RESEARCH ARTICLE

# Developmental Deconvolution for Classification of Cancer Origin

Enrico Moiso[1,2], Alexander Farahani[3], Hetal D. Marble[3], Austin Hendricks[1], Samuel Mildrum[1], Stuart Levine[1], Jochen K. Lennerz[3], and Salil Garg[1,3]

Illustration by Bianca Dunn

**ABSTRACT** Cancer is partly a developmental disease, with malignancies named based on cell or tissue of origin. However, a systematic atlas of tumor origins is lacking. Here we map the single-cell organogenesis of 56 developmental trajectories to the transcriptomes of over 10,000 tumors across 33 cancer types. We deconvolute tumor transcriptomes into signals for individual developmental trajectories. Using these signals as inputs, we construct a developmental multilayer perceptron (D-MLP) classifier that outputs cancer origin. D-MLP (ROC-AUC: 0.974 for top prediction) outperforms benchmark classifiers. We analyze tumors from patients with cancer of unknown primary (CUP), selecting the most difficult cases in which extensive multimodal workup yielded no definitive tumor type. Interestingly, CUPs form groups distinguished by developmental trajectories, and classification reveals diagnosis for patient tumors. Our results provide an atlas of tumor developmental origins, provide a tool for diagnostic pathology, and suggest developmental classification may be a useful approach for patient tumors.

**SIGNIFICANCE:** Here we map the developmental trajectories of tumors. We deconvolute tumor transcriptomes into signals for mammalian developmental programs and use this information to construct a deep learning classifier that outputs tumor type. We apply the classifier to CUP and reveal the developmental origins of patient tumors.

*See related commentary by Wang, p. 2498.*

## INTRODUCTION

Diagnosis of malignancy relies on histopathologic classification of tumor appearance, often alongside other features such as mutation profiling and clinical presentation. However, many tumors display a spectrum of heterogeneous appearances, which may in part reflect unknown differences in their development. Tumor heterogeneity can lead to diagnostic uncertainty, with disagreement among pathologists, overdiagnosis, underdiagnosis, or inability to distinguish "gray zone" cases between tumor types (1–3). Additionally, the cell type of origin for some cancers is unclear, and these malignancies are usually classified based on tissue of occurrence. Further, tumors can dedifferentiate, correlating with more aggressive behavior and complicating diagnostic identification. Cancers of unknown primary (CUP) represent malignancies that are often particularly dedifferentiated and aggressive with poor survival rates. Lack of diagnostic information is one factor that complicates the treatment of many cancers, including CUPs, which are usually treated using nontargeted therapies with harsh toxicities. Understanding developmental pathways dysregulated in malignancies is a major goal in cancer biology and could enable targeted therapeutic interventions guided by more precise diagnosis tools.

Machine learning classifiers have shown promise as new tools when applied to image processing in radiology and histopathology (4–6). However, image classifiers only detect visual features and are sometimes subject to artifacts (7, 8). Classifiers that use molecular features, such as gene expression, have great potential to aid in diagnosis through capturing nonvisual information, and recent approaches have demonstrated value in combining visual and molecular features for classification (9, 10). However, gene expression classifiers have suffered from overfitting due to the high number of features (11–13), which results in poor predictive power on new datasets. Alternatively, selecting small gene panels for measurement also reduces predictive power by not utilizing all the information. A key challenge in utilizing gene expression data to build integrated diagnostic models is how to reduce the number of features while extracting the most relevant information.

To address these challenges, we made use of two comprehensive atlases: The Cancer Genome Atlas (TCGA) and the Mouse Organogenesis Cell Atlas (MOCA; refs. 14, 15). TCGA contains expression data for 33 bulk sequenced tumor and normal tissue types accompanied by diagnostic annotations. MOCA, in turn, contains single-cell expression profiling [single-cell RNA sequencing (scRNA-seq)] dissecting the mammalian organogenesis process during E9.5 to E13.5 after fertilization in mice (corresponding to E22 to E44 in humans) and arranges single cells into developmental trajectories. For covering periods when adult mammalian lineages are developed, the MOCA study represents the most complete atlas of mammalian organogenesis, and MOCA developmental lineages show a very high degree of similarity to their human ortholog lineages (16).

In this analysis, we systematically compare both atlases to define the developmental relationships of human tumors. We

[1]Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts. [2]Broad Institute of Harvard-MIT, Cambridge, Massachusetts. [3]Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts.

**Corresponding Authors:** Salil Garg, Yale School of Medicine, 333 Cedar Street, Clinic Building 0407A, New Haven, CT 06520. Phone: 203-737-7023; E-mail: salil.garg@yale.edu; and Enrico Moiso, Koch Institute for Integrative Cancer Research, 500 Main Street, Room 76-417, Cambridge, MA 02142. E-mail: emoiso@mit.edu

Cancer Discov 2022;12:2566–85

**doi:** 10.1158/2159-8290.CD-21-1443

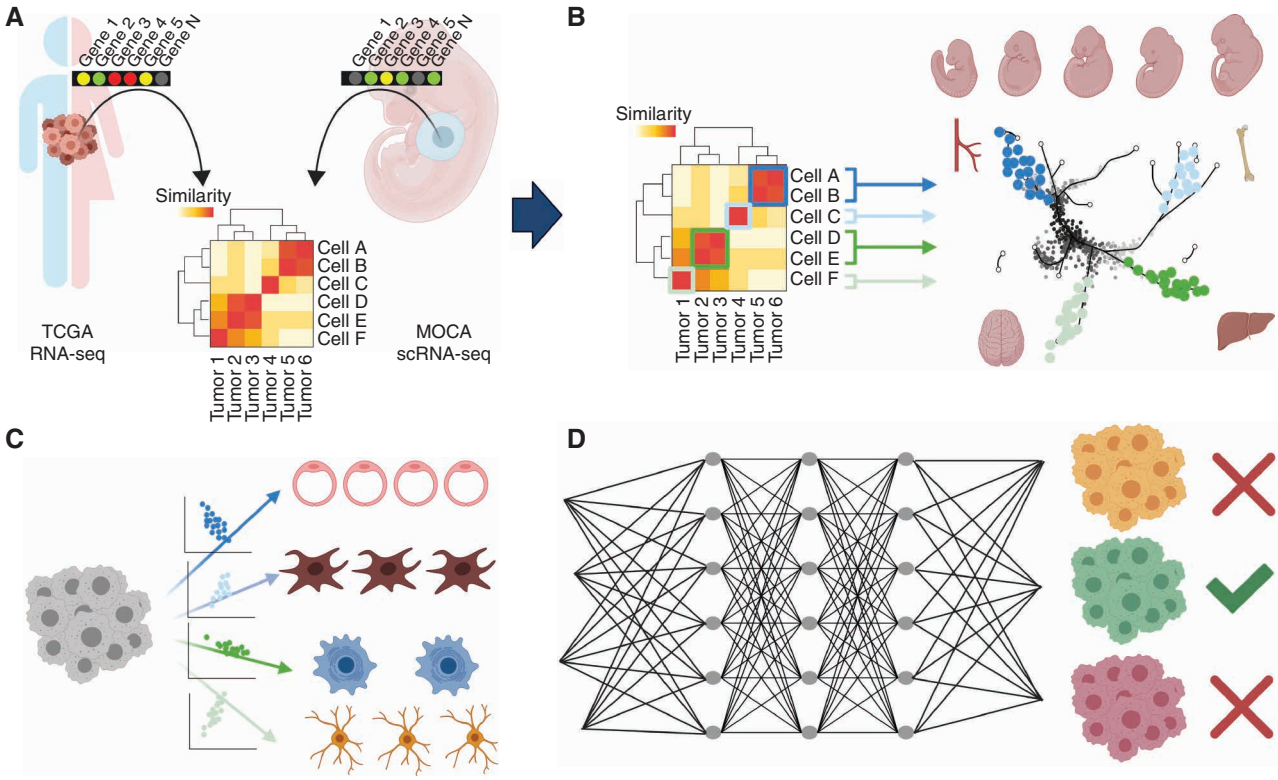



**Figure 1.** Diagnosis of malignancy by developmental deconvolution and machine learning. **A,** A comparison between bulk RNA-seq data from TCGA and scRNA-seq from MOCA was performed, generating a systematic developmental correlation analysis (map) of human tumors. **B,** Each mapped relationship consists of tumor types and developmental subtrajectories, represented at different stages of embryogenesis. **C,** Using this map, bulk gene expression signatures from each tumor sample can be deconvoluted into component developmental trajectories. **D,** Scores for each developmental trajectory at each embryonic time point can be inputted into a multilayer perceptron classifier that outputs tumor type prediction. This classifier is then applied to CUP.

apply deconvolution to bulk tumor transcriptome data and identify the most closely related developmental trajectories for each tumor sample. We input scores for each developmental trajectory into a deep learning algorithm [developmental multilayer perceptron (D-MLP)] that predicts tumor types with up to 99% accuracy. Finally, we apply this tool to CUPs and make predictions for patient samples, narrowing the differential diagnosis with implications for treatment.

## RESULTS

### Systematic Mapping of TCGA Tumors by Developmental Trajectories

An overview of the study is shown in Fig. 1. In brief, we mapped tumors to trajectories belonging to major cell lineages and developmental programs (Fig. 1A and B). This allowed us to deconvolute bulk tumor gene expression signatures into scores for each developmental program (Fig. 1C), which we inputted into a multilayer perceptron that classified tumor type (Fig. 1D).

In order to achieve this goal, first we systematically compared gene expression profiles of TCGA samples with MOCA single cells. We calculated the rank-based correlation coefficient for expressed genes between each TCGA sample and each MOCA single cell (Fig. 2A; see Methods). Altogether, the analysis constituted a systematic comparison between 15,929 genes expressed across 10,388 TCGA samples derived from 9,681 patients (cohort details, Supplementary Table S1) and 1,331,984 single cells derived from the MOCA dataset. In total, 21,217,199,458 datapoints were used to compute 13,836,649,792 correlation coefficients (summary statistics, Supplementary Table S2). We verified that these coefficients represented meaningful association between the two datasets by comparing them to those generated from row-randomized data (Supplementary Fig. S1A–S1C). In the MOCA study, single cells are grouped into 10 main trajectories, which are then divided into 56 subtrajectories, using gene expression similarity and known marker genes (15). Further, for each subtrajectory, MOCA provides Uniform Manifold Approximation (UMAP) coordinates for all the cells of that trajectory. We averaged correlation coefficients for each sample of the same TCGA type (Fig. 2A) and plotted their similarity against cells in each MOCA subtrajectory (Supplementary Fig. S1D). This revealed many expected relationships, such as inhibitory neuronal trajectories showing similarity with low-grade gliomas [brain lower grade glioma (LGG)] but not hepatocellular tumors [liver hepatocellular carcinoma (LIHC)], and vice versa for hepatocyte trajectories (Fig. 2B).

Next, we averaged correlation coefficients across all cells of the same developmental subtrajectory (Fig. 2A) and visualized them as a single similarity score compared against TCGA sample type (Fig. 2C). Hierarchical clustering analysis of these data identified six TCGA sample and six developmental subtrajectory clusters, for a total of 36 (Fig. 2C).

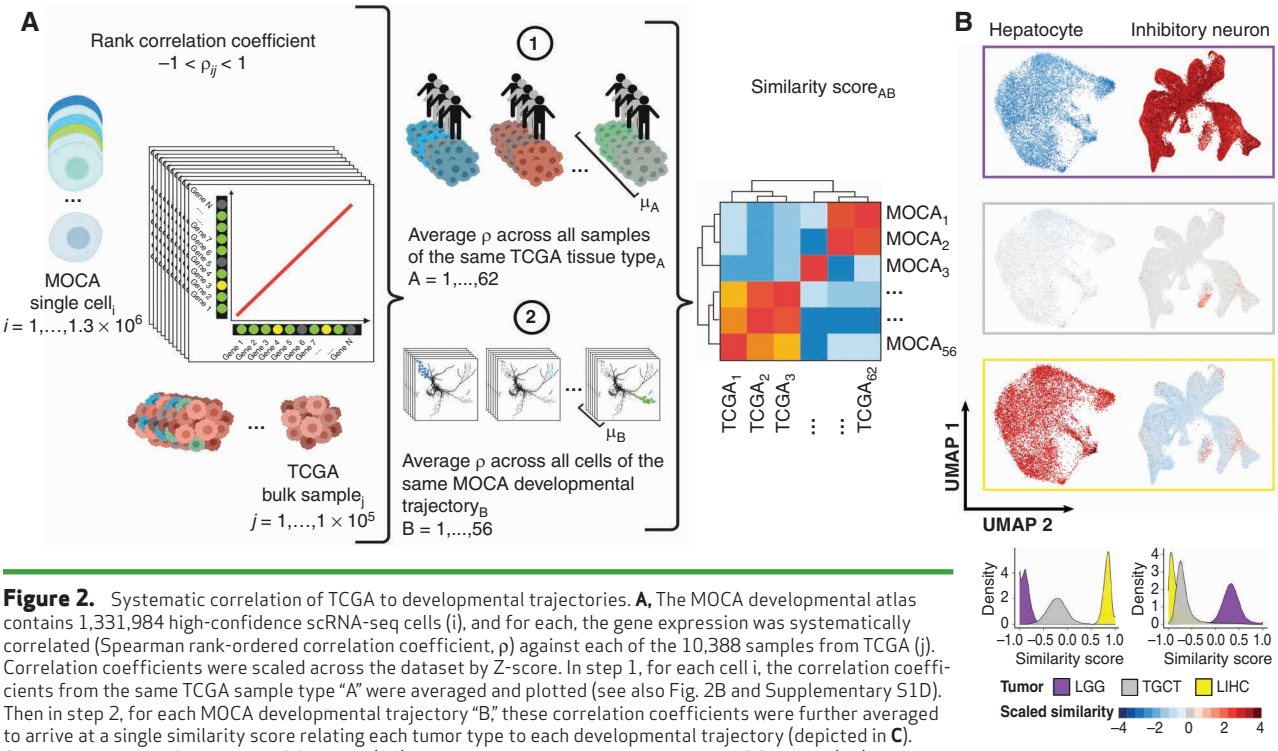



**Figure 2.** Systematic correlation of TCGA to developmental trajectories. **A,** The MOCA developmental atlas contains 1,331,984 high-confidence scRNA-seq cells (i), and for each, the gene expression was systematically correlated (Spearman rank-ordered correlation coefficient, ρ) against each of the 10,388 samples from TCGA (j). Correlation coefficients were scaled across the dataset by Z-score. In step 1, for each cell i, the correlation coefficients from the same TCGA sample type "A" were averaged and plotted (see also Fig. 2B and Supplementary S1D). Then in step 2, for each MOCA developmental trajectory "B," these correlation coefficients were further averaged to arrive at a single similarity score relating each tumor type to each developmental trajectory (depicted in **C**). Sample types are defined in the TCGA study (14) and developmental trajectories in the MOCA study (15) and were used as given. **B,** Cells from the MOCA subtrajectories hepatocyte development and inhibitory neuron development are plotted according to their UMAP coordinates given in that study. The similarity of each cell with selected TCGA sample types is shown, as is the distribution of similarity scores across the cells for each TCGA type (LGG, TGCT, LIHC). Inhibitory neuron trajectory cells showed higher similarity scores with LGG than did hepatocyte trajectory cells, and vice versa for LIHC. TGCT was not significantly related to either trajectory. (*continued on next page*)

Clusters highlighted relationships between tissue types and developmental trajectories: brain-derived samples [LGG, glioblastoma multiforme (GBM)] with neuronal subtrajectories, kidney tumors [kidney chromophobe (KICH), kidney renal papillary cell carcinoma (KIRP)] with renal epithelial trajectories, hepatocellular tumors (LIHC) with hepatocyte trajectories, and testis tumors [testicular germ cell tumors (TCGT)] with germ cell trajectories, among several other expected correlations. Furthermore, we observed expected developmental lineage relationships: Melanoma samples [skin cutaneous melanoma (SKCM)] with neural crest trajectories, carcinoma tumors with epithelial lineages, and mesenchymal tumor types with mesoderm-derived developmental lineages all showed strong similarity (Fig. 2C).

The identification of expected relationships supported the notion that the observed correlations were due to underlying biological relationships and served as partial validation of the method. In order to further validate the observed correlations, we used two approaches. In the first approach, we developed an optimized protocol for transcriptome sequencing from formalin-fixed, paraffin-embedded tissue (FFPE), sequenced the transcriptome for 40 tumors of known types, and compared similarity for developmental trajectories to TCGA. Comparison between the FFPE cohort and the TCGA cohort showed strong agreement (Supplementary Fig. S2; average Spearman ρ = 0.69), validating the method in an independent sample set. In the second approach, we utilized a single-cell atlas of human fetal tissues cataloging later embryonic stages of mid-gestation development (17). We used a representative set of cells provided by the human atlas, correlated them with TCGA sample types, and compared the results with those from the murine atlas. At least five of six TCGA sample-type clusters were observed when using human cells (Supplementary Fig. S3A). Further, when comparing orthologous lineages for similarity with TCGA, a strong agreement was obtained between murine and humans (Supplementary Fig. S3B; Pearson correlation coefficient $R = 0.78$). Reproducibility of correlations between tumors and developmental trajectories across cohorts and species supported the idea that correlations were due to underlying biological relationships. We focused analysis on the MOCA dataset given its use of earlier developmental stages and relative completeness.

## Developmental Time Map Differences between Tumors and Normal Tissue

For several tumor types, TCGA contains matched normal tissue, allowing tumor–normal comparison of developmental relationships (Fig. 2C; Supplementary Fig. S4A and S4B; normal tissue and primary tumor). Normal and malignant tissue types from similar anatomic locations were more likely to be clustered with each other, consistent with previous observations that tumor expression data largely reflect cell type of origin (18). In general, normal tissues tended to have higher scores for vascular endothelial trajectories, highlighting the abnormal vascularization associated with malignancy (19). Tumors characterized by strong inflammatory responses, such as bladder cancer (BLCA) and

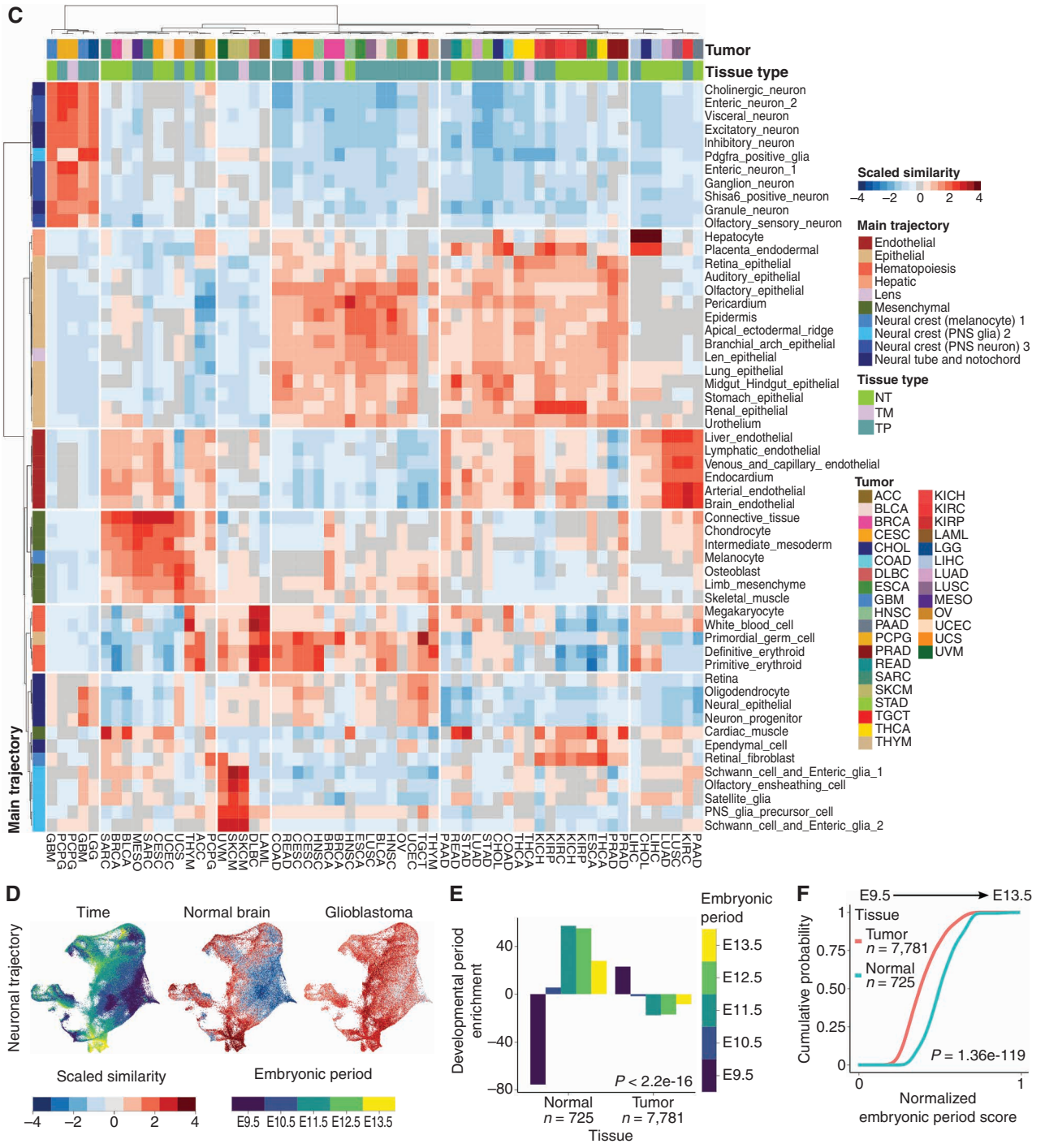



**Figure 2.** ***(Continued)*** **C,** Heat map showing the scaled similarity between every developmental subtrajectory and each TCGA sample type. TCGA samples from normal tissue (NT; green), metastatic tumors (TM; gray), and primary tumors (TP; aqua) are indicated at the top. Main developmental trajectory types are defined by the MOCA dataset for each subtrajectory and shown on the left. Subtrajectory names are listed on the right. Hierarchical clustering of rows (trajectories) and columns (sample types) is shown. PNS, peripheral nervous system. **D,** Neuronal progenitor cells are plotted according to their UMAP coordinates and colored according to their embryonic time point of isolation (left), similarity to normal brain samples (middle), or glioblastoma samples (right). Quantification in Supplementary Fig. S4B. **E,** Pan-cancer comparison of tumor samples with normal samples for enrichment of the embryonic days from which their most similar developmental cells are derived (see Methods). Chi-square ($\chi^2$) testing *P* value is shown. **F,** Cumulative distribution plot of the normalized embryonic period score for all tumor and normal samples in TCGA is shown (see Methods). Kolmogorov–Smirnov testing *P* value is shown. TCGA code names: ACC, adrenocortical carcinoma; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, cholangiocarcinoma; COAD, colon adenocarcinoma; ESCA, esophageal carcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; KICH, kidney chromophobe; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LAML, acute myeloid leukemia; LGG, brain lower grade glioma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; MESO, mesothelioma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PCPG, pheochromocytoma and paraganglioma; PRAD, prostate adenocarcinoma; READ, rectum adenocarcinoma; SARC, sarcoma; SKCM, skin cutaneous melanoma; STAD, stomach adenocarcinoma; TGCT, testicular germ cell tumors; THCA, thyroid carcinoma; THYM, thymoma; UCEC, uterine corpus endometrial carcinoma; UCS, uterine carcinosarcoma; UVM, uveal melanoma.

lung adenocarcinoma (LUAD), showed strong similarity with hematopoietic trajectories when compared with normal samples from the same tissues. Additionally, some tumor–normal comparisons showed a relative loss of differentiation programs upon malignant transformation, such as a reduction in hepatocyte lineage similarity in cholangiocarcinoma (CHOL) compared with normal gallbladder (Supplementary Fig. S4A), consistent with ideas of cancer as a disease of dedifferentiation (20, 21).

To further define whether tumors were dedifferentiated compared with normal tissue, we utilized the fact that the MOCA study dissected the organogenesis process in mice over time, between E9.5 and E13.5. We assessed how relationships between samples and trajectories changed at different embryonic times when comparing tumor and normal tissue. First, we visualized cells from a MOCA developmental trajectory, such as the neuron progenitor trajectory, and their known embryonic time of origin (Fig. 2D, left). Next, we plotted similarity scores for both glioblastoma and normal brain for each cell. Glioblastoma shows a high degree of similarity with neuronal cells from all time points, whereas the normal brain is more similar to neuronal cells from later time points (Fig. 2D; Supplementary Fig. S4B; $P < 2 \times 10^{-16}$ for all time points except E12.5). Next, we extended this analysis to a pan-cancer cohort comprising all TCGA tumor types for which normal and primary tumors were available (Supplementary Table S3). We identified a fixed number of MOCA cells with the highest positive correlation to each sample and noted the embryonic day at which they were isolated. This allowed us to compute enrichment for tumor or normal cells at each embryonic day using categoric testing ($\chi^2$; see Methods). We found a pan-cancer enrichment for earlier embryonic periods in tumors compared with normal tissue (Fig. 2E; $P < 2.22 \times 10^{-16}$, $\chi^2$ test). Finally, for each sample, we grouped the known embryonic time periods of the most enriched MOCA cells by adding them together, allowing us to calculate an "embryonic period score." Tumors were shifted toward a lower embryonic period score (Fig. 2F; $P < 1.4 \times 10^{-119}$ Kolmogorov–Smirnov test), providing pan-cancer confirmation that tumors represented dedifferentiation compared with NT.

Interestingly, tumor–normal comparisons also revealed unexpected or emerging relationships. Recently, some epithelial-derived tumors such as LUAD have been noted to change their phenotypic characteristics in favor of parallel developmental pathways (22, 23). The lung bud develops from the anterior foregut in embryonic week 3. We observed that lung-derived tumors [LUAD and lung squamous cell carcinoma (LUSC)] showed strong similarity with gut-derived trajectories, such as stomach and midgut/hindgut, and contrasted with normal lung tissue that did not show these similarities (Fig. 2C; Supplementary Fig. S4A). Therefore, this may represent reexpression of an earlier embryonic developmental program in these tumors. Additionally, glioblastomas represent a particularly heterogeneous tumor, with the exact cell type(s) of origin and their contributions to pathogenesis relatively unclear (24–26). Our analysis noted a strong correlation of glioblastomas with both main developmental trajectories, neural tube notochord and neural crest peripheral nervous system neuron, whereas other main trajectory lineages did not show such strong similarity (Fig. 2C). This relationship was supported by a correlation between glioblastomas and the analogous human cell types (Supplementary Fig. S3A). Thus, a systematic comparison between tumors and developmental trajectories revealed many specific relationships with both expected and emerging insights.

## Deconvolution of Tumor Transcriptomes into Component Developmental Trajectories

The creation of a correlation map between TCGA samples and developmental trajectories inspired us to attempt a systematic developmental deconvolution of human tumor gene expression. In deconvolution, a recorded signal (bulk gene expression) made of component parts (developmental programs) is deconstructed into individual signals from each component (trajectories at embryonic time points). We used developmental components (DC), a single quantitative measure of each developmental subtrajectory at each time point, to represent the developmental information for every TCGA sample (Supplementary Table S4; Methods). DCs were scaled across all tumor samples and charted on radar plots, which represent information about the developmental period, subtrajectory, and DC score for each sample. A schematic for this plot is shown in Fig. 3A (Supplementary Fig. S5).

To understand how developmental signals from different cell types appeared on the radar plots, we analyzed data from available scRNA-seq studies of human tumors. Altogether, we analyzed data from scRNA-seq studies of 13 different tumor types representing 237 patients (26–38). First, we considered the deconvoluted signal from one cell of a known type. For example, the deconvolution of a single T cell gave a strong signal for white blood cell subtrajectory, which was a part of the hematopoietic main trajectory (Fig. 3B). Comparing signals from multiple T cells (Fig. 3B; Supplementary Figs. S6A and S6B and S7A) showed fairly consistent scores for subtrajectories across different T cells. In contrast, signals for fibroblasts were distinct from T-cell signals and showed enrichment for the mesenchymal subtrajectories intermediate mesoderm and limb mesenchyme (Fig. 3B). Similarly, fibroblast signals were also consistent cell to cell when compared with other fibroblasts (Fig. 3B; Supplementary Figs. S6B and S7A). We examined a variety of annotated normal cell types from all 13 scRNA-seq studies, finding each normal cell type consistent in its developmental signals but distinct from each other (Supplementary Figs. S6B and S7B).

Next, we examined how these signals would combine in the radar plot of one tumor. First, we aggregated the annotated normal cells for a LIHC sample and visualized their signals on the radar plot (Fig. 3C, top left). We did the same for the annotated malignant cells (Fig. 3C, top right), noting that malignant cells showed different DC scores for many trajectories when compared with normal cells, with higher signal for hepatic trajectories and lower signal for endothelial trajectories. We deconvoluted the combined gene expression of all single cells from this tumor (Fig. 3C, bottom) and compared it with the deconvolution of one bulk-sequenced TCGA hepatocellular carcinoma sample (Fig. 3D). Deconvolution of the bulk-sequenced tumor showed strong similarity to the aggregated signals from the scRNA-seq–sequenced tumor, suggesting that developmental deconvolution captured relevant signals from both malignant cells and admixed normal cells.

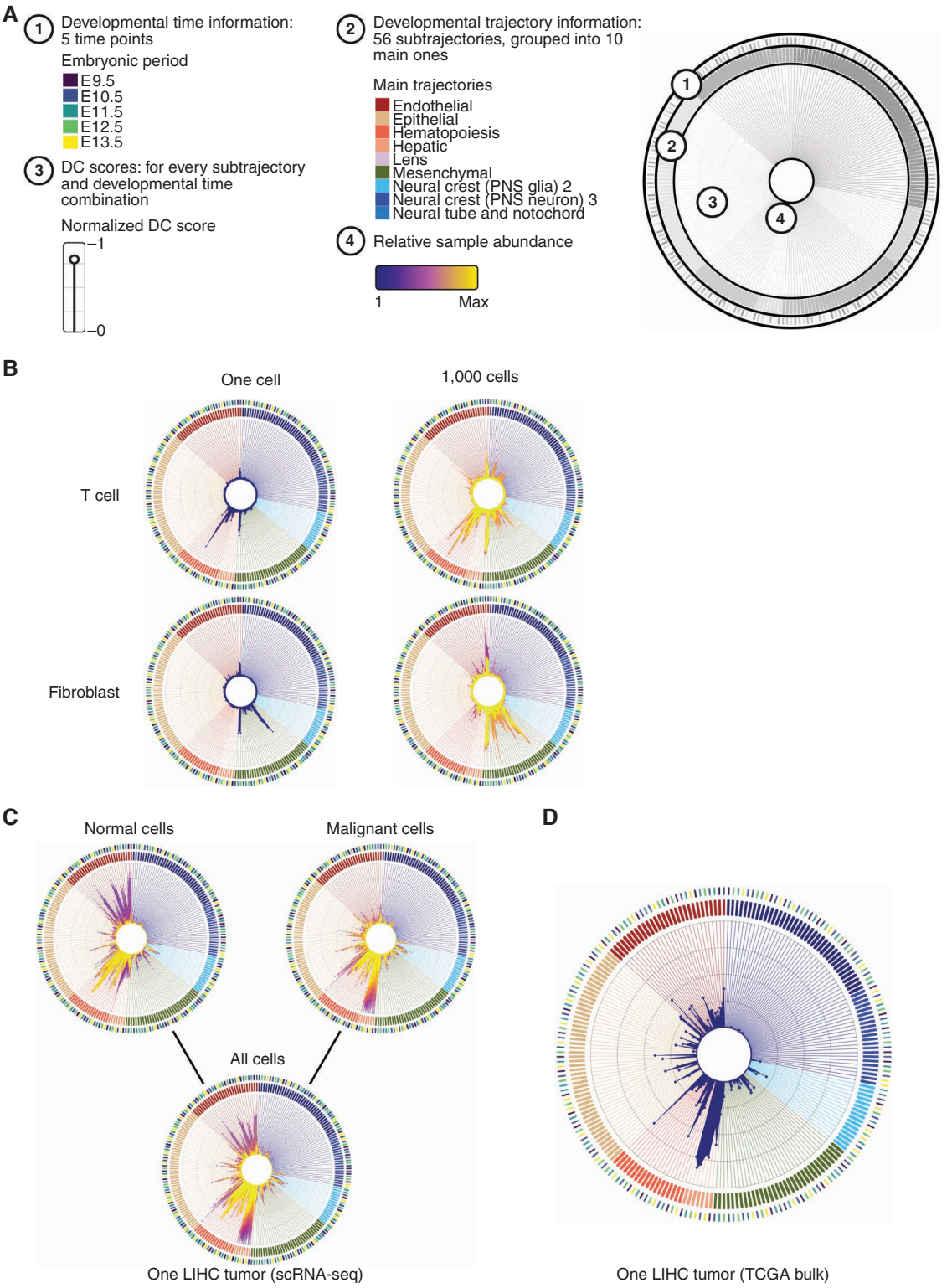



**Figure 3.** Developmental deconvolution captures signals from normal and malignant cells. **A,** Illustration of radar plots for depicting results of developmental deconvolution. Multiple layers of information are present. For each sample, a deconvolution score is generated for each developmental subtrajectory (#3) at each embryonic time point (#1) for a total of 214 scores (DCs). Scores are represented as the distance from the center in the innermost circle. The main trajectory (#2) of each subtrajectory is also indicated and colored, as is the relative number of samples showing each signal within each radar plot or figure panel (#4, purple to yellow). See Supplementary Fig. S5 for subtrajectory order. PNS, peripheral nervous system. **B,** Results of developmental deconvolution performed on one T cell or fibroblast cell from scRNA-seq (left) or the aggregated signal from 1,000 single cells (right). **C,** For one LIHC sequenced by scRNA-seq, the signals for all aggregated normal (nonmalignant) and malignant cells are shown. **D,** Developmental deconvolution of one bulk-sequenced liver hepatocellular tumor is shown for comparison.

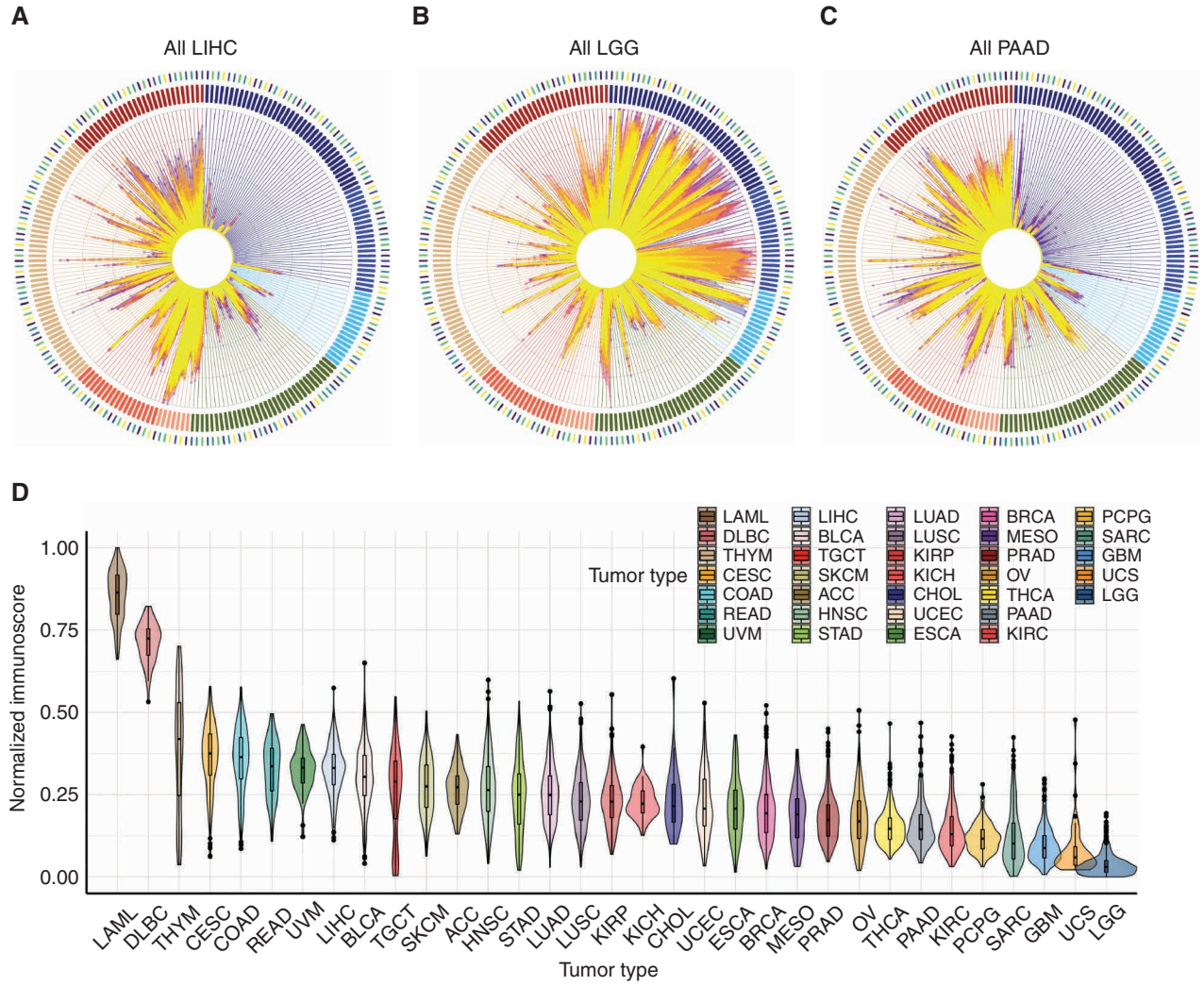



**Figure 4.** Developmental deconvolution of tumor samples. **A,** Radar plot showing the developmental deconvolution signals for all TCGA hepatocellular carcinoma samples. **B,** Radar plot showing deconvolution signals for all LGG samples. **C,** Radar plot showing deconvolution signals for all pancreatic adenocarcinoma (PAAD) samples. **D,** An immune infiltrate score ("immunoscore") was calculated for each TCGA tumor sample as the sum of deconvolution scores for relevant trajectories. The distribution of scores for each TCGA tumor type is shown (box plot center median, box edges 25th to 75th percentiles). (*continued on next page*)

## Developmental Deconvolution of Different Tumor Types Yields Distinct Profiles

In the developmental deconvolution of a TCGA bulk-sequenced hepatocellular carcinoma, we appreciated a strong enrichment in signal for hepatic trajectories, low signal for neuronal and mesenchymal trajectories, and limited signal for hematopoietic and endothelial trajectories (Fig. 3D). This pattern extended to additional hepatocellular carcinoma samples (Supplementary Fig. S8A). However, differences were also noted among LIHC samples, particularly in deconvolution scores for endothelial trajectories. This may reflect differing degrees of tumor vascularization captured through deconvolution and reflected in the DC scores for these trajectories.

We plotted the signal across all TCGA hepatocellular carcinoma samples in a single radar plot (Fig. 4A). LIHC samples were characterized by elevated scores in hepatic trajectories and depletion in neuronal trajectories. In contrast, LGG (Fig. 4B) showed high signal for neuronal trajectories and low signal for hepatic trajectories, consistent with prior correlation analysis (Fig. 2B). Although LGG did not show much sample-to-sample variation in endothelial DCs, perhaps reflecting that these tumors are not well vascularized, they did show variable deconvolution into neuronal trajectories (Supplementary Fig. S8B). This may reflect differences in the anatomic location in the brain from which each tumor was isolated or could reflect patient heterogeneity in the precise developmental context in which each tumor arose. Other tumor types showed distinct but specific patterns, such as pancreatic adenocarcinomas (PAAD), which had high DC scores in epithelial trajectories (Fig. 4C). Radar plots for each tumor type are shown in Supplementary Fig. S9.

Many tumor types are known to be infiltrated by immune cells, and the degree of infiltration for each patient can serve as a predictive biomarker for immune-targeted therapies.

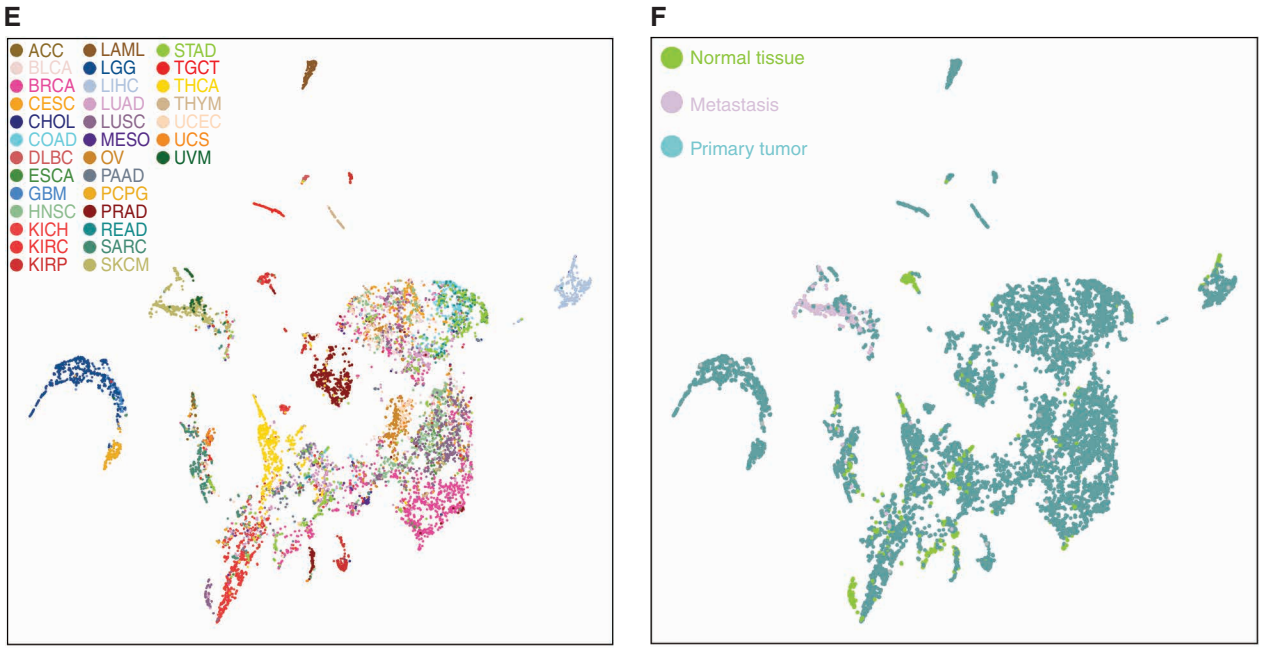



**Figure 4.** ***(Continued)*** **E,** UMAP dimensionality reduction was performed on the 214 developmental deconvolution scores for each TCGA sample and plotted. Tumor type is indicated. See also Supplementary Fig. S10 for each tumor type colored separately. **F,** Identical to **E**, but each sample is colored by sample type (normal, primary tumor, and metastasis). TCGA code names: ACC, adrenocortical carcinoma; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, cholangiocarcinoma; COAD, colon adenocarcinoma; DLBC, lymphoid neoplasm diffuse large B-cell lymphoma; ESCA, esophageal carcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; KICH, kidney chromophobe; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LAML, acute myeloid leukemia; LGG, brain lower grade glioma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; MESO, mesothelioma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PCPG, pheochromocytoma and paraganglioma; PRAD, prostate adenocarcinoma; READ, rectum adenocarcinoma; SARC, sarcoma; SKCM, skin cutaneous melanoma; STAD, stomach adenocarcinoma; TGCT, testicular germ cell tumors; THCA, thyroid carcinoma; THYM, thymoma; UCEC, uterine corpus endometrial carcinoma; UCS, uterine carcinosarcoma; UVM, uveal melanoma.

As our method captured admixed immune cells in tumor samples (Fig. 3; Supplementary Figs. S6 and S7), we analyzed the extent to which each tumor sample showed immune infiltration. We calculated an immune infiltration score as the sum of DCs for immune-related developmental lineages and plotted the distribution of these scores for each tumor type (Fig. 4D, "immunoscore"). As expected, some tumor types showed high immune infiltration. More interestingly, in some cases, the variance among samples of a particular tumor type was not uniform. For example, adrenal cortical carcinomas (ACC) showed moderately strong scores but with low sample-to-sample variation, whereas endometrial carcinomas [uterine corpus endometrial carcinoma (UCEC)] showed lower average scores but with higher sample-to-sample variation (Fig. 4D, compare median and distribution of violin plots). Other tumor types with known variation in immune infiltration across patients, such as bladder cancer [bladder urothelial carcinoma (BLCA)] and LUAD, also showed high variance in the immune infiltration score. This indicated that developmental deconvolution captured the degree of immune infiltration in each sample, and that signals for such infiltration were not uniform across the tumor repertoire.

Together, these analyses suggested developmental deconvolution was effective at separating different tumor types from one another. To confirm this, we plotted DC scores across all TCGA samples using UMAP dimensional reduction (214 DC scores inputted per sample; Methods), annotating by tumor type (TCGA code, Fig. 4E and Supplementary Fig. S10) and by tissue type (normal, primary tumor, metastases; Fig. 4F). Normal tissue and primary tumor were spread throughout the plot and were clustered largely according to tissue of origin. Further, although most tumor types were strongly clustered (e.g., LIHC, light gray, and LGG, medium blue, Fig. 4E and Supplementary Fig. S10), some tumor types [e.g., breast invasive carcinoma (BRCA), magenta] spread throughout the plot. Collectively, this suggested that the TCGA–MOCA correlation mapping could be used to deconvolute tumor gene expression into DCs in a manner that resolved most tumor types.

## Construction of the D-MLP Classifier for Cancer Type Prediction

The ability to resolve different tumor types by developmental deconvolution raised the possibility of designing a supervised machine learning (SML) model to classify malignancies. Previously, SML approaches have been applied to gene expression data but have been limited by difficulties with model overfitting due to the high dimensionality of the transcriptome (~22,000 protein-coding genes; refs. 39, 40). To avoid these issues, some studies have selected small gene subsets (41, 42), but this compromises accuracy and predictive power. We reasoned a classifier based on developmental deconvolution scores would extract the most relevant data from gene expression in the form of embryologic development programs dysregulated in tumors (Fig. 5A). After literature mining (43) and

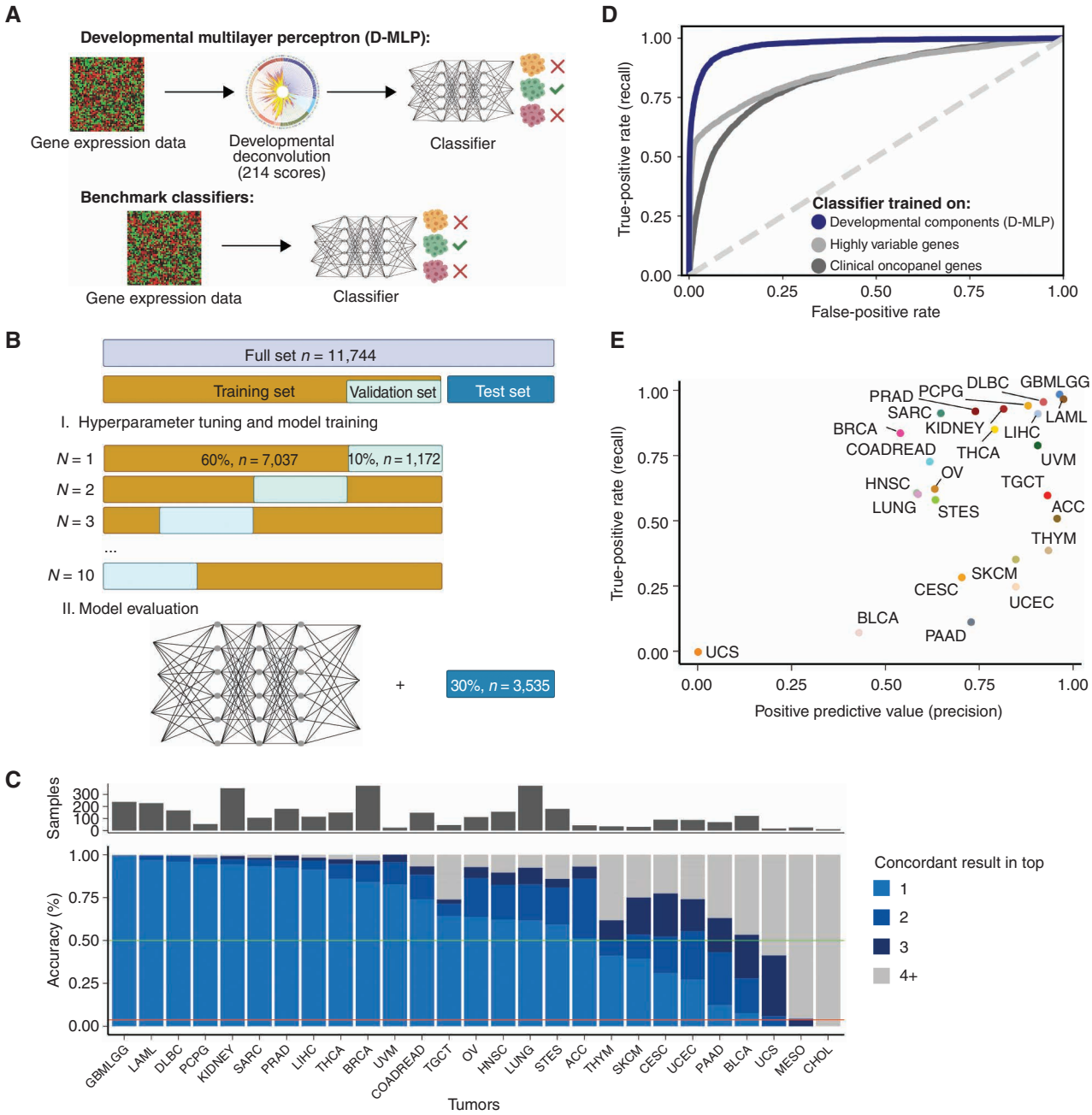



**Figure 5.** Construction and testing of the D-MLP classifier for tumor type. **A,** Schematic for classifier construction and testing. The D-MLP classifier uses developmental deconvolution scores calculated as the similarity of each tumor's gene expression to embryologic developmental trajectories as input. Comparison is made against benchmark classifiers that use gene expression data directly. **B,** Parameter optimization and model training. The full cohort contains samples from multiple cancer studies (TCGA, BEATAML1.0, CGCI-BLGSP, CTSP-DLBCL1, MMRF CoMMpass, CPTAC, and TARGET). Of the full cohort (11,744 samples), (i) 70% of cases were sampled, 60% for training and 10% for validation, in hyperparameter optimization using a 10-fold cross-validation approach to construct the classifier, and (ii) 30% of cases were held out and never seen by the model during training or optimization (test set). The model was assessed in these cases to gauge performance. **C,** Classifier accuracy (concordance) measured against the test set and number of samples is shown for all TCGA tumor types. **D,** Microaveraged ROC plotting the true-positive rate (also known as recall) as a function of false-positive rate for classifier performance for the top prediction. D-MLP classifier performance (blue line) and random guess performance (dashed gray line) are shown. ROC-AUC for the top prediction was calculated as 0.974 ± 0.003. Also shown are ROC performance curves for benchmark classifiers trained on either the most highly variable genes in expression across the training cohort ("highly variable genes," light gray, ROC-AUC: 0.859; 95% confidence interval, 0.829–0.976) or expression of a panel of genes tested in routine clinical assays at our institution ("clinical oncopanel genes," dark gray, ROC-AUC: 0.836; 95% confidence interval, 0.828–0.975). **E,** Precision (positive predictive value) versus recall (true positive rate) performance characteristics for D-MLP for each TCGA tumor type. Note CHOL and mesothelioma (MESO), are omitted as the positive predictive value is undefined for these tumors.
(*continued on next page*)

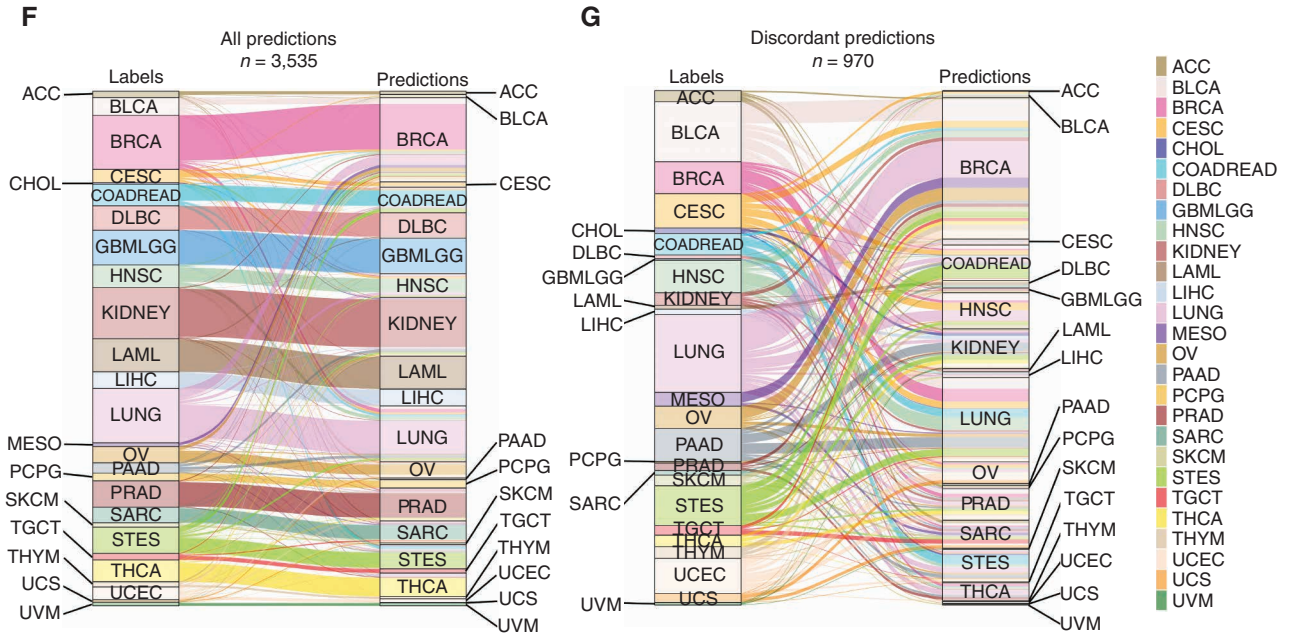



**Figure 5.** ***(Continued)*** **F,** Sankey plot showing classifier results for the top tumor type prediction for all samples ($n = 3,535$). **G,** Sankey plot showing the results for discordant classifications ($n = 1,006$) for top predictions. ACC, adrenocortical carcinoma; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, cholangiocarcinoma; COADREAD, colon adenocarcinoma (COAD) + rectum adenocarcinoma (READ); DLBC, lymphoid neoplasm diffuse large B-cell lymphoma; GBMLGG, glioblastoma multiforme (GBM) + brain lower grade glioma (LGG); HNSC, head and neck squamous cell carcinoma; KIDNEY, kidney chromophobe (KICH) + kidney renal clear cell carcinoma (KIRC) + kidney renal papillary cell carcinoma (KIRP); LAML, acute myeloid leukemia; LIHC, liver hepatocellular carcinoma; LUNG, lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC); MESO, mesothelioma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PCPG, pheochromocytoma and paraganglioma; PRAD, prostate adenocarcinoma; SARC, sarcoma; SKCM, skin cutaneous melanoma; STES, esophageal carcinoma (ESCA) + stomach adenocarcinoma (STAD); TGCT, testicular germ cell tumors; THCA, thyroid carcinoma; THYM, thymoma; UCEC, uterine corpus endometrial carcinoma; UCS, uterine carcinosarcoma; UVM, uveal melanoma.

testing different approaches, we decided on a multilayer perceptron due to its ability to simultaneously perform feature extraction and classification. This class of SML algorithms relies on artificial neurons, or threshold logic units, organized in three classes of layers (input, hidden, and output) and takes advantage of back propagation to increase accuracy.

First, we expanded our cohort beyond TCGA by incorporating tumor transcriptome samples from other cancer cohorts (BEATAML1.0, CGCI-BLGSP, CTSP-DLBCL1, MMRF CoMMpass, CPTAC, and TARGET; refs. 44–48) and FFPE (Supplementary Fig. S2). To incorporate different studies together, we merged subclasses for specific tumors into main classes (Methods), leaving 27 diagnostic categories. This enlarged cohort (11,744 samples; Supplementary Table S5) allowed us to increase sample size using data gathered by different sources. Next, we divided samples into two separate cohorts: a cohort (70% of total, $n = 8,209$) used to construct the model and a separate cohort (30% of total, $n = 3,535$) that was never seen by the model during training or optimization and was used later to test performance (Fig. 5B). After dividing the dataset, we used the first cohort for hyperparameter optimization using a grid search and 10-fold cross-validation approach (see Methods). The final architecture was trained and validated 10 times, each time drawing from only the first cohort (60%, $n = 7,037$ for training with 10%, $n = 1,172$ for validation each iteration; Fig. 5B).

When trained on developmental deconvolution scores (214 per sample), the D-MLP classifier reached an overall "top1" prediction concordance of 73% and "top3" prediction concordance of 90% (Fig. 5C). For many tumor categories [e.g., glioblastoma multiforme + brain lower grade glioma (GBMLGG), acute myeloid leukemia (LAML), lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), PCPG, KICH + kidney renal clear cell carcinoma (KIRC) + KIRP (KIDNEY), and sarcoma (SARC)], concordance rates of 92% to 100% were observed (Fig. 5C). Interestingly, BRCA tumors showed high concordance despite being spread across developmental deconvolution (Supplementary Fig. S10), suggesting the D-MLP classifier was capturing information not readily apparent from deconvolution alone. Overall, 24 of 27 tumor types were classified with a higher degree of concordance than by chance. Lower concordance rates were obtained for some tumors [e.g., colon adenocarcinoma + rectum adenocarcinoma (COADREAD), head and neck squamous cell carcinoma (HNSC), aggregated LUAD and LUSC (LUNG), and esophageal carcinoma (ESCA) + stomach adenocarcinoma (STAD), or STES], and very poor results were obtained for CHOL, uterine carcinosarcoma (UCS), and mesothelioma (MESO; Fig. 5C; Supplementary Fig. S11). Poor CHOL classification may be due to the lack of development of a distinct gallbladder in rodents (49), with no trajectory for this cell type in the MOCA data, poor UCS classification due to few training samples, and poor MESO classification due to mesothelioma arising from toxin exposure (asbestos) with no developmental signature. ROC characteristics for classifier performance were strong, with an ROC-AUC of 0.9740 (microaveraged method

for top prediction, 95% confidence interval 0.971–0.977; Fig. 5D) and with high precision (positive predictive value) and recall (true-positive rate) for most tumor types (Fig. 5E), together validating D-MLP effectiveness.

As further validation, we applied D-MLP to annotated malignant cells from scRNA-seq studies. D-MLP had an overall high performance on these cells, with an ROC-AUC of 0.859 (95% confidence interval, 0.826–0.890), confirming its high accuracy on another sample set (Supplementary Fig. S12A). As noted, tumors are composed of malignant cells admixed with normal stroma. To determine how tumor purity affected classification, we generated *in silico* mixes between malignant cells and normal cells. We combined known ratios of these cells from scRNA-seq, summed their gene expression counts, performed developmental deconvolution, and classified mixes using D-MLP. For simulated samples with >20% tumor purity, we found retention of relatively high accuracy, with an ROC-AUC of 0.787 (95% confidence interval, 0.751–0.826) or better. High performance on many tumor types for the more sparsely sampled scRNA-seq data lent confidence to the robustness of the approach. We supplemented this analysis using analogous *in silico* tumor–normal mixes from our bulk-sequenced cohort, selecting high purity tumors with matched patient normal samples. Similar to analysis using single cells, samples with >20% tumor purity showed high performance as measured by ROC-AUC and classification accuracy (Supplementary Fig. S12B).

Next, we compared our developmental deconvolution approach with benchmark classifiers trained directly on gene expression data. First, we selected the most variably expressed genes across all samples, reasoning that variably expressed genes should have the strongest power to resolve tumor samples from one another. We used the 214 most variable genes to match feature counts with D-MLP, trained a new classifier with reoptimized weights, and evaluated its performance on the test set. This classifier (highly variable genes) performed with an ROC-AUC of 0.859 (95% confidence interval, 0.829–0.976; Fig. 5C and Supplementary Fig. S13A and S13B) under D-MLP performance on the bulk-sequenced cohort. Second, we selected a panel of 214 genes tested in diagnostic clinical assays at our institution (e.g., *EGFR*, *MYC*, *KRAS*; see Methods for details), trained a new classifier with reoptimized weights, and evaluated its performance. This classifier (clinical oncopanel genes) performed with an ROC-AUC of 0.836 (95% confidence interval, 0.828–0.975; Fig. 5C and Supplementary Fig. S13C and S13D), again under D-MLP performance on the test set. Additionally, these benchmark classifier approaches had lower overall accuracy. We conclude that developmental deconvolution yields higher accuracy classification than training directly on gene expression data for similar numbers of input features.

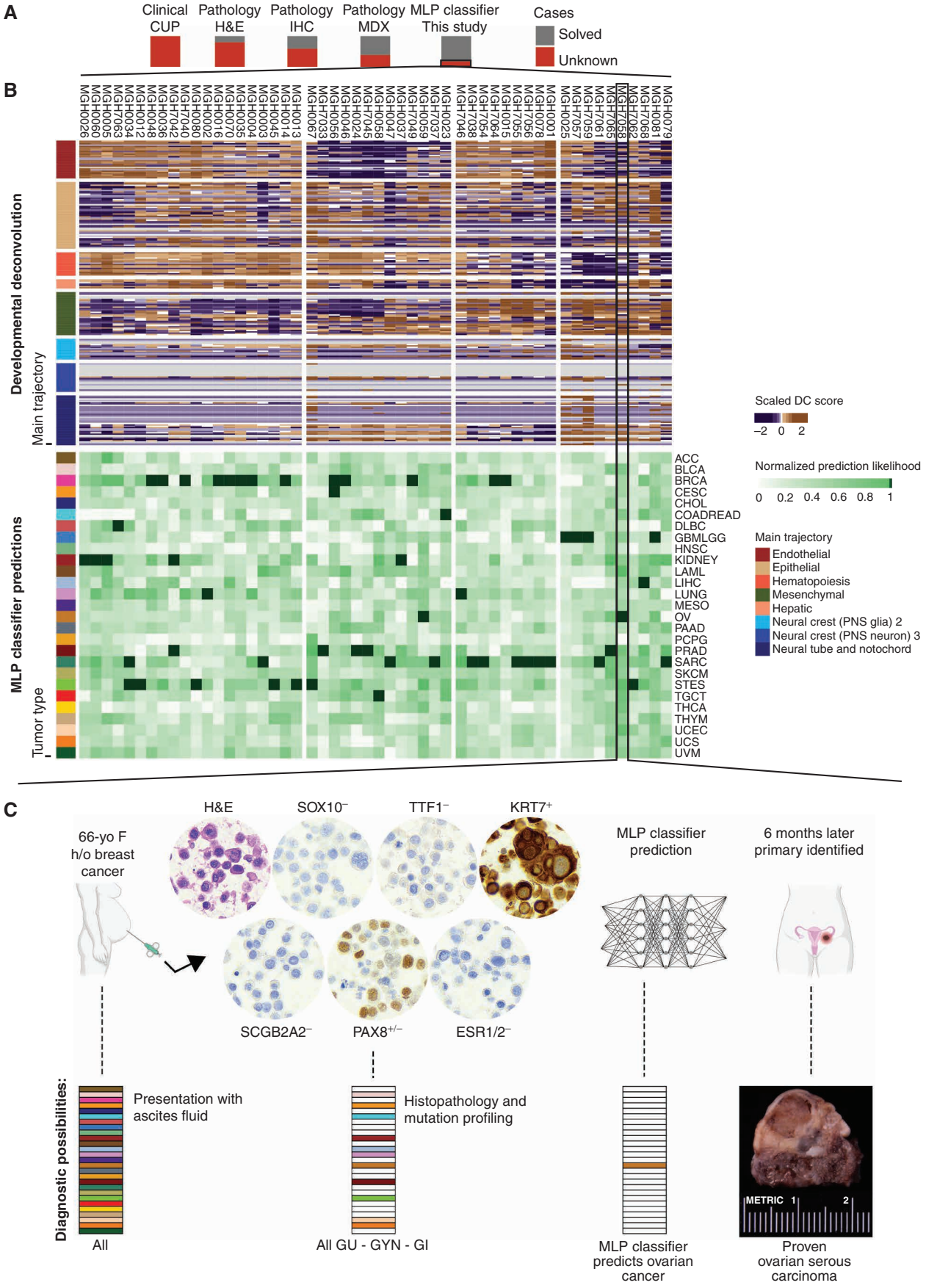
## Analysis of Discordant D-MLP Predictions

In principle, discordant predictions could arise from classifier inaccuracy, false-positive associations, or additional developmental information. Tumor sampling could include nearby tissue, tumors might share previously unappreciated developmental connections, or tumors classified as one histopathologic entity might contain heterogeneity between samples in their developmental origins. To assess these possibilities, we examined discordant classification relationships among tumors (Fig. 5F and G; Supplementary Fig. S11). The adrenal gland rests on top of the kidney within the retroperitoneum. We noted that ACC was often discordantly classified as kidney (12%) or as an adrenal medulla tumor (9%, PCPG). Other examples pointed to tumor heterogeneity. For example, both lung and breast adenocarcinomas are epithelial tumors, both normal tissues are formed by the interaction of ectodermal and mesodermal elements, and both normal tissues continue remodeling into young adulthood (50, 51). We found breast and lung tumors were commonly discordantly classified as each other (Fig. 5F) and shared striking heterogeneity in their signal for epidermis and branchial arch developmental trajectories across samples (Supplementary Fig. S14A). Analysis of classifier predictions showed many discordantly classified samples were close to each other in prediction (Supplementary Fig. S14B), fitting with a model where underlying heterogeneity contributed to discordant classifications. In general, concordant predictions reflected known cancer biology, whereas discordant ones reflected less well-understood connections between tumor types or developmental heterogeneity within tumors.

## Classification of Cancer of Unknown Primary

CUP remains a major clinical problem. Of patients who present to the clinic with CUP, a fraction of diagnoses are resolved using hematoxylin and eosin staining (H&E), a process that can be aided by machine learning tools that rely on image inputs (5). An additional fraction of diagnoses are resolved using IHC and tumor mutation profiling (Fig. 6A). However, a portion of cases fail all currently available modalities and remain true diagnostic dilemmas in need of new approaches (Fig. 6A, far right). In our experience, ~1% of all patient tumors fell into this category at our institution from 2015 to 2021 (Methods). Given the often high-grade, dedifferentiated appearance of these tumors, we reasoned developmental mapping might provide a new diagnostic approach to determining their origins and could provide a classification. We gathered a cohort of 52 such cases representing the most challenging diagnostic dilemmas seen at our institution.

First, we sequenced their transcriptomes and performed developmental deconvolution. Interestingly, developmental deconvolution of CUPs yielded four major clusters characterized by enrichment for different main trajectories (Fig. 6B shows all trajectories and Supplementary Fig. S15A shows enriched trajectories; clusters are numbered left to right): endothelial and hematopoietic (cluster 1), neural tube notochord and hematopoiesis (cluster 2), endothelial and mesenchymal (cluster 3), and mesenchymal and neural tube notochord (cluster 4). Next, we applied D-MLP to each CUP case. The classifier made diagnostic predictions for all patients (Fig. 6B, bottom). Intriguingly, we noted relationships between developmental information and D-MLP prediction for CUPs. Cluster 1 was enriched for BRCA, cluster 2 was not enriched for any particular classification, cluster 3 was enriched for SARC, and cluster 4 was enriched for GBM-LGG (Supplementary Fig. S15B). Together, this suggested that some CUPs could be resolved by transcriptomic profiling, whereas other CUPs may express overlapping developmental programs or be truly dedifferentiated.


A
Clinical CUP
Pathology H&E
Pathology IHC
Pathology MDX
MLP classifier This study
Cases
Solved
Unknown
B
Developmental deconvolution
Main trajectory
MLP classifier predictions
Tumor type
ACC
BLCA
BRCA
CESC
CHOL
COADREAD
DLBC
GBMLGG
HNSC
KIDNEY
LAML
LIHC
LUNG
MESO
OV
PAAD
PCPG
PRAD
SARC
SKCM
STES
TGCT
THCA
THYM
UCEC
UCS
UVM
Scaled DC score
–2 0 2
Normalized prediction likelihood
0 0.2 0.4 0.6 0.8 1
Main trajectory
Endothelial
Epithelial
Hematopoiesis
Mesenchymal
Hepatic
Neural crest (PNS glia) 2
Neural crest (PNS neuron) 3
Neural tube and notochord
C
66-yo F h/o breast cancer
H&E
SOX10⁻
TTF1⁻
KRT7⁺
SCGB2A2⁻
PAX8⁺/⁻
ESR1/2⁻
MLP classifier prediction
6 months later primary identified
Diagnostic possibilities:
Presentation with ascites fluid
Histopathology and mutation profiling
All
All GU - GYN - GI
MLP classifier predicts ovarian cancer
Proven ovarian serous carcinoma
METRIC 1 2

If available clinically, this developmental information could have supplemented diagnostic decision-making. For example, case MGH058 was a 66-year-old female with a history of breast cancer who presented with peritoneal ascites. Fluid was drained and examined by cytology, revealing high-grade features (nuclear pleiomorphism, mitotic figures) on H&E (Fig. 6C). IHC stains were negative for breast markers (mammaglobin-A "*SCGB2A2*," estrogen receptor "*ESR1/2*"), lung marker *TTF1*, and melanoma marker *SOX10*, positive for epithelial markers (keratin, type II cytoskeletal 7, or "*KRT7*"), and variable/weakly positive for genitourinary origin (*PAX8*; Fig. 6C). Molecular profiling was positive for variants in *RB1* and *TP53*. This left a very large differential diagnosis with no further tools to narrow it. Analyzing cells from ascites, the D-MLP classifier gave a strong prediction of ovarian cancer for this case. Six months later, and after extensive additional clinical workup, the patient was found to have a mass proven to be ovarian serous carcinoma. We conclude that deep learning classifiers based on developmental deconvolution could serve as a useful adjunct, impacting diagnosis and clinical decision-making.

## DISCUSSION

Our analysis compared TCGA tumor samples and MOCA single-cell data to construct a developmental map of human tumors, systematically quantifying parallels between cancer biology and developmental programs. We used this map to deconvolute tumors into DCs, which in turn allowed us to construct a D-MLP classifier capable of high-accuracy tumor type prediction. Together, this constitutes a proof of principle for how the integration of developmental and tumor signatures can be used to aid in clarifying the diagnoses of otherwise unclassified tumor entities.

Many clinical cases remain diagnostic challenges, and although image-based tools have shown great promise in narrowing differentials, they rely on visual input. Gene expression has the potential to add orthogonal information, but generating models with true predictive power has been difficult, owing to the challenge in extracting the most relevant information. Our approach used developmental trajectories to dimensionally reduce gene expression data. Projecting tumor data onto axes of reduced dimensionality defined through developmental programs, instead of defined through gene expression, increased the accuracy of classification. Another benefit of a developmental approach is that this focus can reveal new tumor biology or new tumor classification schemes. Based on developmental profiling, we identified four putative clusters of CUPs, which could form the basis of a classification scheme for these tumors. Assembling a larger multi-institutional cohort will allow validation of these categories and may allow the identification of marker genes that could be used to separate them.

However, although powerful, our approach has drawbacks. Not all information germane to tumor classification is developmentally related, leading to inaccuracy. Additionally, although the correlation between TCGA and MOCA showed a strong signal compared with shuffled controls, we cannot formally exclude that noise contributes to classification inaccuracies or leads to false-positive associations. Ultimately, careful experimental validation of novel associations using disease-focused cohorts and cancer model systems will be needed. Another limitation is that the current reference dataset was taken from a different species (*Mus musculus*), as this was the largest single-cell developmental reference atlas available at the time of study capturing developmental periods when adult mammalian lineages are specified. Although mice have proven useful models for human cancer, it will be of great interest to construct new classifiers using human single-cell datasets, particularly those focusing on earlier fetal development, such as may emerge from the Human Cell Atlas. The approach described here will generalize to such datasets as they become available.

The approach in the present study focused on broad categories of malignancies. Yet many current diagnostic dilemmas in pathology focus on distinguishing different entities within one tissue type rather than distinguishing different tissue types from one another. New classifiers, trained on extensive datasets within one tissue type, may be able to distinguish these entities on the basis of their development. Our comparisons suggested underlying differences in developmental components between tumor and normal tissue. Classifiers focused on distinguishing benign entities from malignant ones would be of great use in pathology, especially in histopathologic gray zone cases, as clinical management decisions often turn on whether an entity is classified as benign or malignant.

**Figure 6.** Diagnosis of CUP by a D-MLP classifier. **A,** Patients present clinically with CUP. Some cases are solved by H&E examination of tissue. Additional cases are solved by IHC stains that mark particular tissue types and by molecular techniques (MDX) such as mutation profiling. However, a subset of CUP remains undiagnosed with no primary site assigned after all available techniques. We applied D-MLP to a cohort of 52 such cases from our institution. **B,** Top, the developmental deconvolution profile of each CUP case is shown by plotting DC scores. Trajectories are arranged top to bottom, as they are shown counterclockwise on radar plots (Fig. 3; Supplementary Fig. S5), and colors for main trajectories are as in Figs. 2–5. See Supplementary Fig. S15A for a plot of differential trajectories. Bottom, D-MLP classifier predictions for 52 CUP cases. Note higher confidence predictions are weighted in coloration (dark green). Colors for tumor types are as in Figs. 2–5. PNS, peripheral nervous system. **C,** Case MGH058 is highlighted for further consideration. A 66-year-old female with a history of breast cancer presented with ascites fluid accumulation. Fluid was drained and assessed by standard cytologic/histopathologic workup. Stains are shown for H&E (morphology), mammaglobin-A (*SCG2A2*, breast cancer), estrogen receptor (*ESR1/2*, breast cancer), *TTF1* (lung), *SOX10* (melanoma), keratin, type II cytoskeletal 7 (*KRT7*, epithelial origin), and *PAX8* (broad marker, primarily genitourinary), which ruled out breast cancer but left a broad differential diagnosis encompassing genitourinary (GU), gynecologic (GYN), and some gastrointestinal (GI) malignancies. D-MLP classifier gave a strong prediction of ovarian cancer for this patient's ascites. Six months later, after extensive workup, the patient underwent bilateral salpingo-oophorectomy and was found to have a mass (pictured) identified as ovarian serous carcinoma. ACC, adrenocortical carcinoma; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, cholangiocarcinoma; COADREAD, colon adenocarcinoma (COAD) + rectum adenocarcinoma (READ); DLBC, lymphoid neoplasm diffuse large B-cell lymphoma; GBMLGG, glioblastoma multiforme (GBM) + brain lower grade glioma (LGG); HNSC, head and neck squamous cell carcinoma; KIDNEY, kidney chromophobe (KICH) + kidney renal clear cell carcinoma (KIRC) + kidney renal papillary cell carcinoma (KIRP); LAML, acute myeloid leukemia; LIHC, liver hepatocellular carcinoma; LUNG, lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC); MESO, mesothelioma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PCPG, pheochromocytoma and paraganglioma; PRAD, prostate adenocarcinoma; SARC, sarcoma; SKCM, skin cutaneous melanoma; STES, esophageal carcinoma (ESCA) + stomach adenocarcinoma (STAD); TGCT, testicular germ cell tumors; THCA, thyroid carcinoma; THYM, thymoma; UCEC, uterine corpus endometrial carcinoma; UCS, uterine carcinosarcoma; UVM, uveal melanoma.

In TCGA, not all patients have normal samples matched to tumor, and thus TCGA may not adequately represent unaffected regions of all patients. As larger cohorts of normal samples become available, these will boost diagnostic accuracy when used as reference datasets for classifiers that focus on distinguishing benign from malignant entities.

One challenge in building deep machine learning classifiers for CUPs, especially for the most challenging cases, is that no current gold standard exists against which to compare predictions. Perhaps a combination of models, such as those that analyze both imaging data and molecular features, will prove to be most useful in achieving precision cancer care. Ultimately, prospective studies will need to be done to demonstrate the benefit of machine learning approaches. The results presented here give a developmental map of human tumors and suggest a new tool for decreasing diagnostic uncertainty in pathology with implications for the diagnostic classification of cancer.

## METHODS

### Data Gathering and Sample Cohorts

***MOCA.*** The expression profile and meta information of cells analyzed in ref. 15 RNA-seq (gene_count_cleaned.RDS) and annotation (cell_annotation.csv) data were manually downloaded from https://oncoscape.v3.sttrcancer.org/atlas.gs.washington.edu.mouse.rna/downloads. For this study, we used the expression data of the 1,331,984 high-quality cells defined in the MOCA study. Briefly, MOCA study filtering criteria were as follows: Cells with less than 400 detected mRNA molecules were removed, all detected doublet cells were removed, and all cells from doublet-derived subclusters were removed. In the MOCA study, the authors identified 10 main trajectories and 56 subtrajectories, which were noncontinuous, based on transcriptional similarities between the analyzed cells and literature-curated marker genes. Further information is available in the MOCA annotation file (15).

***TCGA.*** The coding gene expression profile (RNAseqV2_RSEM_genes_normalized_data_Level_3) and clinical information (Merge_Clinical.Level_1.2016012800.0.0) of TCGA samples (release 2016_02_28) were systematically downloaded using the firehose_get v 0.4.1 tool, from the Broad TCGA GDAC site (https://gdac.broadinstitute.org/). This contains the data and the analytic categories used by the TCGA consortium, including for aggregating data for the following tumor types: COADREAD, GBMLGG, KIDNEY, and STES. The authors created LUNG by aggregating LUAD and LUSC as in ref. 52.

***NON-TCGA Sample Cohorts.*** The "NON-TCGA" cohort refers to samples obtained from various cancer studies, for which gene expression profiles and clinical information were retrieved from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). The full list of samples and relative studies used is given in Supplementary Table S5, and their conversion to 27 TCGA diagnostic categories for purposes of classification is given in Supplementary Table S6. Merging of TCGA categories allowed the incorporation of the maximum number of non-TCGA samples into this study, as non-TCGA studies used slightly different diagnostic categories.

***Massachusetts General Hospital Sample Cohort.*** Samples from FFPE tissues were chosen from cases seen in the Center for Integrated Diagnostics in the Department of Pathology at Massachusetts General Hospital (MGH) either with known diagnosis (33 cases) or as CUP (52 cases). Total nucleic acid was isolated from six scraped blank slides using clinically validated protocols. At MGH, over the period from 2015 to 2021, 34,782 tumors were seen as reported by our institution to reporting agencies. Of these, 261 tumors (=0.7%) never received a primary site through their final diagnosis, as estimated by case coding. Fifty-two of these cases were retrievable with extracted nucleic acid for further testing by the developmental deconvolution classifier, and their analysis is the subject of Fig. 6.

***Single-Cell Cancer Studies.*** We used the expression profile of normal and malignant single cells from 13 tumor types across 17 different studies. The list of studies is given in Supplementary Table S3. All of the above studies were used to generate the pseudobulk cohorts used to test how purity affects D-MLP prediction as seen in Supplementary Figs. S7B and S12A. Studies #2, #3, #8, #10, and #11 (Supplementary Table S3) were used for testing developmental deconvolution at the single-cell level as seen in Fig. 3B and Supplementary Fig. S7A. Study #7 was used in Fig. 3C. Filtered, quality-controlled expression data and metadata for the cells used in the study were downloaded from the Curated Cancer Cell Atlas 3CA website (https://www.weizmann.ac.il/sites/3CA/) as cited in ref. 52.

***Human Fetal Organs Single-Cell Dataset.*** Expression data of 377,456 single cells from 15 human fetal organs (HFO; gene_count_sampled.RDS) and relative meta data (df_cell.RDS) were downloaded from https://descartes.brotmanbaty.org/bbi/human-gene-expression-during-development/ as cited in ref. 17.

### Cancer Single-Cell Sequencing Studies Data Processing

From the 17 studies representing 13 different tumor types, we created the following subcohorts:

1. For each of the 205 patients representing 13 different tumor types, we aggregated weighted sum counts of each normal and malignant cell population together (see the next section, "*In silico* generation of tumor–normal mixed sample cohort," for details). This cohort was used to test the D-MLP accuracy at various tumor purity levels.
2. Ten of the most abundant nonmalignant cell types were selected from the single-cell sample cohort. For each patient that had at least one of the selected cell types, the expression counts of all cells of the same cell type were summed together to create a per–cell type, per-patient, pseudobulk expression profile for a total of 860 samples. This cohort was used to test the overall capability of developmental deconvolution to separate different normal cell types as seen in Supplementary Figs. S6B and S7B. The data of this cohort can be found in the following files: normal_pseudobulk_meta.csv, normal_pseudobulk_dc.csv, and data_fig7b.rsave.
3. From a random cohort of 10 patients [two of each of the following tumor types: BRCA, COADREAD, LUNG, ovarian serous cystadenocarcinoma (OV), and PAAD], 100 random T cells and 100 random fibroblasts were chosen from each, forming a total of 2,000 individual unique single cells (1,000 T cells and 1,000 fibroblasts). These were then used to test the quality of the developmental deconvolution on isolated cells as shown in Fig. 3B and Supplementary Figs. S6A and S7A. The files of this cohort can be found in data_fig3bc_fs6a_fs7a.rsave.
4. All the malignant and nonmalignant cells from two LIHC samples were individually correlated with the MOCA dataset to test the effect of normal and malignant cell compartments on DC profiles of an aggregated sample as shown in Fig. 3C. The data are available in the file lihc_single_cell_cohorts.csv.

### In Silico *Generation of Tumor–Normal Mixed Sample Cohort*

To test the effect of tumor purity on classifier accuracy (Supplementary Fig. S12), we took two approaches: (i) mixing of matched primary tumor and normal bulk TCGA samples and (ii) pseudobulk of tumor and normal mixing of single cells of known malignant or nonmalignant type from the same sample. For approach 1, the TCGA contains 678 primary tumor samples for which normal matched samples are available, across 17 different tissues: BLCA, BRCA, cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), CHOL, COADREAD, HNSC, KIDNEY, LIHC, LUNG, PAAD, PCPG, prostate adenocarcinoma (PRAD), SARC, STES, thyroid carcinoma (THCA), thymoma (THYM), and uterine corpus endometrial carcinoma (UCEC). For every primary tumor with a matched normal sample, the gene expression profiles were processed to create a gradient of mixed gene expression ranging from 100% tumor samples to 100% normal tissue with 10% increments (90% tumor + 10% normal, 80% tumor + 20% normal, etc.). See formula (1) in Supplementary Formulas for details. For approach 2, all the counts from malignant and nonmalignant cells from the same patient were summed to create a malignant pseudobulk and a normal pseudobulk. The expression profiles of the pseudobulked normal and malignant compartment were then mixed with an analogous strategy to that used for TCGA tumor–normal patient samples, testing purity content ranges from 100% malignant only pseudobulk to 100% nonmalignant expression profile with 10% increments. See Supplementary Formula (2) for further details. These *in silico*–generated samples were correlated with MOCA cells, deconvoluted, and inputted in the D-MLP classifier. The result of this analysis is shown in Supplementary Fig. S12. Details of these cohorts can be found in the files data_figs12_a.csv (single cell) and data_figs12_b.csv (bulk tumor–normal mix).

### *RNA Extraction from FFPE Clinical Samples*

Libraries were prepared using a modified version of the Takara SMARTer Stranded Total RNA-seq Kit–Pico Input Mammalian kit. In brief, 100 ng of RNA at 10 ng/μL was sonicated using RL230 Covaris sonicator (Covaris), and the resulting material was confirmed using a Fragment Analyzer (Agilent). Ten nanograms of each sonicated sample was prepared using the pico input kit as for FFPE samples using a 1:8 volume reduction on the STP MosquitoHV. Final libraries were validated by Fragment Analyzer and qPCR prior to sequencing on a NovaSeq6000 S4 with 150 nt paired-end reads.

### *RNA-seq Analysis of FFPE-Derived Clinical Samples*

Reads obtained from the sequencing step previously described were processed as follows: A STAR reference genome using GENCODE v35 fasta and gtf files was generated using STAR "genomeGenerate." Next, fastq files were aligned to the genome generated with STAR using two pass mapping (see mgh_sequencing.sh for details on STAR parameters). This step generated bam files compatible with RNA-seq by expectation-maximization (RSEM) gene expression calculation. An RSEM reference was prepared using the rsem-prepare-reference command.

These files along with the bam files were used to calculate gene expression with rsem-calculate-expression using parameters (-p 16, –bam, –paired-end, –no-bam-output, –forward-prob 0.5, and –seed 12345).

Gene expression measured in transcripts per million (TPM) was used to assess the similarity with the MOCA dataset as described in the "Similarity Score Calculation" section. Homo sapiens primary assembly v35_GRCh38.p13 genome and relative annotation GTF file (v35) were downloaded from the GENCODE website (https://www.gencodegenes.org/). STAR v2.7.1a alignment tools, RSEM v1.3.1, R v3.6.0 (https://www.R-project.org/), and Perl v5.24.1 were used for gene expression analysis. The details of the commands and parameters used to generate the gene expression matrix from fastq are also contained in the file mgh_sequencing.sh.

### *Mouse–Human Gene Name Conversion*

To compare murine gene expression from MOCA to human gene expression in TCGA, we standardized gene names between mouse and human using the following approach. The conversion from Mouse Ensembl id (MOCA) to human gene symbol/Entrez id (TCGA) was achieved using BioMart (https://www.ensembl.org/biomart/martview/) Ensembl v95. Human gene symbol/Entrez id (TCGA) to Ensembl id (NON-TCGA) mapping was achieved using *org.Hs.egENSEMBL2EG* from the org.Hs.eg.db (v3.8.2) Bioconductor (v3.9; https://bioconductor.org/) package. The intersection of these two sources was performed using the human gene symbol/Entrez id shared identifier. This process generated a list of translated names, given in Supplementary Table S7. This list was then used as a dictionary for gene names and mouse–human ortholog comparison. This identified 15,929 unique human genes that were used in this study. In the case of multiple mouse gene names mapping to the same human gene name, the average expression levels were calculated across occurrences.

### *Similarity Score Calculation*

The similarity between gene expression profiles from either MOCA or HFO cells and bulk (TCGA, NON-TCGA, and MGH samples) or single cells from cancer datasets was calculated by means of Spearman correlation coefficient [Supplementary Formula (3)], implemented using the *cor* function in R on the expression profile of all shared genes identified as described above for each bulk/single-cell sample and MOCA cell. We decided to use the Spearman correlation coefficient because this nonparametric, rank-based approach is more robust to outliers caused by single-cell transcript dropout and is unaffected by the normalization method, which standardized the use of different gene expression datasets.

Spearman correlation generates a matrix $A$ of correlation coefficients of dimensions $I \times J$, where $I = 1...N$ represents the cells from the MOCA study ($N$ = 1,331,984) and $J = 1...M$ represents the comparing study's sample number [see Supplementary Formula (4)]. The matrix of correlation coefficients for MOCA/TCGA samples, $A_{MOCA/TCGA}$, has $P$ = 10,393 (9,274 primary tumors, 394 metastasis and 725 normal tissues), so $A_{MOCA/TCGA}$ is a 1,331,984 × 10,393 matrix, containing 1.38e+10 correlation coefficients ($A_{HFO/TCGA}$ is a 377,456 × 10,393 matrix containing 3.92e+09 correlation coefficients, $A_{MOCA/MGH}$ is a 1,331,984 × 85 matrix). The correlation coefficient was then used as a metric for the similarity between the samples in the cohorts under exam and was further processed as described in the section "Similarity Score Aggregation, Scaling, and Normalization" next.

The code to calculate the similarity score via the correlation coefficient can be found in the script 01_correlation.R.

### *Similarity Score Aggregation, Scaling, and Normalization*

To generate TCGA-aggregated similarity scores shown in Fig. 2B and C and Supplementary Figs. S1D, S2, S3, S4A, and S15 using dimensions of matrices $A_{MOCA/TCGA}$, $A_{HFO/TCGA}$, and $A_{MOCA/MGH}$, the following steps were performed. Beginning with an $N \times M$ ($N$ = number of MOCA or HFO cells, $M$ = number of samples from TCGA, non-TCGA, or single-cell studies) matrix $A_{IJ}$, where $I = 1,...,N$ and $J = 1,...,M$. Each column $J$ of $A_{MOCA/TCGA}$ or $A_{FHO/TCGA}$ was mean centered and standard deviation ($\sigma$) scaled [see Supplementary Formula (5)], resulting in matrix ($_{scj}A_{IJ}$) calculated using the scale() R function.

The scaled values for every sample belonging to the same tissue type ($B_p = 1...P$), where $P$ is the number of samples belonging to the tissue type $B$, as defined by TCGA, is then averaged, to create the matrix $A'_{IK}$ $N \times K$ [$K$ = number of unique sample types, i.e., 62 for TCGA; see Supplementary Formula (6)], meaning the new matrix has the same number of rows ($N$) but a reduced number of columns ($K$). Every cell belonging to the same subtrajectory $C$ with $Q$ number of cells belonging to the subtrajectory $C$ is then averaged across all aggregated ($C = 1...Q$) tissue type $K$ in $A'_{IK}$ creating the matrix $A''_{LK}$ $L \times K$ [see Supplementary Formula (7)]. $A''_{LK}$ for MOCA and TCGA samples now consists of 56 subtrajectories × 62 tissue type scores (derived from 33 tumor types for which a combination of primary, metastatic, and normal tissues are available). A column-wise mean-centered and standard deviation–scaled version of $A''_{LK}$ is calculated as in scaled similarity scores using Supplementary Formula (5). Scaled values are further min–max normalized [see Supplementary Formula (8)] to change the range of the scaled similarity scores to the plotted interval [e.g., (0–1), or (−1, +1)] as shown in the figures. See script 02_data_recap.R for further details.

### Pan-Cancer Comparisons of Tumor–Normal Tissues and Embryonic Period

The developmental time difference between normal and tumor tissues, shown in Fig. 2E, was calculated according to Supplementary Formula (9). For every TCGA sample from tissue types that contained the expression profile of at least one normal sample, the number of the most strongly correlated cells (top 1,331 cells, sorted by correlation coefficient) were binned by their embryonic period of origin (E9.5–E13.5). This created the matrix T of dimension $I \times J$, where $I$ represents either normal or malignant samples and $J$ represents the five embryonic time periods (E9.5, E10.5, E11.5, E12.5, and E13.5), containing in each matrix entry the number of MOCA cells in the given category $T_{ij.}$ This matrix was then analyzed using the $\chi^2$ test [Supplementary Formula (10)] to produce Fig. 2E. The developmental period enrichment represents the test residuals calculated as (observed − expected)/sqrt(expected). For Fig. 2F, the actual embryonic day (represented as an integer, 9.5, 10.5, and so on) was first multiplied by the number of cells in its relative column $J$ and then these values were added together, providing a per-sample measure characterizing the development period of its transcriptional profile [Supplementary Formula (11)].

### Deconvolution by DCs

To perform the deconvolution into DCs, the following steps were performed [Supplementary Formula (12)]: For each TCGA, NON-TCGA, and MGH bulk and single-cell sample, the MOCA cells were sorted in increasing (lowest to highest) order based on their correlation coefficient. Next, the 1,331 most strongly correlated cells were selected, representing the top ~0.1% of all MOCA cells tested. A rank-based score was then assigned to this selection of cells. The most highly correlated cell was given a score of 1,331, whereas the least correlated score was given a score of 1. These scores were then summed across all cells belonging to the same combination of subtrajectory at a particular developmental time, creating the raw DC score. The raw scores were then transformed by taking the natural logarithm (ln). The correlation between the DCs and sets of samples was calculated using the Kruskal–Wallis rank sum test [Supplementary Formula (13)] using the kruskal.test() R function. Only DCs with a Benjamini–Hochberg adjusted $P < 0.05$ were considered statistically different between groups.

### D-MLP

The D-MLP is a supervised deep learning model trained on natural log-transformed and min–max normalized DC scores that outputs a likelihood score for each of the 27 aggregated TCGA tumor classes. To ensure reproducibility, the minimum and maximum values of the aggregated TCGA, NON-TCGA, and MGH cohorts were calculated and used as the minimum and maximum for all min–max normalization, including for the normalization of the CUP cohort. The model's hyperparameters were identified by means of grid search over the following variables: (i) number of hidden layers, (ii) number of nodes per layer, (iii) type of optimizer function, (iv) type of loss indicator, and (v) number of epochs for which the model is trained as shown in Supplementary Table S8. The architecture selection and the training/validation of the model were performed by a 10-fold cross-validation of a training–validation set split of 60% to 10% of the totality of the TCGA, NON-TCGA, and MGH cohorts. The performance of the D-MLP was tested on the remaining test set (30% of the whole cohort). None of the samples present in the test set were used during the model training or during the architecture selection. Further, the performance of the D-MLP was tested on an independent cohort of single cell–derived pseudo bulk samples, described above. The final model has the following architecture: one input layer with 214 nodes, two hidden layers of 800 and 200 nodes, respectively, and one output layer with 27 nodes. The model was compiled using stochastic gradient descent as optimizer, mean-squared error as loss indicator, and accuracy as metric. The model was trained on the test set for 300 epochs with an early stop function monitoring accuracy score, with a patience of 3. The perceptron was written in python (v3.6.4) using sklearn (v0.19.1) and keras (v2.2.0, with TensorFlow backend). See script d_mlp_classifier.py for code details.

### Classification Analysis

The raw likelihood score resulting from the classification of the test set, CUP cohort, tumor purity cohort, and benchmark cohorts were each analyzed as follows: The output of the D-MLP classifier is a matrix containing a number of rows equal to the number of samples analyzed and columns equal to the 27 classification labels ($N \times M$ matrix). Each sample's top classification (defined by highest likelihood score) is assigned as top1, the next as top2, and so on (up to 4+ as shown in Fig. 5 and Supplementary Fig. S12). For each tumor type ($M$), a frequency table is generated of the number of top1/top2/top3/4+ predictions over $N$ occurrences. To compare the output with the true tumor labels of the $N$ samples, one-hot encoding is used to create another $N \times M$ matrix, where each row $n$ column $m$ entry is 1 if the given sample $n$ is labeled with tumor $m$ in the original dataset, and 0 otherwise. The output classification matrix together with the encoded matrix is then used to calculate the true-positive rate, false-positive rate, and ROC-AUC. The confidence interval for the ROC-AUC scores was calculated over 1,000 bootstraps. See also script test_classifier.py for details. The results of the classification of the various cohorts are shown in Figs. 5 and 6 and Supplementary Figs. S12, S13, and S15.

### Grouping of Discordant Predictions

To assess the distance between top3 classification results of discordantly classified samples and concordantly classified samples, the following "hot-encoder" approach was taken. Each sample's classification response was encoded using a vector of length 81 (27 possible prediction labels with a spot for each of the top3 predictions). Each element of this vector was then populated with a score of 5, 3, 2, or 0, depending on whether the correct answer was in the top1, top2, top3, or top4 (all other predictions) category. A top1 (correctly classified tumor) would have a score of 5 repeated 3 times, a top2 correctly classified sample would have a score of 3 in the range of 28 to 54, and so on. For example, an ACC sample correctly classified in top1 would have a score of 5 in position 1, 28, and 55 (and 0s everywhere else), whereas an ACC sample guessed right

in top2 would have a score of 3 in position 28 and 0s everywhere else. These vectors were then compared using cosine similarity [see Supplementary Formula (14)], and UMAP was used to plot the results of cosine similarity analysis returning the image shown in Supplementary Fig. S14B. See script UMAP_plot_fig4ef_figs10.R for further details.

### CUP Clustering and DC Analysis

Raw DC scores of the CUP cohort were processed as follows. Each DC score was mean centered and standard deviation scaled (Z-scored) across the whole cohort of 52 CUPs. After scaling, Spearman distance was calculated between different samples (see Supplementary Formulas; ref. 15). This distance was evaluated by means of the Dist() function in the amap R package. The function outputs an $N \times N$ matrix with the pairwise distance between $N$ samples. A hierarchical clustering analysis using the "ward.D" algorithm was then calculated on this distance matrix using hclust() R function and cut into four main clusters (chosen based on the observed distance between branch points) using cutree() R function. Statistical analysis of the differential developmental programs between the four clusters was performed by the Kruskal–Wallis test using kruskal.test() in R. Enrichment for specific classifications was performed using the $\chi^2$ test. Seventy DCs with a Bonferroni corrected $P < 0.05$ were considered correlated with at least one cluster and are shown in Supplementary Fig. S15A. See script fig6_figs15.R for further details.

### Benchmark Classifiers and Performance

The performance of the D-MLP classifier was tested against models sharing the same architecture trained on pure gene expression profiles directly without developmental deconvolution. We opted for two sets of genes: (i) clinical oncopanel genes and (ii) highly variable genes. Clinical oncopanel genes represent a list of 251 genes tested in routine clinical cancer care at MGH (assays: SNaPshot, Solid Fusion Assay, Heme SNaPshot); the full list is given in Supplementary Table S8. To match feature counts with the D-MLP classifier, we generated 10 random subsets of 214 genes out of 251. Each of these 214 gene subsets was used to train a benchmark classifier, and the highest performing assessed by top1 and top3 accuracy was directly compared against the developmental based classifier (D-MLP) as shown (Fig. 5C; Supplementary Fig. S13). For the highly variable gene benchmark classifier, the 214 most variably expressed genes assessed by pure variance [using the var() function in R] across the full pan-cancer cohort (TCGA, NON-TCGA, and MGH) taken altogether were used. Expression of these genes in TPM was used as input for the benchmark classifier.

### Similarities between TCGA Correlations with MOCA and HFO

In order to test the specificity and reproducibility of the correlation between the mouse expression profiles and human expression profiles, we adopted the following approach. TCGA expression profiles were correlated with HFO cells as previously described. The HFO cell types were then mapped to the MOCA subtrajectories according to reference (17). The correlation coefficient between the two similarity matrices (each $56 \times 62$, see formulas) was then calculated by cor.test() R implementation. For per TCGA sample type correlations, the same function was applied across the appropriate column of the matrices. See script figs3.R for details.

### Statistical Analysis

Statistical analyses reported in this work were performed using R (v3.6.3). Enrichment (Fig. 2E) was calculated using the $\chi^2$ test and represented as (observed − expected)/sqrt(expected). Statistical differences between cumulative distributions were evaluated using the Kolmogorov–Smirnov tests in Fig. 2F. Pairwise differences between means of continuous variables from different samples were evaluated using Mann–Whitney tests in Supplementary Fig. S4B. ROCs were calculated using the roc_curve and roc_auc functions from Python3 sklearn (v0.22.1) using 1,000 bootstraps to calculate empirical 95% confidence intervals.

### Figures

Box plots, empirical cumulative distribution function plots, violin plots, scatter plots, and bar plots were generated using the ggplot2 (v3.3.2) package. Heat maps were generated with pheatmap (v1.0.12) package. UMAP plots were generated with plot() native R as given in the scripts for individual figures. Radar plots were generated in R using a modified function from the fmsb R package. Sankey plots were generated with ggplot2 and ggalluvial (v0.12.2). Graphical representations were generated using BioRender.com (https://biorender.com/).

### Data and Code Availability

Codes and file intermediates generated in this study are available at https://github.com/emoiso/DevTum. Original data are available on the TCGA, MOCA, and HFO websites (14, 15, 17). Note that the D-MLP artificial intelligence algorithm cannot be used in routine practice, as it is not approved by the FDA.

## Authors' Disclosures

E. Moiso reports other support from the Ludwig Center at MIT's Koch Institute during the conduct of the study, as well as a patent for diagnosis of malignancy using developmental relationships and machine learning pending. S. Garg reports grants from the NCI during the conduct of the study, as well as a patent for diagnosis of malignancy using developmental relationships and machine learning pending. No disclosures were reported by the other authors.

## Authors' Contributions

**E. Moiso:** Conceptualization, data curation, software, formal analysis, validation, investigation, visualization, methodology, writing–original draft, writing–review and editing. **A. Farahani:** Resources. **H.D. Marble:** Resources. **A. Hendricks:** Methodology. **S. Mildrum:** Methodology. **S. Levine:** Methodology. **J.K. Lennerz:** Resources, formal analysis, writing–review and editing. **S. Garg:** Conceptualization, supervision, funding acquisition, investigation, methodology, writing–original draft, writing–review and editing.

## Acknowledgments

We thank Kelli Burke for administrative assistance. We thank Phillip A. Sharp, Jacqueline A. Lees, Amanda J. Whipple, and members of their laboratories for helpful discussions. This work was supported by a Charles W. and Jennifer C. Johnson Clinical Investigator Award (S. Garg), NCI K08-CA237856 (S. Garg), NCI R37-CA225655 (J.K. Lennerz), and NCI P30-CA14051 (Koch Institute core). E. Moiso acknowledges support from a Ludwig Fellowship (Koch Institute). This work was performed under Mass General Brigham Institutional Review Board #2014P000940 and MIT Committee on the Use of Humans as Experimental Subjects E-2066.

The publication costs of this article were defrayed in part by the payment of publication fees. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

## Note

Supplementary data for this article are available at Cancer Discovery Online (http://cancerdiscovery.aacrjournals.org/).

Received October 28, 2021; revised May 31, 2022; accepted August 26, 2022; published first August 30, 2022.

## REFERENCES


1. Esserman LJ, Thompson IM Jr, Reid B. Overdiagnosis and overtreatment in cancer: an opportunity for improvement. JAMA 2013;310:797–8.
2. Hasserjian RP, Ott G, Elenitoba-Johnson KS, Balague-Ponz O, de Jong D, de Leval L. Commentary on the WHO classification of tumors of lymphoid tissues (2008): "gray zone" lymphomas overlapping with Burkitt lymphoma or classical Hodgkin lymphoma. J Hematop 2009;2:89–95.
3. Potts SJ, Krueger JS, Landis ND, Eberhard DA, Young GD, Schmechel SC, et al. Evaluating tumor heterogeneity in immunohistochemistry-stained breast cancer tissue. Lab Invest 2012;92:1342–57.
4. Kohli M, Prevedello LM, Filice RW, Geis JR. Implementing machine learning in radiology practice and research. AJR Am J Roentgenol 2017;208:754–60.
5. Lu MY, Chen TY, Williamson DFK, Zhao M, Shady M, Lipkova J, et al. AI-based pathology predicts origins for cancers of unknown primary. Nature 2021;594:106–10.
6. Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: challenges and opportunities. Med Image Anal 2016;33:170–5.
7. Kothari S, Phan JH, Wang MD. Eliminating tissue-fold artifacts in histopathological whole-slide images for improved image-based prediction of cancer grade. J Pathol Inform 2013;4:22.
8. Rajesh Kumar R, Ajith Kumar VK, Sharath Kumar PN, Sudhamony S, Ravindrakumar R. Detection and removal of artifacts in cervical cytology images using support vector machine. In: 2011 IEEE International Symposium on IT in Medicine and Education; 2011 Dec 9–11; Guangzhou, China. Piscataway (NJ): IEEE; 2011. p. 717–21.
9. Cheng J, Zhang J, Han Y, Wang X, Ye X, Meng Y, et al. Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis. Cancer Res 2017;77:e91–e100.
10. Hao J, Kosaraju SC, Tsaku NZ, Song DH, Kang M. PAGE-Net: interpretable and integrative deep learning for survival analysis using histopathologic images and genomic data. Pac Symp Biocomput 2020;25:355–66.
11. Fakoor R, Ladhak F, Nazi A, Huber M. Using deep learning to enhance cancer diagnosis and classification. Proc Mach Learn Res 2013;28:1–7.
12. Hwang KB, Cho DY, Park SW, Kim SD, Zhang BT. Applying machine learning techniques to analysis of gene expression data: cancer diagnosis. In: Lin SM, Johnson KE, editors. Methods of microarray data analysis. New York: Springer; 2002. p. 167–82.
13. Tan AC, Gilbert D. Ensemble machine learning on gene expression data for cancer classification. Appl Bioinformatics 2003;2:S75–83.
14. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas pan-cancer analysis project. Nat Genet 2013;45:1113–20.
15. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature 2019;566:496–502.
16. Srivatsan SR, McFaline-Figueroa JL, Ramani V, Saunders L, Cao J, Packer J, et al. Massively multiplex chemical transcriptomics at single-cell resolution. Science 2020;367:45–51.
17. Cao J, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, et al. A human cell atlas of fetal gene expression. Science 2020;370:eaba7721.
18. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell 2018;173:291–304.
19. Bussolati B, Grange C, Camussi G. Tumor exploits alternative strategies to achieve vascularization. FASEB J 2011;25:2874–82.
20. Sell S. Cellular origin of cancer: dedifferentiation or stem cell maturation arrest? Environ Health Perspect 1993;101:15–26.
21. Friedmann-Morvinski D, Verma IM. Dedifferentiation and reprogramming: origins of cancer stem cells. EMBO Rep 2014;15:244–53.
22. Marjanovic ND, Hofree M, Chan JE, Canner D, Wu K, Trakala M, et al. Emergence of a high-plasticity cell state during lung cancer evolution. Cancer Cell 2020;38:229–46.
23. Snyder EL, Watanabe H, Magendantz M, Hoersch S, Chen TA, Wang DG, et al. Nkx2-1 represses a latent gastric differentiation program in lung adenocarcinoma. Mol Cell 2013;50:185–99.
24. Chesler DA, Berger MS, Quinones-Hinojosa A. The potential origin of glioblastoma initiating cells. Front Biosci 2012;4:190–205.
25. Fan X, Xiong Y, Wang Y. A reignited debate over the cell(s) of origin for glioblastoma and its clinical implications. Front Med 2019;13:531–9.
26. Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, Rahme GJ, et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. Cell 2019;178:835–49.
27. van Galen P, Hovestadt V, Wadsworth Ii MH, Hughes TK, Griffin GK, Battaglia S, et al. Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity. Cell 2019;176:1265–81.
28. Qian J, Olbrecht S, Boeckx B, Vos H, Laoui D, Etlioglu E, et al. A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. Cell Res 2020;30:745–62.
29. Lee HO, Hong Y, Etlioglu HE, Cho YB, Pomella V, Van den Bosch B, et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. Nat Genet 2020;52:594–603.
30. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell 2017;171:1611–24.
31. Zhang Y, Narayanan SP, Mannan R, Raskind G, Wang X, Vats P, et al. Single-cell analyses of renal cell cancers reveal insights into tumor microenvironment, cell of origin, and therapy response. Proc Natl Acad Sci U S A 2021;118:e2103240118.
32. Sun Y, Wu L, Zhong Y, Zhou K, Hou Y, Wang Z, et al. Single-cell landscape of the ecosystem in early-relapse hepatocellular carcinoma. Cell 2021;184:404–21.
33. Laughney AM, Hu J, Campbell NR, Bakhoum SF, Setty M, Lavallee VP, et al. Regenerative lineages and immune-mediated pruning in lung cancer metastasis. Nat Med 2020;26:259–69.
34. Tirosh I, Izar B, Prakadan SM, Wadsworth MH 2nd, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 2016;352:189–96.
35. Peng J, Sun BF, Chen CY, Zhou JY, Chen YS, Chen H, et al. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res 2019;29:725–38.
36. Dong B, Miao J, Wang Y, Luo W, Ji Z, Lai H, et al. Single-cell analysis supports a luminal-neuroendocrine transdifferentiation in human prostate cancer. Commun Biol 2020;3:778.
37. Jerby-Arnon L, Neftel C, Shore ME, Weisman HR, Mathewson ND, McBride MJ, et al. Opposing immune and genetic mechanisms shape oncogenic programs in synovial sarcoma. Nat Med 2021;27:289–300.
38. Izar B, Tirosh I, Stover EH, Wakiro I, Cuoco MS, Alter I, et al. A single-cell landscape of high-grade serous ovarian cancer. Nat Med 2020;26:1271–9.
39. Grewal JK, Tessier-Cloutier B, Jones M, Gakkhar S, Ma Y, Moore R, et al. Application of a neural network whole transcriptome-based pan-cancer method for diagnosis of primary and metastatic cancers. JAMA Netw Open 2019;2:e192597.
40. Zhao Y, Pan Z, Namburi S, Pattison A, Posner A, Balachander S, et al. CUP-AI-Dx: a tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence. EBioMedicine 2020;61:103030.
41. Hayashi H, Takiguchi Y, Minami H, Akiyoshi K, Segawa Y, Ueda H, et al. Site-specific and targeted therapy based on molecular profiling

by next-generation sequencing for cancer of unknown primary site: a nonrandomized phase 2 clinical trial. JAMA Oncol 2020;6:1931–8.
42. Kerr SE, Schnabel CA, Sullivan PS, Zhang Y, Singh V, Carey B, et al. Multisite validation study to determine performance characteristics of a 92-gene molecular cancer classifier. Clin Cancer Res 2012;18:3952–60.
43. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd International Conference on Machine Learning; 2006 Jun 25–29; Pittsburgh, PA. New York: Association for Computing Machinery; 2006. p. 161–8.
44. Tyner JW, Tognon CE, Bottomly D, Wilmot B, Kurtz SE, Savage SL, et al. Functional genomic landscape of acute myeloid leukaemia. Nature 2018;562:526–31.
45. Grande BM, Gerhard DS, Jiang A, Griner NB, Abramson JS, Alexander TB, et al. Genome-wide discovery of somatic coding and noncoding mutations in pediatric endemic and sporadic Burkitt lymphoma. Blood 2019;133:1313–24.
46. Phelan JD, Young RM, Webster DE, Roulland S, Wright GW, Kasbekar M, et al. A multiprotein supercomplex controlling oncogenic signalling in lymphoma. Nature 2018;560:387–91.
47. Keats JJ, Craig DW, Liang W, Venkata Y, Kurdoglu A, Aldrich J, et al. Interim analysis of the MMRF CoMMpass trial, a longitudinal study in multiple myeloma relating clinical outcomes to genomic and immunophenotypic profiles. Blood 2013;122:532.
48. Edwards NJ, Oberti M, Thangudu RR, Cai S, McGarvey PB, Jacob S, et al. The CPTAC data portal: a resource for cancer proteomics research. J Proteome Res 2015;14:2707–13.
49. Higashiyama H, Uemura M, Igarashi H, Kurohmaru M, Kanai-Azuma M, Kanai Y. Anatomy and development of the extrahepatic biliary system in mouse and rat: a perspective on the evolutionary loss of the gallbladder. J Anat 2018;232:134–45.
50. Schittny JC. Development of the lung. Cell Tissue Res 2017;367:427–44.
51. Javed A, Lteif A. Development of the human breast. Semin Plast Surg 2013;27:5–12.
52. Moiso E, Provero P. Cancer metabolic subtypes and their association with molecular and clinical features. Cancers 2022;14:2145.
53. Gavish A, Tyler M, Simkin D, Kovarsky D, Gonzalez Castro LN, Halder D, et al. The transcriptional hallmarks of intra-tumor heterogeneity across a thousand tumors. bioRxiv 2021:2021.12.19.473368.