Single-cell transcriptomics reveals a conserved embryonic progenitor in different human cancers

Anna A. Khozyainova[1,†], Vera G. Subrakova[1,†], Maxim E. Menyailo[1,2], Daria I. Zhigalina[3], Tatiana N. Kireeva[3], Adelya A. Galiakberova[2], Mikhail A. Berestovoy[4], Elena E. Kopantseva[2], Anastasia A. Korobeynikova[1], Erdem B. Dashinimaev[2], Dmitry M. Loos[1], Maria S. Tretyakova[1], Ustinia A. Bokova[1], Nikolay A. Skryabin[3,†], Evgeny V. Denisov[1,2†,*]

[1]Cancer Research Institute, Tomsk National Research Medical Center, Russian Academy of Sciences, Tomsk, Russia

[2]Research Institute of Molecular and Cellular Medicine, Peoples' Friendship University of Russia (RUDN University), Moscow, Russia

[3]Research Institute of Medical Genetics, Tomsk National Research Medical Center, Russian Academy of Sciences, Tomsk, Russia

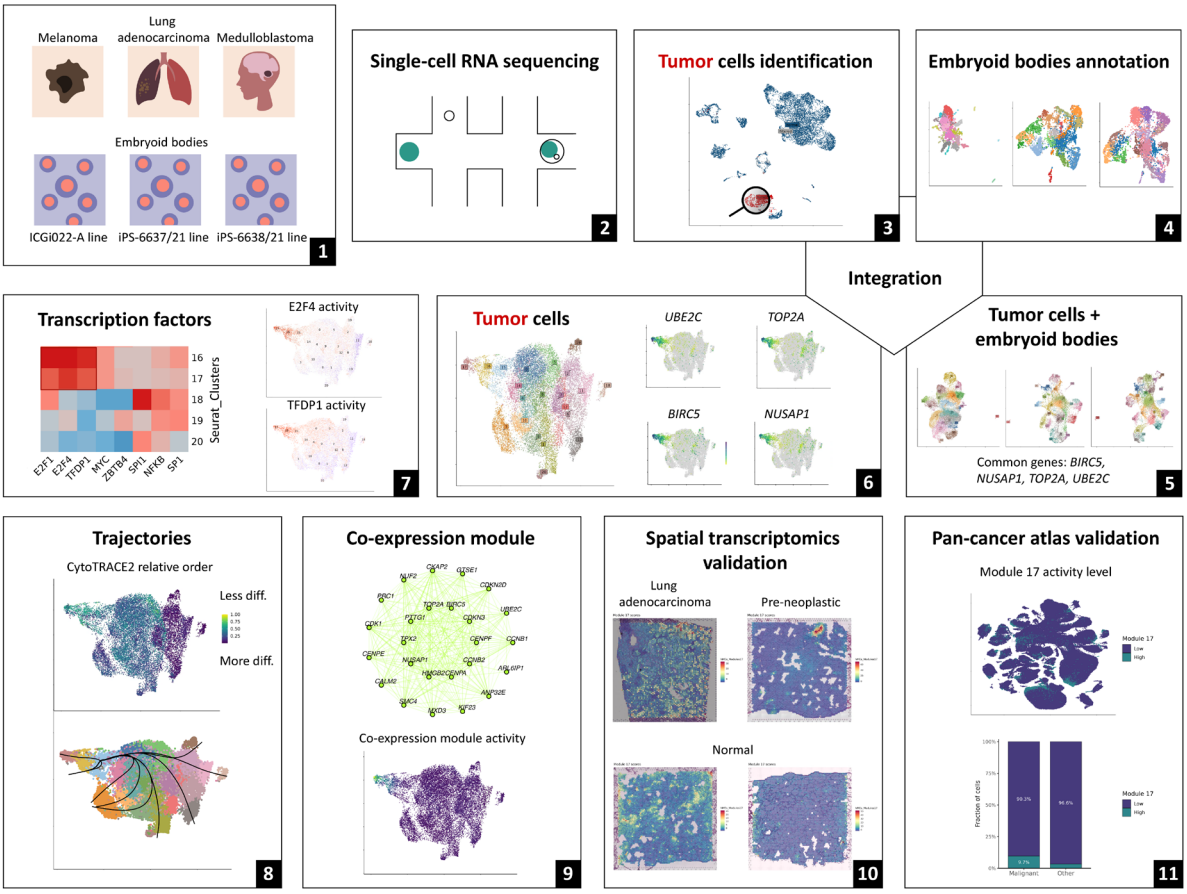
[4]Federal Center for Brain Research and Neurotechnologies of the Federal Medical and Biological Agency of Russia, Moscow, Russia

*Correspondence: d_evgeniy@oncology.tomsk.ru

[†]Equal contribution

Abstract

Cellular plasticity, a critical feature of embryonic development, is often reactivated in cancer, enabling tumor cells to acquire stem-like properties and facilitate uncontrolled growth. This study compares cellular plasticity in cancer and early human development by integrating single-cell data from three cancers and embryoid bodies. The analysis identifies a shared, highly proliferative progenitor-like state, driven by an E2F-TFDP transcriptional axis and enriched in key cell-cycle regulators *UBE2C*, *TOP2A*, *BIRC5*, and *NUSAP1*. Functionally linked to cell-cycle progression and DNA repair, this state acts as a lineage hub within tumors. Spatial transcriptomics revealed that this state is enriched in lung adenocarcinoma and in lung tissue with basal cell hyperplasia, a potential precursor to squamous cell carcinoma. Validation across eight cancer types demonstrates a prevalence of this program in tumors versus normal cells. Collectively, these findings define a conserved mechanism of tumor plasticity with potential for early detection and targeted therapy.



Introduction

Cellular plasticity is a fundamental property of embryonic development, enabling dynamic transitions between cellular states during lineage specification, morphogenesis, and organogenesis [1,2]. These processes are accompanied by the remodeling of transcriptional and epigenetic circuits and are regulated by networks of signaling cascades and transcription factors that ensure the coordination of embryonic development programs [3,4]. Contemporary single-cell and spatial transcriptomics approaches have enabled the characterization of universal transcriptional states and key regulatory axes in early human development [5,6].

Remarkably, many of the molecular circuits that govern embryonic cell plasticity are reactivated in pathological contexts, most notably in cancer. Increasing evidence suggests that oncogenesis co-opts developmental plasticity programs to fuel uncontrolled proliferation, invasion, metastasis, and therapy resistance [7,8]. This convergence highlights a profound biological parallel: the mechanisms that orchestrate normal development could be aberrantly deployed to drive malignant progression, including angiogenesis, epithelial-mesenchymal transition (EMT), acquisition of stem-like properties, and immune evasion, among other processes reminiscent of early embryogenesis [9-12].

Recently, single-cell transcriptomic studies across diverse cancer types have shown that tumor cells adopt states resembling those in normal development and remain restricted to the lineage hierarchies of their tissues of origin [13-16]. This perspective has given rise to the concept of developmental constraints, in which cancer cells, despite their genomic instability and phenotypic

flexibility, remain restricted to transcriptional states accessible within the organism's original developmental map [17]. Tumor cells do not generate arbitrary cell identities; rather, they traverse adjacent progenitor and differentiated states that reflect lineage relationships established during embryogenesis. Viewing cancer plasticity through the lens of embryonic developmental programs could offer some new insights, for example, explain why cancers typically follow lineage-dependent trajectories imposed by their tissue of origin. More importantly, this approach can reveal unique biological pathways tumor cells depend on, pointing us toward new strategies to stop tumor growth and overcome treatment resistance [11,18,19]. However, a limited understanding of human post-implantation embryogenesis that is severely hampered by the ethical and technical impossibility of direct observation or experimental manipulation at this stage represents a critical barrier to fully decoding and targeting the cancer-specific vulnerabilities [20,21].

One potential solution involves the use of embryoid bodies (EBs), which recapitulate key features of early development, including the presence of cells from the three germ layers and the expression of characteristic markers [22]. In particular, studies of cell mechanics in EBs demonstrate that upon aggregation, pluripotent cells self-organize into structured, multicellular assemblies that undergo morphogenetic processes reminiscent of early embryo development, including cellular differentiation, epithelial-mesenchymal transitions, lumen formation, and establishment of primitive tissues, resembling embryonic tissues [23]. Moreover, recent data indicate that human EBs can generate a diverse spectrum of cell types and developmental trajectories, and reproduce aspects of lineage specification and cell-type heterogeneity observed *in vivo* [24,25]. Thus, the EBs system offers a tractable, ethically acceptable, and experimentally controllable platform to model events of early human embryogenesis, bridging the gap between *in vivo* embryo research and *in vitro* cellular models. Consequently, EBs could serve as a developmental reference framework enabling direct comparison of tumor transcriptional states with those characteristic of early embryonic cell populations.

The present study aims to identify common cell states, transcriptional modules, and regulatory axes shared by embryogenesis and oncogenesis. To achieve this, we compared the transcriptional profiles of embryoid bodies and three ontogenetically distinct tumors: non-small cell lung cancer, melanoma, and medulloblastoma using single-cell RNA sequencing (scRNA-seq) integrative analysis, followed by validation via spatial transcriptomics and an independent pan-cancer scRNA-seq atlas.

Results

1. Characterization of embryoid body and tumor single-cell datasets

To identify common molecular programs shared between embryogenesis and oncogenesis, the analysis included samples representing two distinct biological contexts: normal early human development and malignant growth. As a model of early human embryogenesis, embryoid bodies were generated through spontaneous differentiation of three independent iPSC lines: ICGi022-A (sample EB1), iPS-6637/21 (sample EB2), and iPS-6638/21 (sample EB3). The presence of derivatives from all three germ layers (ectoderm, mesoderm, and endoderm) was confirmed by immunocytochemical staining (Figure 1A-C).

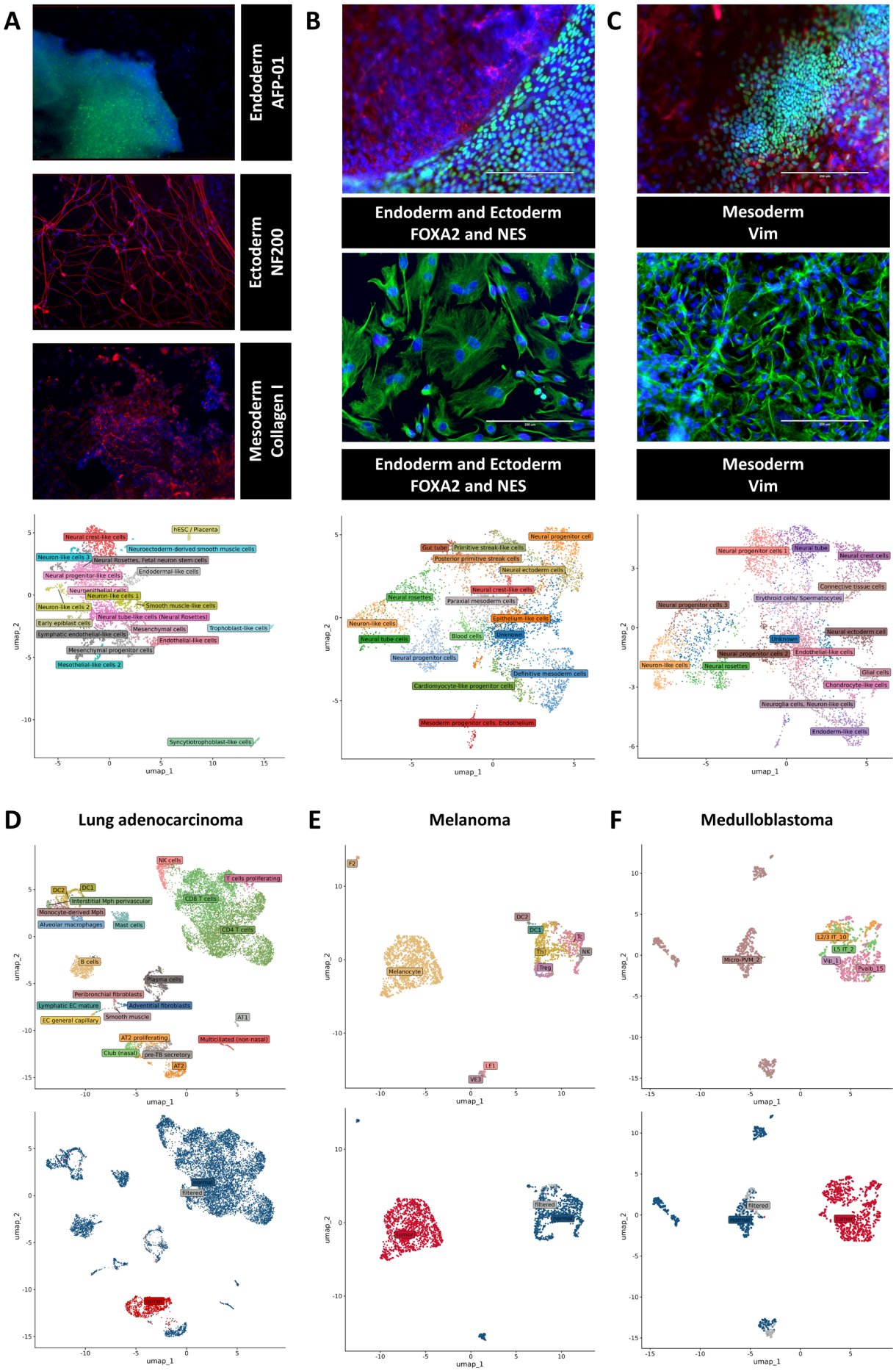

A
Endoderm AFP-01
Ectoderm NF200
Mesoderm Collagen I
B
Endoderm and Ectoderm FOXA2 and NES
Endoderm and Ectoderm FOXA2 and NES
C
Mesoderm Vim
Mesoderm Vim
D
Lung adenocarcinoma
E
Melanoma
F
Medulloblastoma
umap_1
umap_2

Figure 1. Characterization of tumor and EBs scRNA-seq datasets. Upper panel: Immunofluorescence staining of derivatives representing the three germ layers – ectoderm, mesoderm, and endoderm – in embryoid bodies generated from three independent iPSC lines: ICGi022-A (A), iPS-6637/21 (B), and iPS-6638/21 (C). Middle panel: UMAP visualization of manually annotated cell types across the three embryoid body samples. Lower panel: Representative examples of tumor cell annotation and tumor cell identification for three ontogenetically distinct cancer types. For each cancer type – lung adenocarcinoma (D), melanoma (E), and medulloblastoma (F) – the upper panels show cell type annotations generated using CellTypist, while the lower panels display tumor cell calling based on large-scale copy-number alterations inferred.

Single-cell RNA-seq profiling of all EBs provided a high-resolution view of their cellular architecture, revealing a spectrum of transcriptionally distinct populations corresponding to all three germ layers. Across samples, the EBs contained neuroectoderm-associated (including neural progenitors, neural rosettes, neuron-like and neuroepithelial-like cells), mesoderm-associated (such as mesenchymal and endothelial-like cells), endoderm-like and multiple intermediate progenitor populations (Figure 1A-C).

To identify molecular programs that are truly common between embryogenesis and oncogenesis, the analysis included three ontogenetically and clinically distinct cancer types: lung adenocarcinoma, melanoma, and medulloblastoma. Each tumor dataset was processed independently to identify major cell populations. Initial cell-type annotation was performed using the CellTypist classifier with reference models appropriate for the corresponding cancer type. Tumor cells were subsequently identified by inferring large-scale copy-number alterations, which enabled reliable distinction between malignant and normal cells. Representative examples of tumor cell identification for each cancer type are shown in Figure 1D-F.

2. Integration of embryoid bodies with different cancers reveals shared cell states

To identify shared cell states, pairwise integration was performed between each cancer type individually (melanoma, lung adenocarcinoma, and medulloblastoma) and EBs. Shared clusters were defined as those containing cells from all three EBs samples and tumor specimens within a given cancer type.

For the melanoma-EBs integration, nine clusters containing cells from all three EBs and all three melanoma samples were identified as shared: 0, 1, 2, 3, 4, 9, 12, 13, and 14 (Figure 2A). Conserved markers were examined first using the combined melanoma and EBs-derived cells, filtered with the following thresholds: adjusted p-value < 0.05, log2FC > 0.5, and pct.1 > 0.25. This analysis highlighted distinct signatures: elevated expression of *LGALS3* in cluster 4, DNA replication and repair-associated genes (*TYMS*, *PCLAF*, *DUT*, *PCNA*, *HIST1H4C*) in cluster 12, and mitotic regulators (*ASPM*, *CDC20*, *UBE2C*, *CCNB1*, *MAD2L1*) in cluster 14 (Supplementary Table 1).

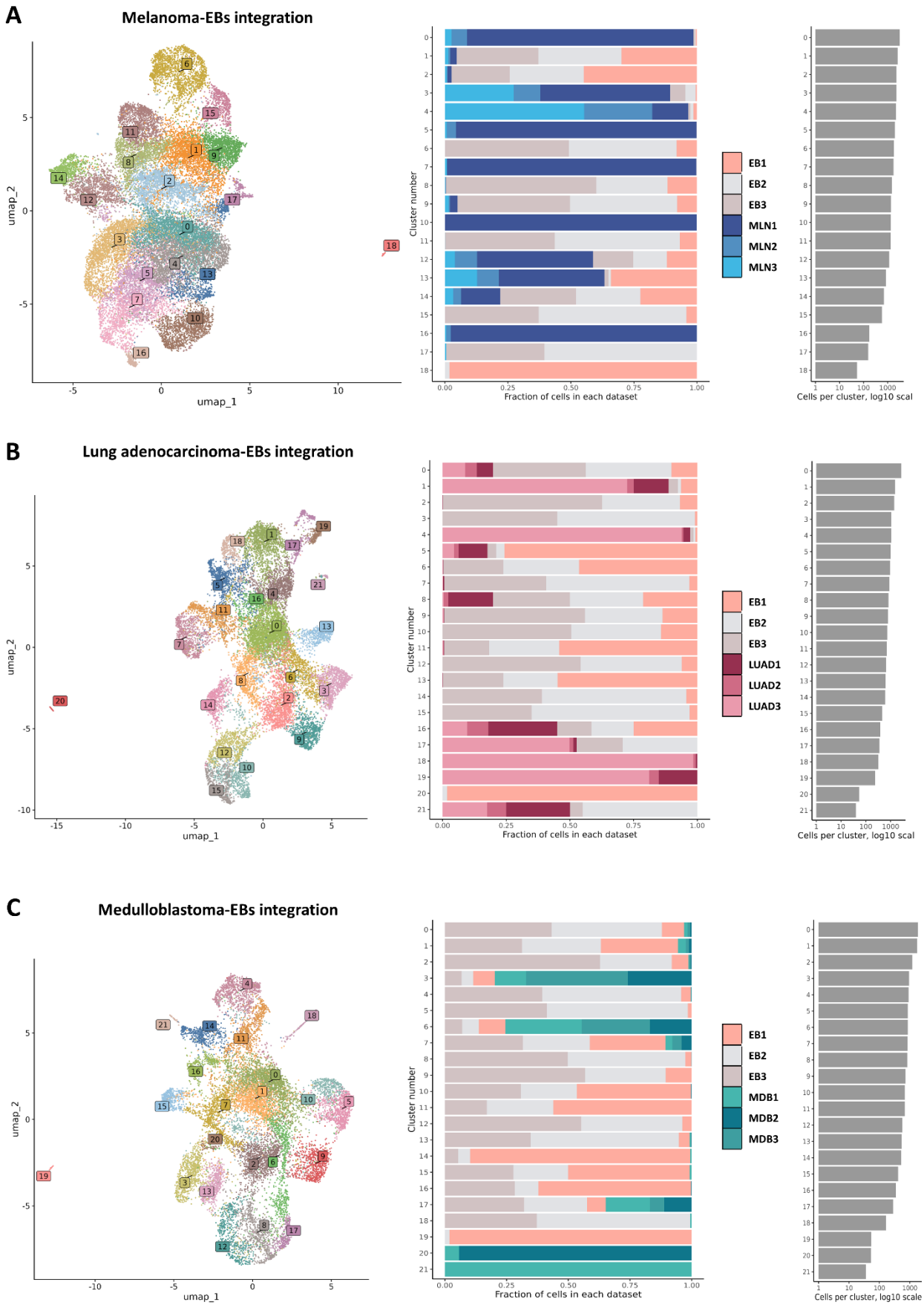


Figure 2. Integration of embryoid body-derived cells with tumor cells from three distinct cancer types. For each cancer type – melanoma (A), lung adenocarcinoma (B), and medulloblastoma (C) – three visualizations are shown. Left panel: UMAP embeddings of the integrated tumor-EBs

datasets with cluster labels. Middle panel: fraction of cells constituting the cluster by each sample. Right panel: total number of cells per cluster on a log10 scale.

To further resolve melanoma-specific conserved programs within these shared clusters, differential expression analysis was repeated using only melanoma-derived cells (adjusted p-value < 0.05, log2FC > 0.5, pct.1 > 0.5). This revealed transcriptional regulators (*ID2*, *RGS2*, *HES1*) in cluster 3, translation-associated genes (*EEF1G*, *RPS3A*, *RPL5*, *RPS4X*) in cluster 4, DNA replication factors (*RRM2*, *CDC45*, *TYMS*, *TK1*, *PCLAF*) in cluster 12, interferon-stimulated *IFIT* genes in cluster 13, and mitotic regulators (*PLK1*, *NEK2*, *DLGAP5*, *UBE2C*, *MKI67*) in cluster 14.

For the lung adenocarcinoma-EBs integration, clusters 1 and 8 contained cells from all three EBs and all three cancer samples (Figure 2B). Conserved markers were first evaluated using the combined tumor- and EBs-derived cells (adjusted p-value < 0.05, log2FC > 0.5, pct.1 > 0.25). Cluster 1 showed epithelial and adhesion-associated genes (*GRHL2*, *SVIL*), whereas cluster 8 displayed a proliferative module including *BIRC5*, *UBE2C*, *TOP2A*, *NUSAP1*, and *CDK1*.

To identify lung adenocarcinoma-specific conserved markers within these clusters, analysis was repeated using only tumor-derived cells (adjusted p-value < 0.05, log2FC > 0.5, pct.1 > 0.5). This revealed dominant expression of ribosomal genes in cluster 0, regulators (*NR3C2*, *FTX*, *ARHGAP24*, *SAMD4A*) in cluster 1, the cytoskeletal gene *FLNA* in cluster 5, and core proliferative genes (*UBE2C*, *TOP2A*, *TK1*, *ZWINT*) in cluster 8.

For the medulloblastoma-EBs integration, clusters 3, 7, and 17 contained cells from all three EBs and all three cancer samples (Figure 2C). Conserved markers were first evaluated using the combined tumor- and EBs-derived cells (adjusted p-value < 0.05, log2FC > 0.5, pct.1 > 0.25). In this analysis, cluster 3 expressed neuronal structural genes (*STMN2*, *TUBB2B*, *MLLT11*, *TUBB3*, *TUBA1A*), cluster 7 displayed mitotic and proliferative regulators (*DLGAP5*, *CENPE*, *BIRC5*, *ASPM*, *TOP2A*), and cluster 17 was marked by the synaptic adhesion gene *NRXN1*.

To identify transcriptional programs conserved specifically within medulloblastoma cells of these shared clusters, differential expression analysis was repeated using only tumor-derived cells (adjusted p-value < 0.05, log2FC > 0.5, pct.1 > 0.5). This revealed DNA replication and repair genes (*EXO1*, *UHRF1*, *CDC45*, *BRIP1*) in cluster 2, cytoskeletal and chromatin regulators (*MARCKSL1*, *TUBA1A*, *H3F3A*) in cluster 3, neuronal differentiation markers (*MYT1L*, *DACH1*, *TRIO*, *EBF1*, *DSCAML1*) in cluster 6, mitotic genes (*CDC25C*, *KIF18B*, *DLGAP5*, *KIF4A*, *ASPM*) in cluster 7, and synaptic and signaling genes (*CACNA1E*, *GABRB1*, *PPP2R2B*, *SGIP1*, *ELMO1*) in cluster 17.

Comparison of conserved DEGs across the three tumor-EBs integrations revealed a strong convergence on a common proliferative program. In analyses based on the combined tumor- and EBs-derived cells, the shared clusters consistently expressed core cell-cycle and mitotic regulators, including *UBE2C*, *TOP2A*, *BIRC5*, and *NUSAP1*. Examination of tumor-restricted conserved markers across the same set of clusters identified a partially overlapping group of genes, featuring *RPS3A*, *ZSWIM6*, *MT-ND1*, *MT-CO3*, as well as *UBE2C* and *TOP2A*, which appeared in both comparisons.

These findings demonstrate that tumor cells with distinct cellular origins converge on a conserved proliferative transcriptional module with substantial overlap to early embryonic programs.

3. Integration of different cancer types reveals embryonic-like clusters

To explore cross-tumor similarities, tumor cells from all three cancer types were subjected to a joint integration, resulting in 21 transcriptional clusters (Figure 3A). Projection of the conservative DEGs identified in the tumor-EBs integrations (*UBE2C*, *TOP2A*, *BIRC5*, and *NUSAP1*) revealed that their expression was predominantly localized to clusters 16 and 17. This indicates that these clusters represent the common proliferative compartment across lung adenocarcinoma, melanoma, and medulloblastoma (Figure 3B). Notably, both clusters contained tumor cells from all three cancer types, confirming their pan-cancer representation (Supplementary Figure 1). Given that the same proliferative signature (*UBE2C*, *TOP2A*, *BIRC5*, *NUSAP1*) was consistently detected in tumor-EBs integrative analyses, clusters 16 and 17 can be interpreted as embryonic-like proliferative states conserved across distinct tumor types.

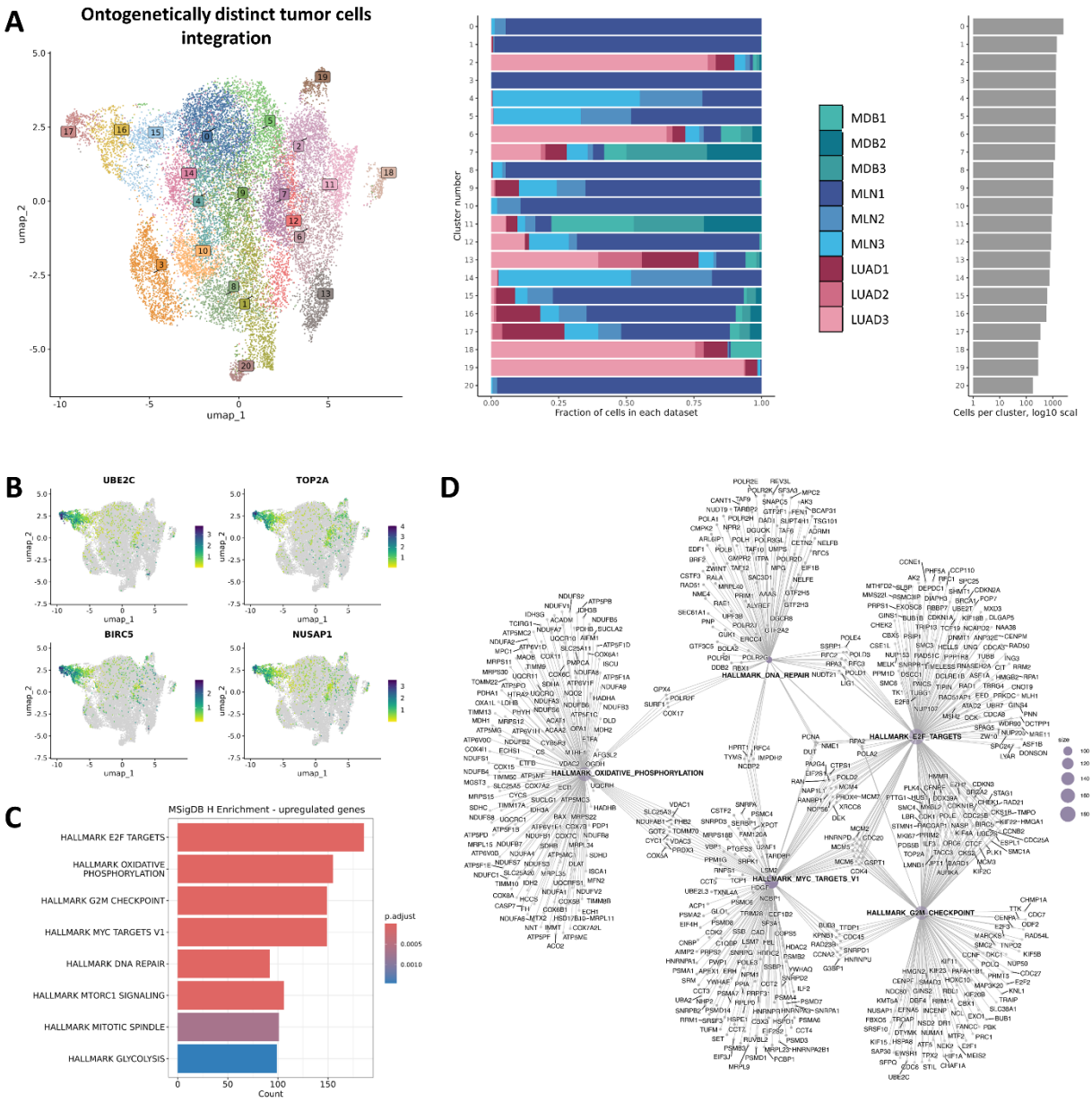


Figure 3. A shared embryonic-like proliferative program across lung adenocarcinoma, melanoma, and medulloblastoma. (A) Left: UMAP embedding of tumor cells from lung adenocarcinoma, melanoma, and medulloblastoma following joint integration, resulting in 21 transcriptional clusters. Middle: fraction of cells contributed by each tumor type to every cluster. Right: total number of cells per cluster (log10 scale). (B) Feature plots showing expression of conserved proliferative markers identified in the EB-tumor integrations (*UBE2C*, *TOP2A*, *BIRC5*, and *NUSAP1*). (C) Hallmark pathway enrichment analysis of genes jointly upregulated in clusters 16 and 17 relative to all other tumor cells. (D) Network representation of enriched Hallmark categories jointly upregulated in clusters 16 and 17, with edges reflecting shared genes between pathways.

To exclude the possibility that clusters 16 and 17 are driven primarily by effects of cell cycle, an additional control integration was performed with cell cycle-associated variation regressed out. Under these conditions, clusters corresponding to the original clusters 16 and 17 remained stable and clearly separated in the UMAP, rather than dispersing across the embedding (Supplementary

Figure 2). This observation indicates that clusters 16 and 17 are not merely a consequence of cell cycle activity, but represent stable proliferative states conserved across tumor types.

Building on these results, we next performed functional annotation of clusters 16 and 17 to clarify their biological role and regulatory dependencies.

4. Functional characterization of the embryonic-like proliferative clusters shared across distinct cancers

Functional enrichment analysis was performed separately for clusters 16 and 17 to clarify their biological features. Cluster 16 demonstrated strong enrichment for cell-cycle-associated GO categories, including chromosome segregation, nuclear and mitotic division, organelle fission, and DNA replication (Supplementary Figure 3A). KEGG pathway analysis was consistent with this proliferative profile, with cell cycle emerging as the top pathway and additional hits corresponding to replication-associated and DNA damage response pathways (Supplementary Figure 3B).

Cluster 17 showed a similarly pronounced proliferative signature. GO analysis highlighted chromosome segregation, mitotic nuclear division, sister chromatid segregation, organelle fission, and microtubule cytoskeleton organization involved in mitosis (Supplementary Figure 3C). KEGG analysis identified the cell cycle as the dominant pathway, together with enrichment in G2/M checkpoint, p53 signaling, and cellular senescence pathways (Supplementary Figure 3D).

To further characterize the shared transcriptional program of clusters 16 and 17, we performed Hallmark enrichment analysis using the genes jointly upregulated in these clusters relative to all other tumor cells (Figure 3C). The combined DEG set showed significant enrichment for HALLMARK_E2F_TARGETS, HALLMARK_G2M_CHECKPOINT, HALLMARK_MYC_TARGETS_V1, HALLMARK_DNA_REPAIR, HALLMARK_MITOTIC_SPINDLE, HALLMARK_OXIDATIVE_PHOSPHORYLATION, and HALLMARK_MTORC1_SIGNALING. To assess how these pathways relate to one another, we constructed a network representation linking enriched Hallmark categories through their shared genes (Figure 3D). The network revealed that HALLMARK_DNA_REPAIR and HALLMARK_MYC_TARGETS_V1 form major connectivity hubs, each displaying links to all other enriched Hallmark programs. In contrast, HALLMARK_E2F_TARGETS formed a strong but more selectively connected module, linked primarily to HALLMARK_G2M_CHECKPOINT and HALLMARK_MYC_TARGETS_V1, but not connected to HALLMARK_OXIDATIVE_PHOSPHORYLATION.

Together, these findings confirm that clusters 16 and 17 represent transcriptionally conserved proliferative compartments across all three cancer types.

5. E2F-TFDP transcriptional factors axis in the embryonic-like proliferative clusters 16 and 17

To identify upstream regulators of the shared proliferative program, we inferred transcription-factor activity using the decoupleR framework and ranked TFs by normalized enrichment score (NES). The activity heatmap of the top 50 most variable transcription factors revealed a clear segregation of clusters 16 and 17 from all other malignant populations (Figure 4A). This separation was driven by coordinated activation of the cell-cycle-associated TFs, most prominently E2F1,

E2F4, and their dimerization partner TFDP1, which showed the highest activity levels (NES > 1.5, $p < 0.05$) (Figure 4B-D). Elevated activity of additional proliferation-related regulators, including MYC, FOXM1, PLAGL2, and HINFP, further supports the presence of an E2F-centered transcriptional module.

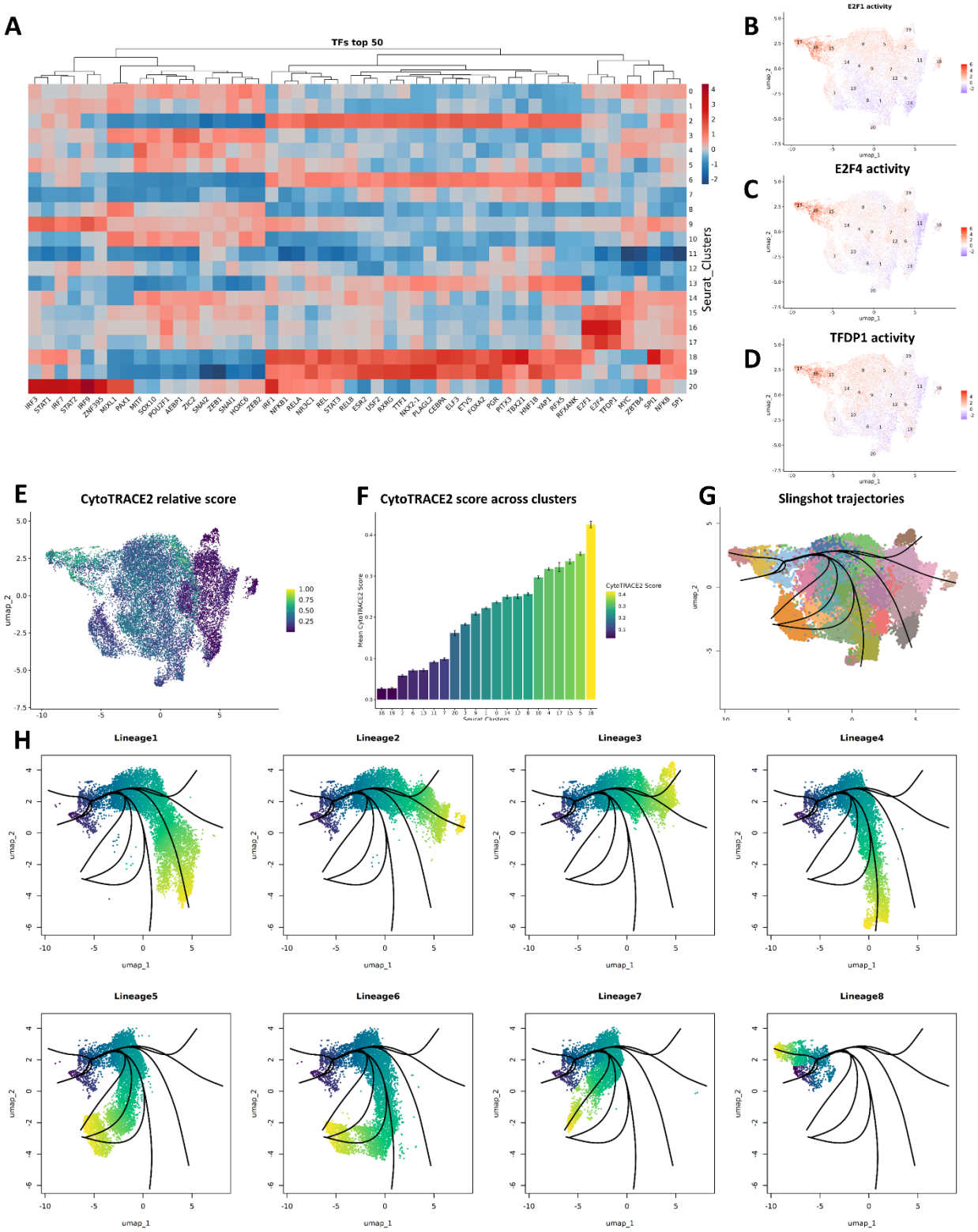


Figure 4. E2F-TFDP-driven proliferative program and pseudotime hierarchy of embryonic-like tumor states. (A) Heatmap of transcription-factor activity scores (top 50 most variable TFs) inferred using decoupleR. (B–D) UMAP projections of inferred activity for E2F1, E2F4, and TFDP1. (E) UMAP visualization of CytoTRACE2 scores across all cells. (F) Distribution of

CytoTRACE2 scores across clusters, shown as mean values with standard error bars. (G) Slingshot-inferred global lineage structure with cluster 16 identified as the principal root node. (H) Eight pseudotime lineages reconstructed by Slingshot.

In contrast, TFs associated with immune signaling (e.g., IRF, STAT) and differentiation-related factors (e.g., MITF, CEBPA, HOXC6) showed markedly reduced activity in clusters 16 and 17, underscoring the selective engagement of a cell-cycle-dominated regulatory state. Together, these findings define an upstream regulatory architecture, in which the E2F1-E2F4-TFDP1 axis forms the central driver of the proliferative identity. This transcription-factor landscape aligns closely with the Hallmark enrichment analysis, supporting the conclusion that clusters 16 and 17 represent a conserved, E2F-TFDP-driven proliferative state shared across lung adenocarcinoma, melanoma, and medulloblastoma.

6. Pseudotime trajectory inference reveals cluster 16 as one of the primary progenitor-like states

To reconstruct differentiation hierarchies within the tumor cell compartment, we first applied the CytoTRACE2 tool to rank clusters based on their developmental potential. Clusters 5, 16, and 17 exhibited the highest CytoTRACE scores (> 0.75), indicating that these clusters are in less differentiated, progenitor-like states (Figure 4E). A Kruskal-Wallis test confirmed significant differences in CytoTRACE2 scores across all clusters ($p < 2.2e-16$). Pairwise comparisons further identified clusters 5, 16, and 17 as having significantly higher CytoTRACE2 scores compared to other clusters, consistent with a progenitor-like phenotype. These clusters exhibited the highest mean CytoTRACE2 scores, supporting their characterization as less differentiated, progenitor-like states (Figure 4F, Supplementary Figure 4).

Next, we performed trajectory inference using Slingshot on the RPCA-integrated UMAP embedding, using clusters 5, 16, and 17 as potential lineage origins. Slingshot identified cluster 16 as the global root, reconstructing eight major pseudotime lineages that radiated outward along the UMAP manifold (Figure 4G, Supplementary Figure 5). These lineages exhibited smooth trajectories, spanning the full spectrum of transcriptional states observed in the tumor compartment, with both progenitor-like and more differentiated states represented.

Slingshot trajectory analysis identified eight distinct lineages within the tumor cell population, each representing a unique differentiation path (Figure 4H, Table 1). In all lineages, cluster 16 was positioned at the earliest pseudotime points, suggesting that it represents a progenitor-like state at the root of differentiation. Notably, cluster 17, despite its high CytoTRACE score, was positioned slightly downstream of cluster 16 in most trajectories, indicating it may represent a proliferative intermediate rather than an independent progenitor state.

Table 1. Slingshot-inferred pseudotime lineages across tumor cell populations

Slingshot lineage	Pseudotime cluster path
1	16 → 15 → 0 → 5 → 12 → 7 → 6 → 13

2	16 → 15 → 0 → 5 → 2 → 11 → 18
3	16 → 15 → 0 → 5 → 2 → 19
4	16 → 15 → 0 → 9 → 1 → 20
5	16 → 15 → 0 → 4 → 10 → 3
6	16 → 15 → 0 → 9 → 1 → 8
7	16 → 15 → 0 → 14
8	16 → 17

7. hdWGCNA reveals proliferation-associated module in cluster 16

To uncover coordinated transcriptional programs driving tumor cell plasticity, we applied high-dimensional weighted gene co-expression network analysis (hdWGCNA) to a highly proliferative tumor subpopulation (cluster 16) from our integrated multi-cancer scRNA-seq atlas. A signed co-expression network was constructed, resulting in 25 distinct co-expression modules (Supplementary Figure 6). Among these, Module 17 emerged as the most biologically relevant due to its strong association with known proliferation markers. This module contained classic proliferation markers such as *TOP2A*, *MKI67*, *BIRC5*, *UBE2C*, and *CENPF*.

To explore the gene network further, we examined the top 25 hub genes in this module, which were highly interconnected and included key regulators of mitosis, such as *TOP2A*, *BIRC5*, *PRC1*, *CDK1*, *CENPF*, *CENPE*, *SMC4*, and *KIF23*, all of which are crucial for mitotic spindle function, chromosome alignment, segregation, and overall cell cycle progression (Figure 5A). We further explored the functional significance of this gene set and found that it was strongly enriched for biological pathways related to Mitotic Spindle, G2/M checkpoint, and E2F targets, as defined by the Hallmark MSigDB 2020 (Supplementary Figure 3E). To confirm the module's association with a proliferative state, we performed UCell-based scoring of the 25 hub genes. This analysis revealed that these genes exhibited strong activity in cluster 16 and, apart from that, the module was highly active in cluster 17 (Figure 5B). Consistent with these findings, a UMAP projection of Module 17 activity further showed that its signal is spatially concentrated almost exclusively within clusters 16 and 17, reinforcing the conclusion that Module 17 defines a tightly restricted, highly proliferative embryonic-like tumor state (Figure 5C).

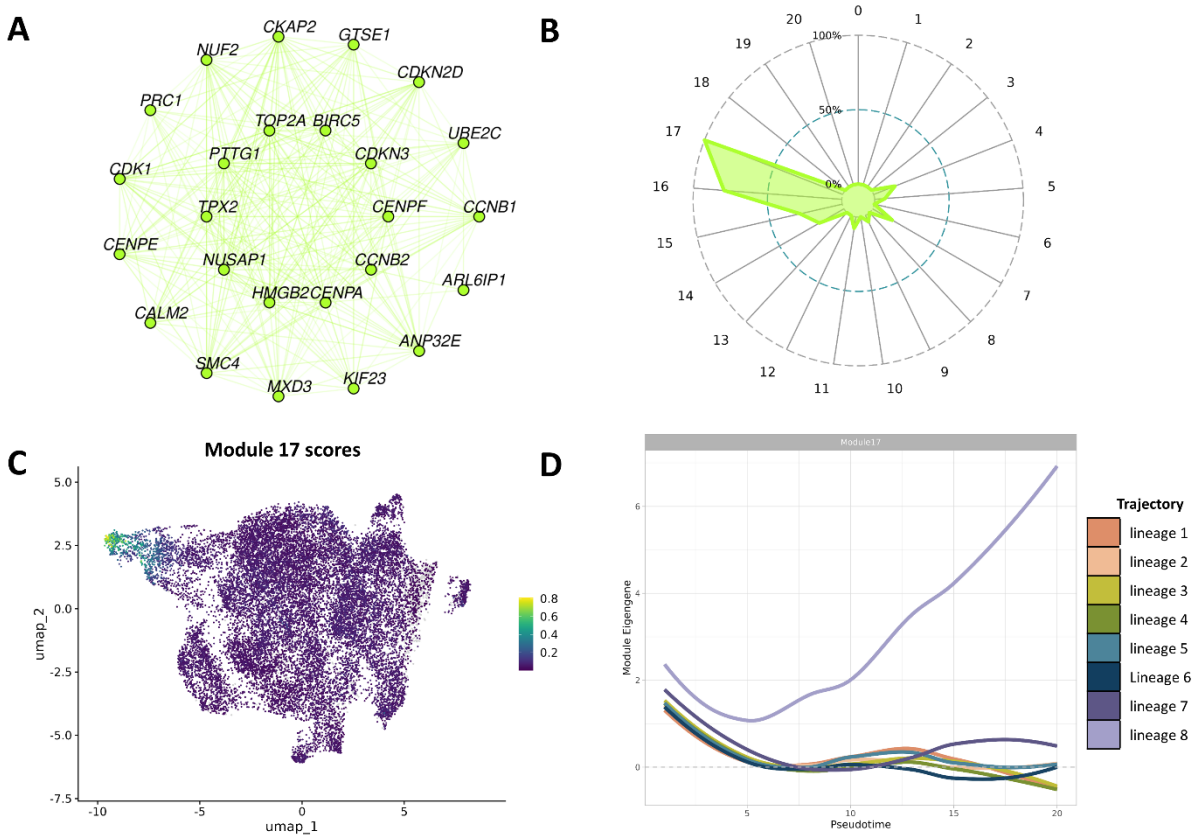


Figure 5. Structure, functional enrichment, and dynamic behavior of co-expression Module 17. (A) Network representation of the top 25 hub genes. (B) Module 17 activity across tumor cell clusters (UCell scoring). (C) UMAP projection of Module 17 activity scores (UCell scoring). (D) Changes in Module 17 activity along pseudotime.

In addition, we evaluated the expression of the co-expression Module 17 in the context of pseudotime analysis. Trajectory analysis using Slingshot, initiated from high-CytoTRACE clusters (including cluster 16), revealed that the activity of Module 17 peaked early in pseudotime and progressively declined along differentiation lineages (Figure 5D). This pattern is consistent with the transient activation of a proliferative progenitor-like state during tumor cell plasticity, suggesting that Module 17 plays a key role in regulating early-stage tumor cell proliferation.

8. Spatial validation of pan-cancer embryonic-like tumor cell states

To explore the spatial organization and microenvironmental context of the highly proliferative embryonic-like tumor cell states corresponding to clusters 16 and 17 in various cancers, we transferred cluster labels onto Visium spatial transcriptomics datasets from lung adenocarcinoma samples and healthy lung tissue from patients with no known disease (non-smoker male, ex-smoker male, and smoker female).

In the lung adenocarcinoma No.3 Visium sample, anchor-based label transfer (FindTransferAnchors followed by TransferData) revealed a consistent, spatially restricted distribution of clusters 16 and 17. Prediction scores indicated significant enrichment of these proliferative states within rosette-like epithelial structures with expanded luminal spaces (Figure

6A). Histologically, these structures were characterized by enlarged lumina, often filled with homogeneous eosinophilic material resembling edematous fluid, sometimes mixed with fresh erythrocytes. The tumor cells in these regions appeared morphologically similar to adjacent cells but were slightly smaller with occasional hyperchromatic nuclei. The morphological distinction was more prominent than the cytological one: adjacent tumor cells typically formed solid nests or classical rosettes, while regions associated with clusters 16 and 17 displayed rosettes with wider luminal spaces and looser epithelial arrangements along their margins.

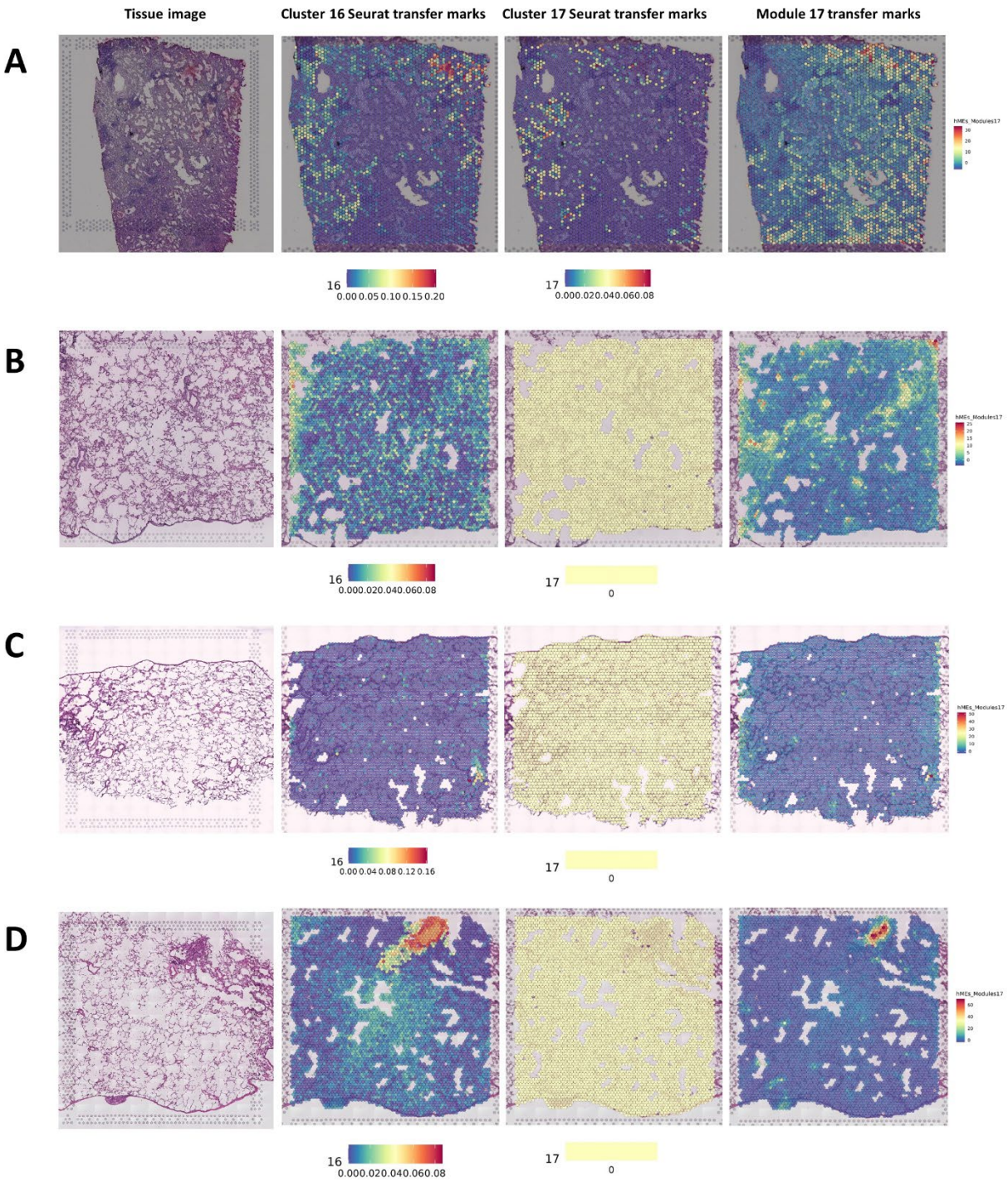


Figure 6. Spatial distribution of proliferative states in lung adenocarcinoma and healthy lung samples. A–D: Tissue images of lung samples (first column) from different conditions: (A) Lung

adenocarcinoma No.3 tissue image, (B) Healthy lung tissue from a non-smoker male, (C) Healthy lung tissue from an ex-smoker male, (D) Healthy lung tissue from a smoker female. The second, third, and fourth columns display the spatial localization of cluster 16 transfer marks, cluster 17 transfer marks, and Module 17 transfer marks, respectively.

Although these structures are not typically considered distinct morphological subtypes in diagnostic practice, their consistent spatial co-localization with transcriptionally defined proliferative modules suggests that they may represent a previously unrecognized subpopulation within the malignant compartment. This finding, combined with both molecular and morphological divergence from surrounding tumor tissue, highlights these regions as a potential manifestation of tumor cell plasticity, warranting further investigation.

Additionally, data from the lung adenocarcinoma No.3 sample was integrated with the hdWGCNA results to visualize the spatial distribution of the Module 17. We projected the identified co-expression modules identified previously onto the Visium dataset, allowing us to visualize the spatial localization of Module 17 activity. The spatial projection revealed that Module 17 was localized near the regions marked by cluster 16 transfer. This spatial distribution confirmed the restricted activation of the proliferation-associated program in specific tumor regions, supporting the hypothesis that Module 17 is associated with cluster 16 (Figure 6A). Taken together, these findings indicate that Module 17 is intricately linked to the proliferative activity of tumor cells, reinforcing its role in tumor cell plasticity and progression. The spatially restricted expression of Module 17 in proliferative regions highlights its potential as a molecular signature of cell division and tumor growth.

In two healthy lung Visium datasets (non-smoker male and ex-smoker male), prediction scores for clusters 16 and 17 were either low or undetectable (Figure 6B,C). This suggests that these tumor cell states are not present under physiological conditions or as a result of chronic smoking alone. However, in the smoker female dataset, both cluster 16 and Module 17 were localized near each other, with elevated prediction scores (Figure 6D). Importantly, histologic assessment of the same tissue revealed basal cell hyperplasia, a reactive expansion of the airway basal cell compartment that has been described as a putative preneoplastic lesion in the bronchial epithelium.

In summary, integration of spatial transcriptomics with the multi-cancer single-cell reference revealed recurrent, spatially confined tumor cell states (clusters 16 and 17 or Module 17) across lung adenocarcinoma. These states, characterized by high proliferative and plastic potential, are preferentially localized to invasive margins and microenvironmental stress niches, offering new insights into tumor biology and potential therapeutic targets.

9. Validation of Module 17 in a pan-cancer scRNA-seq dataset

A publicly available pan-cancer single-cell RNA-seq atlas (Zenodo record 14511579) was used for external validation of Module 17, the highly proliferative embryonic-like transcriptional program. The dataset includes 355 941 tumor and normal cells across eight tumor types: melanoma, melanoma brain metastasis, basal cell carcinoma, triple-negative breast cancer, HER2+ breast cancer, ER+ breast cancer, clear-cell renal cell carcinoma, hepatocellular carcinoma, and

intrahepatic cholangiocarcinoma. The atlas was provided as a preprocessed Seurat object with cell-type annotations, malignant/non-malignant classification, and UMAP coordinates (Figure 7A, B).

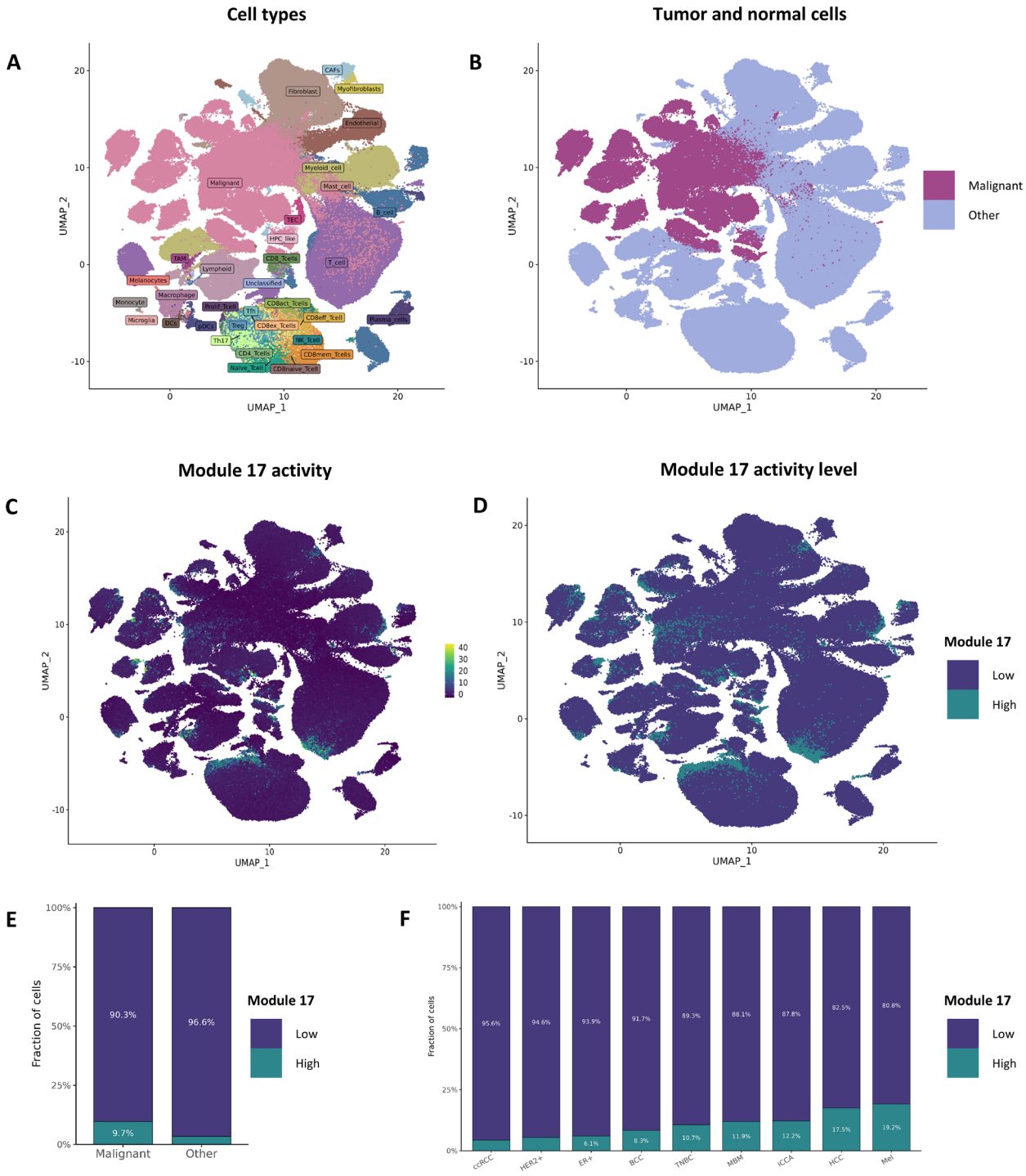


Figure 7. Validation of Module 17 activity in a pan-cancer single-cell RNA-seq atlas. (A) Detailed cell type annotation. (B) Tumor and normal cells. (C) Module 17 activity across the pan-cancer atlas. (D) Module 17-high and Module 17-low cells based on a 95th percentile threshold of Module 17 activity. (E) Fraction of Module 17-high and Module 17-low cells among tumor and normal cells. (F) Fraction of Module 17-high and Module 17-low cells across cancer entities. ccRCC – clear-cell renal cell carcinoma; HER2+ – HER2+ breast cancer; ER+ – ER+ breast cancer; BCC – basal cell carcinoma; TNBC – triple-negative breast cancer; MBM – melanoma brain metastasis; iCCA – intrahepatic cholangiocarcinoma; HCC – hepatocellular carcinoma; Mel – melanoma.

Module 17 eigengene values were calculated for each cell in the pan-cancer atlas, and their distribution was used to define a high Module 17 score (Figure 7C). A threshold at the 95th percentile of Module 17 values was applied, and cells exceeding this cutoff were classified as Module 17-high. This provided a consistent criterion for identifying cells with the strongest activation of the Module 17 program.

Module 17-high cells were detected in both tumor and normal cells within the pan-cancer atlas (Figure 7D). At the same time, their representation differed between these populations, with a higher fraction of Module 17-high cells observed among tumor cells compared with normal cells (Figure 7E; Fisher's exact test: odds ratio 3.04, 95% CI 2.95-3.13, $p < 2.2\times10^{-16}$). Importantly, Module 17-high cells were present in all cancers included in the atlas, although their relative abundance varied between cancer types (Figure 7F). In addition, Module 17-high cells were not evenly distributed across cellular populations, being more frequently detected in proliferative cell states and only rarely observed in most differentiated cell types (Supplementary Figure 7).

To identify conserved transcriptional features distinguishing tumor Module 17-high cells from normal Module 17-high cells across cancer types, differential expression analysis was performed with adjusted p-value < 0.05, log2FC > 0.25, and a detection ratio (pct.1/pct.2) ≥ 2 and identified fourteen genes (*EFNA1*, *ERBB3*, *HES1*, *YAP1*, *MDK*, *CTTN*, *TOM1L1*, *RTKN*, *PFN2*, *TSPAN6*, *S100A13*, *S100A16*, *TUSC3* and *CRNDE*).

Taken together, Module 17 activity is present in both normal and tumor cells across multiple cancer types, indicating that this transcriptional program is not restricted to malignant cells. However, cells with high Module 17 activity are significantly more frequent among tumor cells and are characterized by enrichment of specific genes.

Discussion

Oncogenesis substantially mirrors aspects of embryogenesis demonstrating enhanced proliferation, migration, plasticity, and immune evasion [18,26,27]. Putative regulators shared between oncogenesis and embryogenesis, include members of the paired box (PAX) transcription factor family, the DUX4 transcription factor, and the Kaiso and noncanonical Wnt pathways [26-28]. However, these insights have been derived from indirect comparisons rather than from analyses integrating embryonic and tumor cells within a unified transcriptional framework. In this study, a direct integrative analysis was performed by comparing transcriptional profiles of lung adenocarcinoma, melanoma, and medulloblastoma cells with embryoid body-derived cells, which serve as a model of early human development. This approach enabled the identification of tumor progenitor-like states, provided insight into mechanisms of cellular plasticity, and allowed the discovery of their potential regulators.

Integrating the tumor cells from each cancer type with embryoid body-derived cells resulted in clusters containing cells from both biological entities and characterized by elevated expression of the *UBE2C*, *TOP2A*, *BIRC5*, and *NUSAP1* genes. Notably, in the subsequent integration performed using tumor cells only, expression of *UBE2C*, *TOP2A*, *BIRC5*, and *NUSAP1* was largely confined to a limited number of tumor cell populations, interpreted in this study as highly proliferative embryonic-like population (clusters 16 and 17). Elevated *NUSAP1* levels are linked to chemoresistance, increased tumor size, and lymphatic metastasis [29,30]. High *TOP2A* expression

correlates with enhanced metastatic potential, EMT, and reduced distant metastasis-free survival [31,32]. *BIRC5* is upregulated in a wide range of cancers and is consistently associated with poorer overall survival [33]. *UBE2C* overexpression predicts shorter disease-free survival, promotes proliferation, and is linked to poor differentiation and lymph node invasion [34,35]. Notably, such associations extend to cancer types beyond those analyzed here, including chronic lymphocytic leukemia, luminal breast cancer, thyroid carcinoma, and oral squamous cell carcinomas, further underscoring the broad clinical relevance of these cell-cycle-associated genes.

Transcription-factor activity analysis identified three key regulators of these embryonic-like tumor cell populations: E2F1, E2F4, and TFDP1. These findings align with recent evidence implicating the E2F–TFDP axis as a central regulatory hub of proliferative and progenitor cell programs in both physiological and pathological contexts. For instance, TFDP1 was shown to be essential for proliferation of human hematopoietic stem and progenitor cells (HSPCs), with E2F4 identified as its binding partner required for post-transplant hematopoiesis [36]. In malignancies, elevated E2F4 and TFDP1 expression marks cancer stem-like subpopulations and correlates with chemoresistance in colorectal cancer [37]. In hepatocellular carcinoma, E2F4 promotes proliferation, invasiveness, and upregulates key cell-cycle genes including *TOP2A* and *BIRC5,* the same genes enriched in the proliferative clusters that are common between cancers and EBs [38]. In glioma stem cells, E2F4 regulates DNA-repair and replication-stress pathways via enhancer activation, underpinning enhanced survival and self-renewal [39].

A highly proliferative embryonic-like population (cluster 16) has been identified as a primary progenitor-like state across different cancer types, with CytoTRACE2 scores indicating a less differentiated phenotype and progenitor-like state in the cluster. Slingshot trajectory inference suggested cluster 16 as a critical lineage origin for tumor differentiation, whereas high-dimensional weighted gene co-expression network analysis revealed co-expression Module 17 with 25 hub genes, including *TOP2A*, *NUSAP1*, *BIRC5*, and *UBE2C* described above. Interestingly, the activity of Module 17 peaked early in pseudotime and declined as trajectories differentiated, reinforcing the idea that this module represents a dedifferentiated cell state that can be reactivated or sustained during tumor progression, contributing to the maintenance of a pool of proliferative progenitor-like cells.

Importantly, validation in an independent pan-cancer scRNA-seq atlas supports the broader relevance of the identified Module 17 program. The Module 17 activity in both normal and tumor cells suggests that this transcriptional program reflects a conserved cellular state rather than a tumor-specific signature. At the same time, tumor cells with high Module 17 activity are characterized by the enrichment of *EFNA1*, *ERBB3*, *HES1*, *YAP1*, and *MDK* genes, associated with tumor growth and proliferation [40-44], *CTTN*, *TOM1L1*, *RTKN*, *PFN2*, and *TSPAN6* genes, involved in cell motility and invasive behavior [45-49] , and *S100A13* and *S100A16* genes, implicated in angiogenesis [50,51]. Together, these observations indicate that, although Module 17-associated program may reflect a broadly conserved cellular state, this is preferentially activated and functionally repurposed in malignant contexts, probably contributing to oncogenesis.

Clusters 16 and 17 and Module 17 were predominantly detected in tumor tissue and were largely absent from normal physiological conditions, with the exception of the smoker female lung tissue dataset, where they were concentrated in a region characterized by basal cell hyperplasia, a lesion often preceding squamous cell carcinoma of the lung [52]. The previous study showed that in airway epithelium carcinogens provoke competition among basal cells, leading to clonal

expansions and an increase in cells in transitional state, decreased luminal differentiation, ultimately driving preinvasive lesions and following malignant transformation [53]. Given this data, the detection of Module 17 may mark a pre-malignant, clonally expanding, progenitor-like compartment in the airway epithelium long before malignancy is morphologically evident.

This study has several limitations. First, EBs recapitulate major features of early lineage differentiation, but do not fully reproduce morphogenetic processes of post-implantation embryogenesis. However, ethical and regulatory constraints substantially limit the use of human embryos for single-cell analysis at comparable developmental stages. Future studies could address these limitations by employing more advanced stem cell-derived models, such as gastruloids, blastoids, and synthetic embryo models, which offer greater fidelity to early embryonic architecture and morphogenesis [24]. These systems could be integrated in similar experimental frameworks to further validate and extend our findings. Second, this work represents a pilot integrative analysis focused on three cancer types: lung adenocarcinoma, melanoma, and medulloblastoma. Although these malignancies differ substantially in their biological origins, expanding the analysis to a broader spectrum of adult and pediatric tumors, as well as to larger cohorts, will be essential to determine the generality and tissue-independence of the conserved proliferative program and its regulatory axis.

In conclusion, our integrative pan-cancer single-cell analysis demonstrates that tumor cells reactivate an embryonic-like proliferative program centered on the E2F-TFDP axis and encoded by a conserved co-expression module enriched for mitotic and cell-cycle regulators. This program, represented most prominently by co-expression of 25 signature genes (module 17), appears to function as a progenitor-like hub state, from which diverse tumor lineages emerge. Its presence across ontogenetically distinct cancers, together with its spatial enrichment in pre-malignant airway lesions, suggests that this transcriptional module may represent a fundamental, tissue-independent axis of tumor cell plasticity and a promising candidate for future efforts in early detection, risk stratification, and therapeutic targeting.

Methods

Embryoid bodies preparation

EBs from ICGi022-A cell line

EBs were obtained from induced pluripotent stem cells (iPSCs) of a healthy individual (ICGi022-A cell line), provided by the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk [54]. PSCs colonies of ICGi022-A cell line were detached using 0.15% type IV collagenase (Thermo Fisher Scientific, USA) and then transferred to 30-mm dishes coated with 1% agarose to initiate the process of undirected differentiation. These iPSCs were cultivated for 18 days in an iPSC medium devoid of growth factor basic fibroblast growth factor (bFGF): Dulbecco's Modified Eagle Medium (DMEM)/Nutrient Mixture F-12 (Thermo Fisher Scientific, USA) supplemented with 15% Knockout Serum Replacement (KoSR), 1× MEM Non-Essential Amino Acids Solution (Gibco, USA), 0.1 mM 2-mercaptoethanol (Gibco, USA), 1× penicillin-streptomycin (Gibco, USA), and 1× GlutaMAX.

EBs from iPS-6638/21 and iPS-6637/21 cell lines

iPS-6638/21 and iPS-6637/21 cell lines were derived from healthy female and male donor fibroblasts using Sendai virus reprogramming kit. PSCs were seeded at a density of 2000 cells/cm² on 6 cm Petri dishes coated with 2% Matrigel in mTeSR Plus medium supplemented with 5 µM ROCK Inhibitor Y-27632 (STEMCELL Technologies, Canada). Once the iPSCs formed individual small colonies, the dishes were incubated with a dispase solution (3 mg/ml in DMEM/F12; HiMedia, India) at 37°C for 20-30 minutes to detach the colonies as intact aggregates. The detached colonies were transferred to a 15 ml tube with DPBS without calcium and magnesium and centrifuged at 100 g for 1-2 min. The pellet was carefully resuspended in mTeSR Plus and transferred to ultra-low binding Petri dishes for suspension culture in mTeSR Plus supplemented with 5 µM ROCK Inhibitor Y-27632 (STEMCELL Technologies, Canada). A partial medium change was performed every other day, gradually replacing mTeSR Plus with embryoid body medium including DMEM/F12 (PanEco, Russia), 20% FBS (Capricorn Scientific, Germany), 2 mM Glutargo (Capricorn Scientific, Germany), 1 mM Sodium Pyruvate (Gibco, USA), 0.1 mM β-Mercaptoethanol (Sigma-Aldrich, USA), 100x NEAA (Capricorn Scientific, Germany), and PenStrep (50 U/ml; 50 µg/ml; Gibco, USA).

Immunofluorescence staining of EBs

EBs of ICGi022-A cell line were transferred to gelatin-coated 6-well plates with mouse embryonic fibroblasts for an additional eight days. EBs were fixed in 3% paraformaldehyde and incubated with primary antibodies overnight at 4 °C. Following washes with PBS containing 0.1% Tween-20, samples were incubated with Alexa Fluor-conjugated secondary antibodies diluted in blocking buffer. Nuclei were counterstained with DAPI (Sigma-Aldrich, USA) in antifade mounting medium (Vectashield, Vector Laboratories, USA). Fluorescence images were acquired using an Axio Imager Z2 microscope (Zeiss, Germany) equipped with ISIS software (MetaSystems, Germany).
After 14 days of suspension culture, EBs from lines iPS-6638/21 and iPS-6637/21 were collected and plated into wells of a 24-well plate, pre-coated with 2% Matrigel, at a density of 6 EBs per well in the embryoid body medium. The medium was changed 2-3 times per week. Following additional 14 days of spontaneous differentiation, the resulting cultures were fixed for subsequent immunocytochemical analysis as described above. The antibodies used for immunocytochemistry are listed in Supplementary Table 1.

Generation of single-cell suspensions

EBs from ICGi022-A cell line

The EBs were transported on ice in 2 mL of DMEM/F12 culture medium (Thermo Fisher Scientific, USA) supplemented with 20% fetal bovine serum (FBS) (Global Kang Biotechnology, China). The experimental sample consisted of approximately 20-25 EBs. The EBs were centrifuged at 300g for 10 min, and the supernatant was removed and replaced with 5 mL of DMEM/F12 containing the enzyme cocktail from the Human Tumor Dissociation Kit (100 µL Enzyme H, 50 µL Enzyme R, and 12.5 µL Enzyme A; Miltenyi Biotec, Germany). The EBs were incubated for 30 min at 37 °C in the gentleMACS Octo Dissociator with Heaters (Miltenyi Biotec, Germany). After dissociation, the cell suspension was passed through a 70-µm cell strainer, centrifuged at 300g for 5 min, and resuspended in 1 mL of fresh DMEM/F12. Cells were counted

using 0.4% Trypan Blue Solution (Thermo Fisher Scientific, USA) on a LUNA-II Automated Cell Counter (Logos Biosystems, South Korea). The experimental sample yielded approximately 1×10^6 cells, with a viability of ~80%. The scRNA-seq experiment included 16,500 cells.

EBs from iPS-6638/21 and iPS-6637/21 cell lines

The EBs were transported on ice in 2 mL of the embryoid body medium. Each experimental sample contained approximately 20-25 EBs. The EBs were centrifuged at 300g for 5 minutes, and their culture medium was replaced by 2 mL of DMEM High Glucose (ServiceBio, China) with the enzyme cocktail from the Human Tumor Dissociation Kit (100 µL Enzyme H, 50 µL Enzyme R, 12.5 µL Enzyme A; Miltenyi Biotec, Germany). The EBs were incubated for 30 minutes at 37 °C and 1400 rpm in the thermoshaker TS-100C Smart (Biosan, Latvia). After the incubation, the resulting dissociated cells were sifted through the 40-micron cell strainer, centrifuged at 300g for 5 minutes, and the cell medium was replaced by fresh 1 mL of DMEM High Glucose (ServiceBio, China). The cells were counted using an acridine orange/propidium iodide stain buffer (Logos Bioscience, Korea) on LUNA-FL Dual Fluorescence Cell Counter (Logos Bioscience, Korea). The two experimental samples had around 10^6 cells with approximately 71% viability. The scRNA-seq experiment included 16500 cells.

Single-cell RNA sequencing library preparation

The single-cell transcriptome libraries were prepared according to the standard protocol by the Single Cell 3′ Reagent Kit v3 (10x Genomics, USA). GEM generation was performed on a Chromium iX/X instrument (10x Genomics, USA) for the EBs samples from iPS-6638/21 and iPS-6637/21 lines and on a Chromium Controller (10x Genomics, USA) for the ICGi022-A EBs sample. Reverse transcription and cDNA amplification were carried out using a QuantGene 9600 thermal cycler (Bioer, China) for the EBs samples from iPS-6638/21 and iPS-6637/21 lines, and a SimpliAmp Thermal Cycler (Thermo Fisher Scientific, USA) for the EBs sample from ICGi022-A line. Final libraries were diluted to 3.6 pM and sequenced on the GenolabM platform (GeneMind, China) with the following run configuration: Read 1 – 28 cycles, Read 2 – 90 cycles.

Publicly available scRNA-seq datasets

Raw sequencing data used in this study were obtained from the Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra). scRNA-seq data for lung adenocarcinoma were retrieved from the SRA Study SRP347260 and included samples SRR17008551, SRR17008552, and SRR17008559. Medulloblastoma scRNA-seq data were obtained from the SRA Study SRP274357, comprising samples SRR12354570, SRR12354583, and SRR12354584. Melanoma scRNA-seq data were retrieved from the SRA Study SRP389673, including samples SRR20791201, SRR20791205, and SRR20791206. All datasets were downloaded as raw sequencing data. A detailed overview of the tumor samples included in the analysis is shown in Supplementary Table 2.

Single-cell RNA-seq data processing

Raw sequencing reads from lung adenocarcinoma, melanoma, medulloblastoma and EBs samples were analyzed with the Cell Ranger 7.1.0 pipeline (10x Genomics, USA). The resulting feature-barcode matrices were imported into the R programming environment (4.5.1) for further analysis with Seurat (5.3.0) [55]. Doublet correction was performed using the scDblFinder tool (1.22.0) with default parameters [56]. Cells with fewer than 500 detected genes or more than 20% mitochondrial content were excluded from the EBs datasets. For the tumor samples, cells with fewer than 200 detected genes, identified as doublets/multiplets, or with >20% mitochondrial content were excluded. Cell-level normalization and scaling were performed using SCTransform v2. The number of principal components (PCs) retained for downstream analysis was determined as follows: 50 PCs for EBs samples and 15 PCs for tumor samples. Neighbor graphs were constructed, and clustering was performed using the FindNeighbors and FindClusters functions with default parameters. Clustering resolution was set to 0.8 for tumor samples and to 1.2 for EBs samples. Dimensionality reduction and two-dimensional visualisation were performed using UMAP on the selected PCs.

Cell type annotation

The resulting clusters of tumor samples were annotated into major cell types with CellTypist (1.6.3) using the Human Lung Cell Atlas as a reference model for lung adenocarcinoma samples, the Adult Human Skin Atlas for melanoma samples, and the Adult Human Middle Temporal Gyrus Atlas for the medulloblastoma samples [57-60]. In all samples, tumor cells were identified with SCEVAN (1.0.3) with immune cells used as a reference [61]. Additionally, tumor cell annotation was confirmed with the inferCNV package (1.24.0) (https://github.com/broadinstitute/inferCNV). Immune cells in each dataset (CD8 T cells, B cells, CD4 T cells, DC2, NK cells for lung adenocarcinoma and melanoma samples, Micro-PVM_2 for medulloblastoma samples) were designated as normal controls to establish a baseline CNV profile. Suggested tumor cells were analyzed with default parameters with cutoff set at 0.1.

Manual cell type annotation for EBs was performed based on differentially expressed genes (DEGs). Gene sets characteristic of each cluster were queried against the following databases and electronic resources: GeneAnalytics (https://ga.genecards.org), GeneCards (https://www.genecards.org), LifeMap Discovery (https://discovery.lifemapsc.com), WebGestalt (https://www.webgestalt.org), and The Human Protein Atlas (https://www.proteinatlas.org). Additionally, EBs samples were annotated with CellTypist using the Pan Fetal Human Atlas [62].

Integration of EBs and cancer datasets

To investigate transcriptional similarities between early embryogenesis and oncogenesis, EBs cells and tumor cells from three cancer types, previously identified using SCEVAN, were integrated using the IntegrateLayers function with the RPCAIntegration method. Integrations were performed in two modes: pairwise integrations of tumor cells from each individual cancer type (lung adenocarcinoma, melanoma, and medulloblastoma separately) with EBs cells and a joint integration of tumor cells from all three cancer types. As a control analysis to exclude potential effects of cell cycle-associated variation, an additional joint integration of tumor cells from all three cancer types was performed with cell cycle effects regressed out. Following each integration, the data were scaled, and PCA was performed. Neighbor graphs were constructed using the first

30 PCs, followed by UMAP dimensionality reduction and Louvain clustering at resolution 1.0. UMAP feature plots showing gene expression were visualised with scCustomise package (3.2.0) (https://github.com/samuel-marsh/scCustomize/tree/master).

Differential gene expression analysis

Identification of DEGs between clusters containing both tumor and EBs cells was performed using the FindConservedMarkers function with grouping.var set to be a unique sample ID. For DEGs from clusters 16 and 17 from the integrated dataset of three cancers, similarly, FindConservedMarkers function with grouping.var set to be a unique sample ID was used. DEGs were identified using the following criteria: log2FC > 0.3 and adjusted p-value < 0.05 and used for the Gene Ontology (GO) and KEGG enrichment analyses. In addition, a list of DEGs from these clusters was identified with FindMarkers with default parameters, filtered to include DEGs with adjusted p-value < 0.05, and used for the over-representation of MSigDB Hallmark (H) gene sets analysis.

Gene set enrichment analysis

Gene Set Enrichment Analysis (GSEA) was performed on the list of DEGs from clusters 16 and 17, as well as on the 25 hub genes from Module 17 using the clusterProfiler (4.16.0) package [63]. The enricher function was used to test significant over-representation of MSigDB Hallmark (H) gene sets among significantly upregulated and downregulated genes (adjusted p-value < 0.05), using all detected genes in the dataset as the background. Additionally, DEGs from clusters 16 and 17 were profiled using Gene Ontology (GO) and KEGG pathway enrichment analyses with enrichGO and enrichKEGG functions with default parameters. Enrichment results were visualized using functions from the enrichplot (1.28.4) package (https://github.com/YuLab-SMU/enrichplot).

Cell differentiation trajectories

To predict the least differentiated cells and the assumed starting point of differentiation in tumor cells, the CytoTRACE2 algorithm (1.1.0) was used with default settings [64]. Clusters 16, 17 and 5 with the highest in all dataset CytoTrace potency scores (more than 0.75) were defined as the root clusters for subsequent trajectory analysis. To investigate possible differentiation trajectories and infer the pseudotemporal cell ordering, the Slingshot algorithm (2.16.0) was used [65]. The algorithm was used with the existing cluster labels (reducedDim="umap") to define the graph structure of the trajectory. The root clusters (start.clus=c(5,16,17)) were selected based on the highest CytoTRACE potency scores, and cellular distances were computed dist.method = "mnn".

Transcription factor activity inference

Regulatory programs at the transcription factor (TF) level were revealed with the decoupleR package (2.14.0) [66]. Reference dataset of human TF-target gene regulatory network was retrieved with decoupleR::get_collectri function, containing interactions derived from literature curation and ChIP-seq data. Activity scores were estimated using the univariate linear model method

(run_ulm, minsize = 5), as suggested in the vignette. Per-cluster mean TF activities were computed by averaging scaled scores within each Seurat cluster. The top 50 most variable TFs across clusters were identified by ranking the standard deviation of cluster-wise mean activities. Heatmaps of cluster-level TF activities were generated using pheatmap (1.0.13) (https://github.com/raivokolde/pheatmap) with column-wise (TF-wise) z-scoring.

hdWGCNA co-expression network analysis

Single-cell high-dimensional weighted gene co-expression network analysis (hdWGCNA) was performed using the R package hdWGCNA (0.4.07) to examine co-expression modules in the cluster 16 of integrated tumor cells dataset [67,68]. A signed co-expression network was constructed with soft-thresholding power ($\beta = 18$), which was selected based on approximate scale-free topology ($R^2 > 0.8$) using TestSoftPowers. The hdWGCNA analysis was performed following the default workflow, including network construction (ConstructNetwork), followed by module detection via hierarchical clustering and dynamic tree cutting (default parameters). Module eigengenes (MEs) and harmonized module eigengenes (hMEs) were computed across all cells using ModuleEigengenes with sample-level harmonization. Intramodular connectivity (kME) was calculated within the target population using ModuleConnectivity. Hub genes were identified as the top 25 genes ranked by ModuleConnectivity within each module. Module activity scores were computed using UCell on the top 25 hub genes per module (ModuleExprScore, method = "UCell"). Differential module activity across slingshot pseudotime was evaluated using built-in PlotModuleTrajectory function.

The co-expression modules were projected onto the Visium datasets using the ProjectModules function with default parameters. Spatial visualizations were generated using SpatialFeaturePlot based on values from the GetMEs function, which returns the connectivity values for the co-expression module.

Publicly available spatial transcriptomic datasets

The processed Visium FF lung adenocarcinoma No.3 sample was retrieved from the DBKERO database (https://kero.hgc.jp/Ad-SpatialAnalysis_2024.html). Filtered matrices and tissue images of healthy patients with no known lung disease were retrieved from the BioStudies portal (accession number S-BSST1410) and included samples V10T31-019-A1 (non-smoker male), V10T31-015-A1 (ex-smoker male), and V19S23-092-A1 (smoker female) (https://www.ebi.ac.uk/biostudies/studies/S-BSST1410).

Spatial transcriptomics data processing and annotation transfer

The spatial datasets were imported into the R programming environment (4.5.1) for further analysis with Seurat (5.3.0). The FF lung adenocarcinoma No. 3 sample from the DBKERO database was provided as pre-filtered and was used without additional quality control filtering. For samples from BioStudies S-BSST1410, spots with nFeature_Spatial > 250, nCount_Spatial > 500, and percent.mt < 10% were retained for downstream analysis. Normalization was performed using SCTransform. Dimensionality reduction using PCA, clustering, and UMAP visualization were performed using the first 30 PCs.

Reference-based label transfer was performed using Seurat's anchor-based integration framework. Cluster identities corresponding to clusters 16 and 17 obtained from scRNA datasets were transferred to the spatial datasets using the TransferData function.

Validation of Module 17 in a pan-cancer scRNA-seq dataset

To validate the behavior of Module 17 in an independent dataset, a publicly available preprocessed pan-cancer scRNA-seq atlas of 355 941 cells (https://doi.org/10.5281/zenodo.14511579) was utilized [69]. The atlas was loaded as a fully normalized and annotated Seurat object with precomputed UMAP coordinates, malignant/non-malignant labels, and detailed cell type classification.

Module 17, previously identified in the discovery dataset using the hdWGCNA workflow, was transferred to the pan-cancer atlas using the ProjectModules function. The WGCNA-derived gene weights of Module 17 were mapped onto the external dataset, and the module eigengene was recalculated with GetMEs. The resulting per-cell eigengene values were stored as Modules17 and used as the Module 17 activity score in all downstream analyses.

To identify cells with high Module 17 activity, a quantile-based thresholding strategy was applied. The distribution of Module 17 values was assessed across all cells in the atlas, and the 95th percentile was selected as the cutoff defining the Module 17-high population.

Conserved markers distinguishing malignant and non-malignant Module 17-high cells were identified using the FindConservedMarkers function, with cancer type set as the grouping variable. For each cancer type, differential expression was calculated using logfc.threshold = 0.25 and min.pct = 0.1. Genes were retained only if adjusted p-values were < 0.05 across all cancer types, and if the detection ratio (pct.1/pct.2) was ≥ 2 consistently in each cancer type.

Statistics

Data visualization was performed using the ggplot2 package (3.5.2) [70]. The distribution of CytoTRACE2 scores across clusters was illustrated using a bar plot, where the height of each bar represents the mean CytoTRACE2 score for each cluster. Error bars were added to represent the standard error of the mean, calculated as the standard deviation divided by the square root of the sample size. Statistical analysis was conducted using the R programming environment (4.5.1). The normality of CytoTRACE2 score distribution was assessed using the Anderson-Darling test, and since it did not follow normal distribution, nonparametric tests were employed. The Kruskal-Wallis rank sum test was applied to compare CytoTRACE2 scores across the clusters, followed by pairwise Wilcoxon rank-sum tests with Benjamini-Hochberg adjustment for multiple testing. Statistical significance was defined as follows: not significant (ns) for $p > 0.05$, significant at $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. The enrichment of Module 17-high cells within tumor and normal cells was assessed by constructing a 2×2 contingency table based on atlas annotations, and statistical significance was evaluated using Fisher's exact test.

Author Contributions

Anna A. Khozyainova: conceptualization, methodology, formal analysis, validation, writing - original draft, data curation, visualization, supervision. Vera G. Subrakova: methodology, formal analysis, validation, writing - original draft, visualization. Maxim E. Menyailo: methodology, investigation, writing - original draft. Daria I. Zhigalina: methodology, investigation, writing - original draft. Tatiana N. Kireeva: investigation. Adelya A. Galiakberova: investigation. Mikhail A. Berestovoy: resources. Elena E. Kopantseva: investigation. Anastasia A. Korobeynikova: investigation. Erdem B. Dashinimaev: investigation. Dmitry M. Loos: validation. Maria S. Tretyakova: investigation. Ustinia A. Bokova: investigation. Nikolay A. Skryabin: conceptualization, methodology, writing - review & editing, supervision, project administration. Evgeny V. Denisov: conceptualization, methodology, writing - review & editing, supervision, project administration

Acknowledgments

The authors are grateful to S.M. Zakian (Dr. Sci. Biol., Prof., RAS, Head of the Laboratory of Developmental Epigenetics, Institute of Cytology and Genetics, SB RAS), as well as to A.A. Malakhova (Ph.D. Biol.) and E.V. Grigor'eva (Ph.D. Biol.) for providing the induced pluripotent stem cell line ICGi022-A for this research.

Conflicts of Interest

The authors declare no conflicts of interest

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Funding:

The study was supported by a grant from the Tomsk National Research Medical Center “Cellular plasticity as a general molecular program of embryo- and oncogenesis” (#24-grant-MNI).

References

1. Sathyanarayanan A, Ing-Simmons E, Chen R, et al. Early developmental plasticity enables the induction of an intermediate extraembryonic cell state. Science advances. 2022;8(44):eabl9583.

2. Linneberg-Agerholm M, Sell AC, Redó-Riveiro A, et al. The primitive endoderm supports lineage plasticity to enable regulative development. Cell. 2024;187(15):4010-4029. e16.

3. Wilkinson AL, Zorzan I, Rugg-Gunn PJ. Epigenetic regulation of early human embryo development. Cell Stem Cell. 2023;30(12):1569-1584.

4. Lammers NC, Galstyan V, Reimer A, Medin SA, Wiggins CH, Garcia HG. Multimodal transcriptional control of pattern formation in embryonic development. Proceedings of the national academy of sciences. 2020;117(2):836-847.

5. Zhao C, Plaza Reyes A, Schell JP, et al. A comprehensive human embryo reference tool using single-cell RNA-sequencing data. Nature methods. 2025;22(1):193-206.

6. Xiao Z, Cui L, Yuan Y, et al. 3D reconstruction of a gastrulating human embryo. Cell. 2024;187(11):2855-2874. e19.

7. Mehta A, Stanger BZ. Lineage plasticity: the new cancer hallmark on the block. Cancer research. 2024;84(2):184-191.

8. Shi Z-D, Pang K, Wu Z-X, et al. Tumor cell plasticity in targeted therapy-induced resistance: mechanisms and new strategies. Signal Transduction and Targeted Therapy. 2023;8(1):113.

9. Wang Z, Su Y, Zhao L, et al. Pan-Cancer Single-Cell Transcriptomic Analysis Reveals Divergent Expression of Embryonic Proangiogenesis Gene Modules in Tumorigenesis. Cancer Medicine. 2024;13(21):e70373.

10. Jolly MK, Somarelli JA, Sheth M, et al. Hybrid epithelial/mesenchymal phenotypes promote metastasis and therapy resistance across carcinomas. Pharmacology & therapeutics. 2019;194:161-184.

11. Manzo G. Similarities between embryo development and cancer process suggest new strategies for research and therapy of tumors: a new point of view. Frontiers in cell and developmental biology. 2019;7:20.

12. Sharma A, Bleriot C, Currenti J, Ginhoux F. Oncofetal reprogramming in tumour development and progression. Nature Reviews Cancer. 2022;22(10):593-602.

13. Couturier CP, Ayyadhury S, Le PU, et al. Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. Nature communications. 2020;11(1):3406.

14. Okonechnikov K, Joshi P, Sepp M, et al. Mapping pediatric brain tumors to their origins in the developing cerebellum. Neuro-oncology. 2023;25(10):1895-1909.

15. Young MD, Mitchell TJ, Vieira Braga FA, et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. science. 2018;361(6402):594-599.

16. Rajbhandari N, Hamilton M, Quintero CM, et al. Single-cell mapping identifies MSI+ cells as a common origin for diverse subtypes of pancreatic cancer. Cancer cell. 2023;41(11):1989-2005. e9.

17. Patel AS, Yanai I. A developmental constraint model of cancer cell states and tumor heterogeneity. Cell. 2024;187(12):2907-2918.

18. Tu S-M, Aydin AM, Maraboyina S, et al. Stem cell origin of cancer: Implications of oncogenesis recapitulating embryogenesis in cancer care. Cancers. 2023;15(9):2516.

19. Sun H, Kumari N, Melzer ME, Backman V, Goyal Y. Navigating Nongenetic Plasticity in Cancer Drug Resistance. Annual Review of Cancer Biology. 2025;9

20. Rivron NC, Arias AM, Pera MF, Moris N, M'hamdi HI. An ethical framework for human embryology with embryo models. Cell. 2023;186(17):3548-3557.

21. Rossant J, Tam PP. Early human embryonic development: Blastocyst formation to gastrulation. Developmental cell. 2022;57(2):152-165.

22. Rhodes K, Barr KA, Popp JM, Strober BJ, Battle A, Gilad Y. Human embryoid bodies as a novel system for genomic studies of functionally diverse cell types. Elife. 2022;11:e71361.

23. Zeevaert K, Elsafi Mabrouk MH, Wagner W, Goetzke R. Cell mechanics in embryoid bodies. Cells. 2020;9(10):2270.

24. Zhigalina DI, Denisov EV, Lebedev IN, Skryabin NA. Embryoid bodies as a model system for exploring early human embryonic development. Journal of Assisted Reproduction and Genetics. 2025:1-20.

25. Brickman JM, Serup P. Properties of embryoid bodies. Wiley Interdisciplinary Reviews: Developmental Biology. 2017;6(2):e259.

26. Fu B, Ma H, Wang L, et al. Embryonic Origins of Cancer: Insights from Double Homeobox 4 Regulation. Biomolecules. 2025;15(5):721.

27. Cofre J, Saalfeld K, Abdelhay E. Cancer as an embryological phenomenon and its developmental pathways: a hypothesis regarding the contribution of the noncanonical Wnt pathway. The Scientific World Journal. 2019;2019(1):4714781.

28. Wang Q, Fang WH, Krupinski J, Kumar S, Slevin M, Kumar P. Pax genes in embryogenesis and oncogenesis. Journal of cellular and molecular medicine. 2008;12(6a):2281-2294.

29. Han Y, Hu X, Yun X, et al. Nucleolar and spindle associated protein 1 enhances chemoresistance through DNA damage repair pathway in chronic lymphocytic leukemia by binding with RAD51. Cell Death & Disease. 2021;12(11):1083.

30. Ge Y, Li Q, Lin L, et al. Downregulation of NUSAP1 suppresses cell proliferation, migration, and invasion via inhibiting mTORC1 signalling pathway in gastric cancer. Cell Biochemistry and Function. 2020;38(1):28-37.

31. An X, Xu F, Luo R, et al. The prognostic significance of topoisomerase II alpha protein in early stage luminal breast cancer. BMC cancer. 2018;18(1):331.

32. Wu J, Li W, Zhang X, et al. Expression and potential molecular mechanism of TOP2A in metastasis of non-small cell lung cancer. Scientific Reports. 2024;14(1):12228.

33. Fäldt Beding A, Larsson P, Helou K, Einbeigi Z, Parris TZ. Pan-cancer analysis identifies BIRC5 as a prognostic biomarker. BMC cancer. 2022;22(1):322.

34. Xiang C, Yan H-c. Ubiquitin conjugating enzyme E2 C (UBE2C) may play a dual role involved in the progression of thyroid carcinoma. Cell Death Discovery. 2022;8(1):130.

35. Liu P-F, Chen C-F, Shu C-W, et al. UBE2C is a potential biomarker for tumorigenesis and prognosis in tongue squamous cell carcinoma. Diagnostics. 2020;10(9):674.

36. Tran NT, Graf R, Acevedo-Ochoa E, et al. In vivo CRISPR/Cas9-mediated screen reveals a critical function of TFDP1 and E2F4 transcription factors in hematopoiesis. Leukemia. 2024;38(9):2003-2015.

37. Matsubara J, Li YF, Koul S, et al. The E2F4 transcriptional repressor is a key mechanistic regulator of colon cancer resistance to irinotecan (CPT-11). bioRxiv. 2025;

38. Ma Z, Li Q, Wang W, Deng Z. Transcription factor E2F4 facilitates SUMOylation to promote HCC progression through interaction with LIN9. International Journal of Oncology. 2024;65(4):1-15.

39. Sun Y-m, Zhang Y-m, Shi H-l, et al. Enhancer-driven transcription of MCM8 by E2F4 promotes ATR pathway activation and glioma stem cell characteristics. Hereditas. 2023;160(1):29.

40. Zhang Y, Zhang J, Pan G, et al. Effects of EFNA1 on cell phenotype and prognosis of esophageal carcinoma. World Journal of Surgical Oncology. 2021;19(1):242.

41. Duan B, Zhu Z, You B, et al. Overexpression of ERBB3 promotes proliferation, migration, and angiogenesis in nasopharyngeal carcinoma. International Journal of Clinical and Experimental Pathology. 2019;12(8):2931.

42. Wang J, Zhu M, Zhu J, et al. HES1 promotes aerobic glycolysis and cancer progression of colorectal cancer via IGF2BP2-mediated GLUT1 m6A modification. Cell death discovery. 2023;9(1):411.

43. Yang Z, Zhang M, Xu K, et al. Knockdown of YAP1 inhibits the proliferation of osteosarcoma cells in vitro and in vivo. Oncology reports. 2014;32(3):1265-1272.

44. Ito S, Sakakibara-Konishi J, Sato M, et al. Midkine promotes tumor growth and attenuates the effect of cisplatin in small cell lung cancer. Cancer Medicine. 2025;14(13):e71034.

45. Ito A, Shimazu T, Maeda S, et al. The subcellular localization and activity of cortactin is regulated by acetylation and interaction with Keap1. Science signaling. 2015;8(404):ra120-ra120.

46. Qu G-q, Lu Y-m, Liu Y-f, et al. Effect of RTKN on progression and metastasis of colon cancer in vitro. Biomedicine & Pharmacotherapy. 2015;74:117-123.

47. Cao Q, Liu Y, Wu Y, et al. Profilin 2 promotes growth, metastasis, and angiogenesis of small cell lung cancer through cancer-derived exosomes. Aging (Albany NY). 2020;12(24):25981.

48. Chevalier C, Roche S, Bénistant C. Vesicular trafficking regulators are new players in breast cancer progression: Role of TOM1L1 in ERBB2-dependent invasion. Molecular & Cellular Oncology. 2016;3(4):e1182241.

49. Zhang C, Du F-h, Wang R-x, et al. TSPAN6 reinforces the malignant progression of glioblastoma via interacting with CDK5RAP3 and regulating STAT3 signaling pathway. International Journal of Biological Sciences. 2024;20(7):2440.

50. Miao S, Qiu T, Zhao Y, et al. Overexpression of S100A13 protein is associated with tumor angiogenesis and poor survival in patients with early-stage non-small cell lung cancer. Thoracic cancer. 2018;9(9):1136-1144.

51. Wang N, Wang R, Tang J, et al. Calbindin S100A16 promotes renal cell carcinoma progression and angiogenesis via the VEGF/VEGFR2 signaling pathway. Contrast Media & Molecular Imaging. 2022;2022(1):5602011.

52. Pankova OV, Tashireva LA, Rodionov EO, et al. Premalignant changes in the bronchial epithelium are prognostic factors of distant metastasis in non-small cell lung cancer patients. Frontiers in Oncology. 2021;11:771802.

53. Gómez-López S, Alhendi AS, Przybilla MJ, et al. Aberrant basal cell clonal dynamics shape early lung carcinogenesis. Science. 2025;388(6752):eads9145.

54. Malakhova A, Grigor'eva E, Pavlova S, et al. Generation of induced pluripotent stem cell lines ICGi021-A and ICGi022-A from peripheral blood mononuclear cells of two healthy individuals from Siberian population. Stem Cell Research. 2020;48:101952.

55. Hao Y, Stuart T, Kowalski MH, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nature biotechnology. 2024;42(2):293-304.

56. Germain P-L, Lun A, Meixide CG, Macnair W, Robinson MD. Doublet identification in single-cell sequencing data using scDblFinder. f1000research. 2022;10:979.

57. Domínguez Conde C, Xu C, Jarvis LB, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science. 2022;376(6594):eabl5197.

58. Sikkema L, Ramírez-Suástegui C, Strobl DC, et al. An integrated cell atlas of the lung in health and disease. Nature medicine. 2023;29(6):1563-1577.

59. Reynolds G, Vegh P, Fletcher J, et al. Developmental cell programs are co-opted in inflammatory skin disease. Science. 2021;371(6527):eaba6500.

60. Braun E, Danan-Gotthold M, Borm L, et al. Comprehensive cell atlas of the first-trimester developing human brain. Science 382, eadf1226. 2023.

61. De Falco A, Caruso F, Su X, Iavarone A, Ceccarelli M. A variational algorithm to detect the clonal copy number substructure of tumors from scRNA-seq data. Nat. Commun. 14, 1074. 2023.

62. Suo C, Dann E, Goh I, et al. Mapping the developing human immune system across organs. Science. 2022;376(6597):eabo0510.

63. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics: a journal of integrative biology. 2012;16(5):284-287.

64. Kang M, Gulati GS, Brown EL, et al. Improved reconstruction of single-cell developmental potential with CytoTRACE 2. Nature methods. 2025:1-6.

65. Street K, Risso D, Fletcher RB, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC genomics. 2018;19(1):477.

66. Badia-i-Mompel P, Vélez Santiago J, Braunger J, et al. decoupleR: ensemble of computational methods to infer biological activities from omics data. Bioinformatics advances. 2022;2(1):vbac016.

67. Morabito S, Reese F, Rahimzadeh N, Miyoshi E, Swarup V. hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data. Cell reports methods. 2023;3(6)

68. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics. 2008;9(1):559.

69. Gondal MN, Cieslik M, Chinnaiyan AM. Integrated cancer cell-specific single-cell RNA-seq datasets of immune checkpoint blockade-treated patients. Scientific Data. 2025;12(1):139.

70. Wickham H. Data analysis. ggplot2: elegant graphics for data analysis. Springer; 2016:189-201.