2017).

Supplementary data

Supplementary Tables

1. Non-malignant AT2 cell metaprograms and Gavish et al. cancer cell metaprograms used in analysis
2. top 50 DE genes per malignant subpopulation
3. Comparison of the malignant cell subpopulations (top significant metaprograms)
4. Gene signatures for figure S2
5. Ligand-Receptor results (source: Malignant cell , target: T cell, Fibroblasts)
6. Comparison of Clinical trials used in the study
7. Multivariable cox covariate results

Supplementary figures

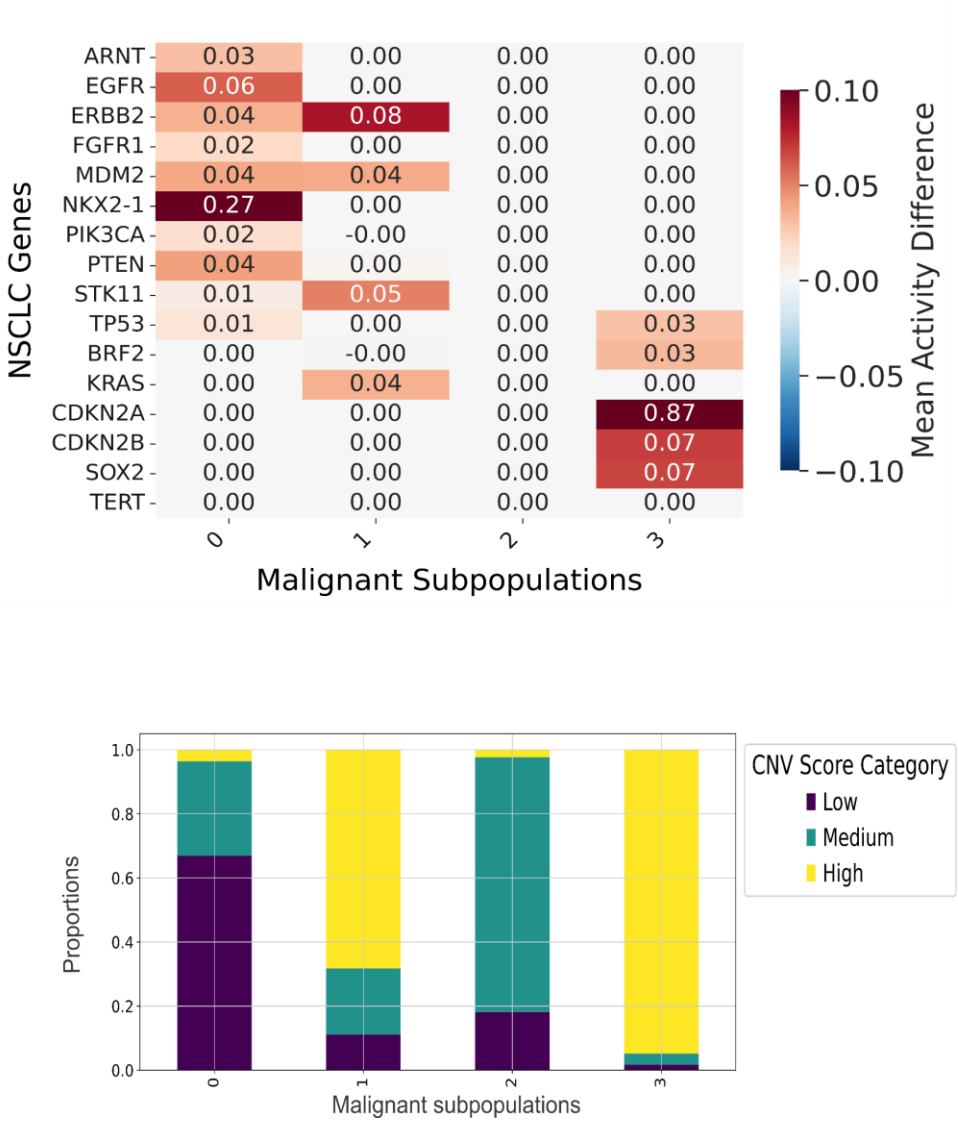


Figure S1: Expression Profile of Canonical NSCLC Driver Genes and Copy Number Variation Burden Across Malignant Subpopulations. (Top) Heatmap displaying mean activity differences in canonical non-small cell lung cancer (NSCLC) driver genes and tumor suppressors across the four malignant subpopulations (Groups 0–3) in LUAD samples. **(Bottom)** Stacked bar plot illustrating the proportional distribution of Copy Number Variation (CNV) score categories (Low, Medium, High) across the four malignant groups, highlighting elevated genomic instability in Groups 1 and 3.

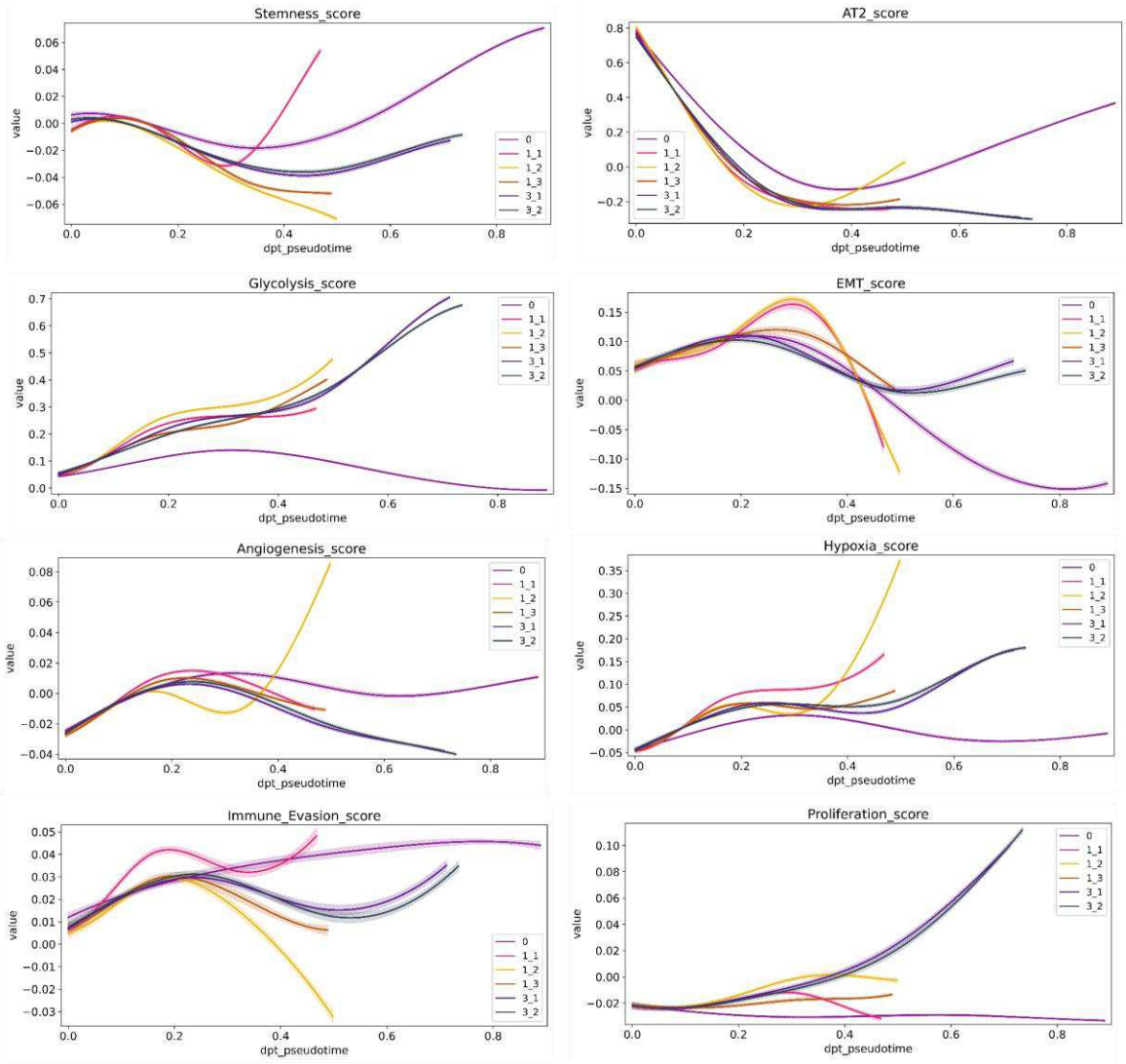


Figure S2: Dynamic Characterization of Functional Signatures and Biological Processes Across Malignant Trajectories. Lineage-specific pseudotime dynamics of key biological processes and functional scores across the identified terminal populations (0, 1_1, 1_2, 1_3, 3_1, and 3_2). Generalized Additive Models (GAMs) depict the smoothed expression and functional activity scores plotted against diffusion pseudotime (DPT), illustrating key divergent trajectories. Metrics tracked include the AT2 lineage score, stemness, glycolysis, epithelial-mesenchymal transition (EMT), angiogenesis, hypoxia, immune evasion, and proliferation scores. Trajectories highlight the specific functional characteristics of Group 1-derived populations, such as early EMT/hypoxia/angiogenesis peaks in population 1_2 and elevated stemness in population 1_1, compared to the preserved lineage identity in population 0 and proliferative focus in populations 3_1 and 3_2.

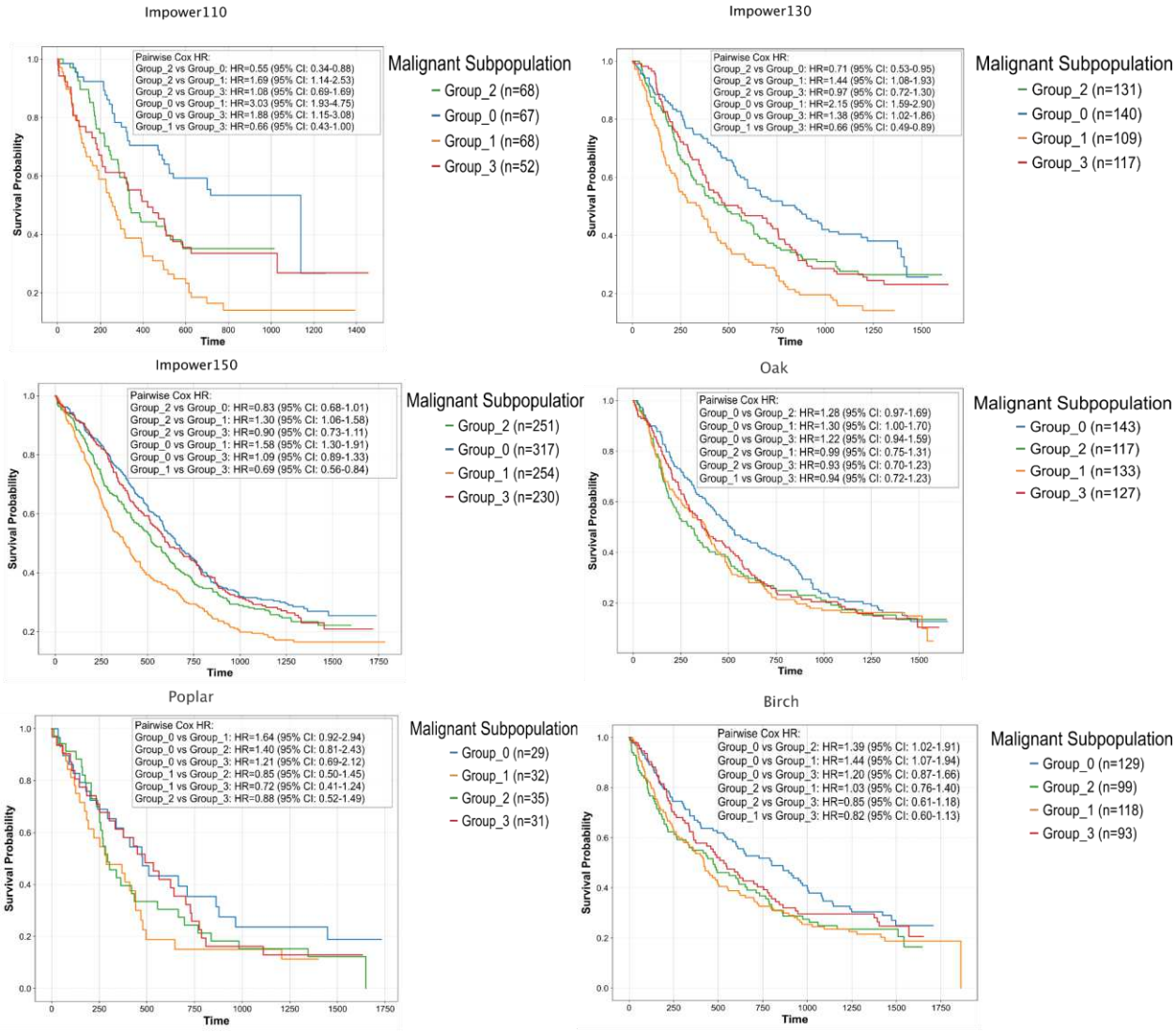


Figure S3: Prognostic Stratification of Malignant Subpopulation Signatures Across Six Clinical Trial Cohorts. Kaplan–Meier overall survival curves comparing patients assigned to dominant malignant subpopulation identities (Groups 0–3) across six clinical trial cohorts (IMpower110, IMpower130, IMpower150, OAK, POPLAR, and BIRCH), demonstrating that Group 1 consistently exhibits the poorest survival while Group 0 confers a favorable prognosis. Pairwise Cox proportional hazard ratios (HR with 95% confidence intervals) and the sample size *n* for each group are provided within each cohort inset.

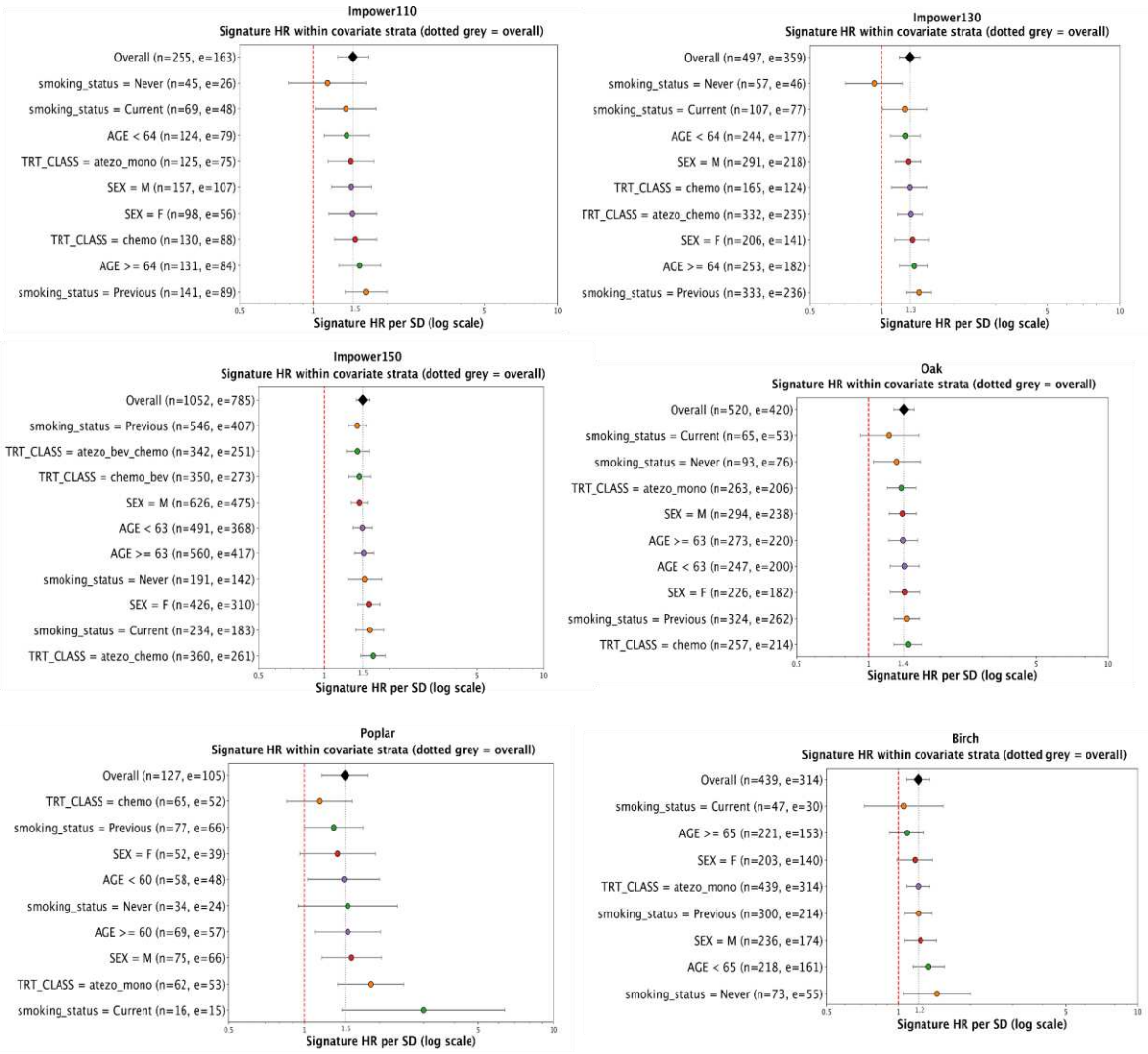


Figure S4: Subgroup Analysis Demonstrating Prognostic Independence of the Group 1 Signature Across Clinical Covariates. Forest plots displaying hazard ratios (HR per 1 SD increase on a log scale) and 95% confidence intervals for overall survival stratified by key clinico-pathological variables across six clinical trial cohorts. Stratifications include smoking status (never, current, previous), age subgroups, sex, and treatment class. The overall cohort estimate is indicated by the black diamond and dotted line, confirming that the signature's prognostic effect remains consistently robust across demographic and clinical subgroups. Sample size *n* and event count *e* are provided for each stratum.

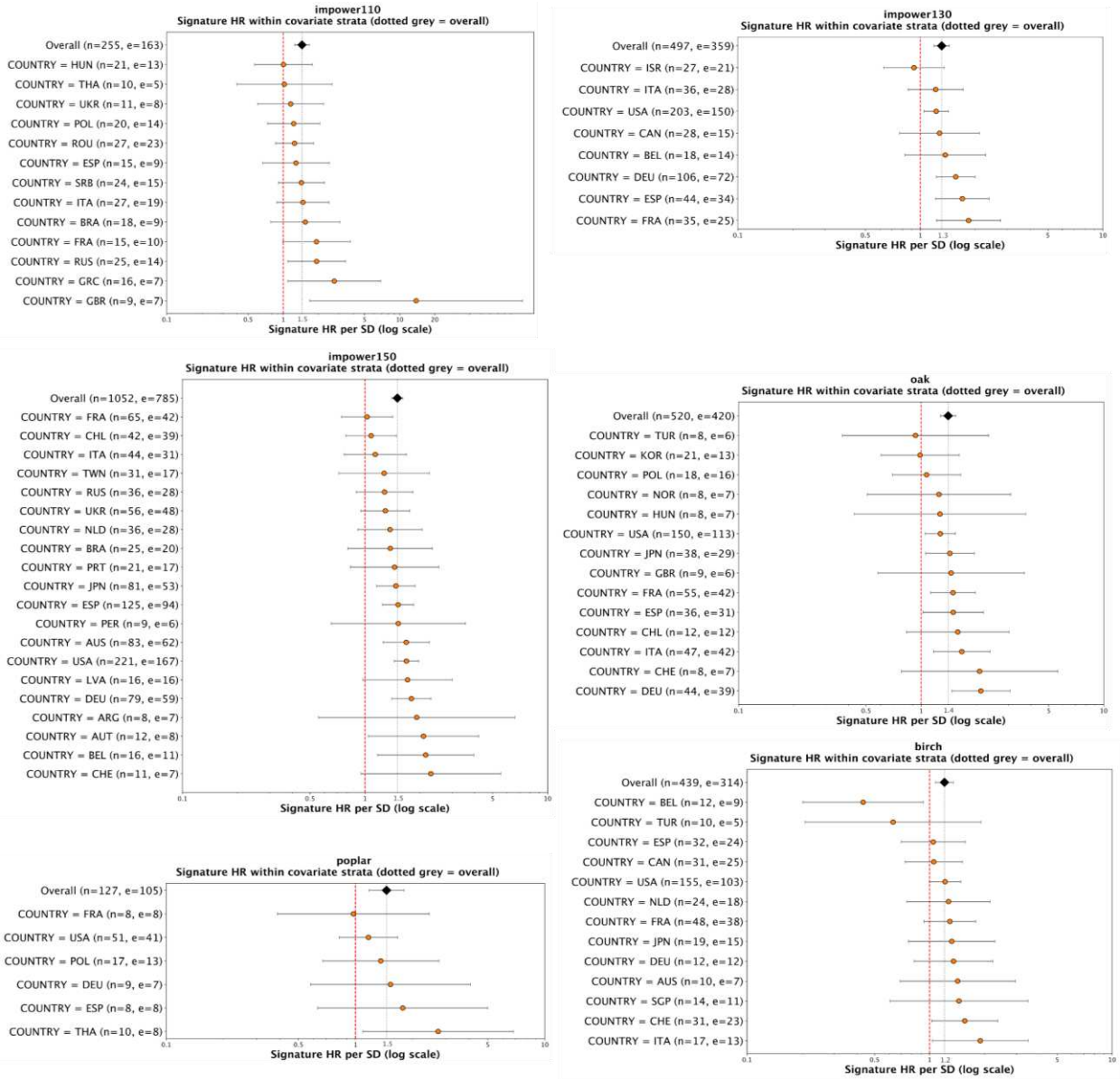


Figure S5: Geographic Independence of the Group 1 Signature Across Trial Locations. Forest plots displaying country-stratified hazard ratios (HR per 1 SD increase on a log scale) and 95% confidence intervals for overall survival across six clinical trial cohorts. The overall cohort estimate is represented by the black diamond and dotted grey line, demonstrating that geographic origin does not confound or disrupt the prognostic performance of the signature. Sample size *n* and event count *e* are provided for each country stratum.

Supplementary Files

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

- s6ComparisonofClinicaltrialsusedinthestudy.xlsx
- s1NonmalignantAT2cellmetaprogramsandGavishetal.cancercellmetaprogramsusedinanalysis.xlsx
- s5LigandReceptorresultssourceMalignantcelltargetTcellFibroblasts.xlsx
- s7multivariablecoxcovariateresults.xlsx
- s2top50DEgenespermalignantsubpopulation.xlsx
- s4signaturesusedpseudotime.xlsx
- s3Comparisonofthemalignantcellsubpopulationstopsignificantmetaprograms.xlsx