communications biology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42003-025-07942-y

scMalignantFinder distinguishes malignant cells in single-cell and spatial transcriptomics by leveraging cancer signatures

Check for updates

Qiaoni Yu[1,2], Yuan-Yuan Li[1,3] & Yunqin Chen[1,2,3]

Single-cell RNA sequencing (scRNA-seq) is a powerful tool for characterizing tumor heterogeneity, yet accurately identifying malignant cells remains challenging. Here, we propose scMalignantFinder, a machine learning tool specifically designed to distinguish malignant cells from their normal counterparts using a data- and knowledge-driven strategy. To develop the tool, multiple cancer datasets were collected, and the initially annotated malignant cells were calibrated using nine carefully curated pan-cancer gene signatures, resulting in over 400,000 single-cell transcriptomes for training. The union of differentially expressed genes across datasets was taken as the features for model construction to comprehensively capture tumor transcriptional diversity. scMalignantFinder outperformed existing automated methods across two gold-standard and eleven patient-derived scRNA-seq datasets. The capability to predict malignancy probability empowers scMalignantFinder to capture dynamic characteristics during tumor progression. Furthermore, scMalignantFinder holds the potential to annotate malignant regions in tumor spatial transcriptomics. Overall, we provide an efficient tool for detecting heterogeneous malignant cell populations.

Tumors are highly heterogeneous diseases, driven by molecular disturbances at the genetic, epigenetic, and gene expression levels within tumor cells[1]. Accurately characterizing tumor heterogeneity is crucial for understanding the mechanisms of cancer development and devising effective and durable therapeutic strategies[2]. Single-cell RNA sequencing (scRNA-seq) technology has revolutionized the field of cancer biology[3] by enabling the measurement of the complete transcriptome of each cell within a tumor tissue, thus facilitating the identification of distinct cell types and states[4].

Identifying malignant cells and distinguishing them from other non-malignant cells is a crucial step in dissecting tumor heterogeneity from cancer scRNA-seq experiments[3,5,6]. This process aids in gaining a deep understanding of transcriptional reprogramming underlying the malignant transformation of cells during cancer development. Copy number variation (CNV) inference methods are commonly employed to identify malignant cells[7,8], and their reliability strongly depends on how well the observed gene expression deviations correlate with underlying copy number changes rather than other biological and technical factors[1]. Furthermore, CNV inference often requires the specification of appropriate normal reference cells, hindering the complete automation of the analysis process. Therefore, relying solely on inferred CNVs has limitations in annotating malignant cells, especially in cancers with minimal genomic structural variation or other unknown origins[5,9]. Regarding the former, a potential alternative is to identify malignant cells through smaller-scale genetic alterations, such as single nucleotide variants (SNVs). However, detecting SNVs from scRNA-seq data requires sufficient read coverage over the expressed exons, making it primarily applicable to the scRNA-seq protocols that capture full-length rather than 3' or 5' transcripts[1].

Given these limitations, recent advancements have introduced automated approaches for identifying malignant cells through supervised learning, utilizing training data with cell identities provided by the original studies[6,10,11]. However, the performance of these supervised learning methods would be greatly compromised by the lack of optimal reference datasets that accurately annotate the malignancy status of each cell[1,12,13]. Additionally, current approaches[6,10] often select genes that consistently exhibit differential expression across datasets as model features, potentially overlooking substantial inter-tumor heterogeneity. Consequently, there is

[1]Shanghai-MOST Key Laboratory of Health and Disease Genomics, Shanghai Institute for Biomedical and Pharmaceutical Technologies, Shanghai, China. [2]Shanghai Genbase Biotechnology Co., Ltd, Shanghai, China. [3]These authors jointly supervised this work: Yuan-Yuan Li, Yunqin Chen.

e-mail: liyuanyuan@sibpt.com; yunqin2016@alumni.sjtu.edu.cn

an urgent need for specialized methods that are guided by both high-quality data and well-established knowledge, incorporated with accurately labelled datasets and robust feature selection strategy to effectively identify heterogeneous malignant cells from cancer scRNA-seq data.

This study introduces scMalignantFinder, a machine learning-based automated classifier, specially designed to distinguish malignant cell from their originating normal cells, rather than from all other non-malignant cells as existing methods usually do. We systemically reviewed multiple scRNA-seq datasets with cell type annotations and calibrated the initially annotated malignant cells using nine carefully curated cancer gene signatures that exhibited consistent transcriptional patterns across diverse cancer types to construct the gold standard training set. The union set of differentially expressed genes (DEGs) between the calibrated malignant cells and normal cells across datasets was adopted to build a classification model. scMalignantFinder demonstrated superior performance compared to current automated methods on independent test sets ranging cancer cell lines, non-cancer tissues, and eleven cancer single-cell datasets covering nine cancer types. The capability of scMalignantFinder to predict malignancy probability empowers it to capture dynamic characteristics during tumor progression. Moreover, we extended the classifier to discover malignant spots in spatial transcriptomics (ST) data without retraining, achieving high concordance with pathologists' annotations across multiple cancer ST slices. These findings underscore scMalignantFinder as a versatile and generalizable tool for investigating malignant cell biology in cancer research.

Results

An overview of scMalignantFinder

We developed scMalignantFinder to distinguishing malignant cells from their originating normal cells. Our focus was on carcinomas, which originate from epithelial tissue and accounts for 80 ~ 90% of all cancer cases. To achieve this, we collected five publicly available single-cell RNA-seq (scRNA-seq) datasets with cell type annotations[14–18], encompassing four cancer types: lung, colorectal, liver, and gastric cancer. These datasets, containing expression profiles over 400,000 malignant and normal epithelial cells, served as the basis for constructing our training set (Supplementary Data 1). scMalignantFinder was designed through two major steps: (1) defining the training dataset by calibrating malignant cells using cancer hallmark-based gene signatures, and (2) selecting model features by taking the union of significantly differentially expressed genes (DEGs) across datasets (Fig. 1).

A key challenge in constructing a training set for supervised learning in the context of identifying malignant cells is the lack of a gold-standard annotation for them. To address this, we adopted a gene signature-based approach to uniformly calibrate the malignant cells across collected datasets (Fig. 1). Cancer hallmarks represent a set of common functional capabilities that human cells acquire during the transition from normal to tumorigenic states[19], making them potential common markers of malignant cells. Based on this rationale, we curated 29 gene signatures associated with cancer hallmarks from both knowledge-based and data-driven sources[20–22].

Gene set activity analysis was applied to the five scRNA-seq datasets and bulk RNA-seq datasets of 16 cancer types from The Cancer Genome Atlas (TCGA) (Supplementary Data 1). Differential activity analysis was then performed for each gene signature between malignant and normal epithelial cells, as well as tumor and adjacent normal samples. Nine out of the 29 gene signatures exhibited consistent trends across different datasets at both single-cell and bulk levels, thus representing pan-cancer transcriptional characteristics (Supplementary Fig. 1; Supplementary Data 2). Among these signatures, eight were upregulated in both malignant cells and tumor samples, representing cellular processes such as cell cycle, DNA damage, and DNA repair; while one signature, known to be enriched in

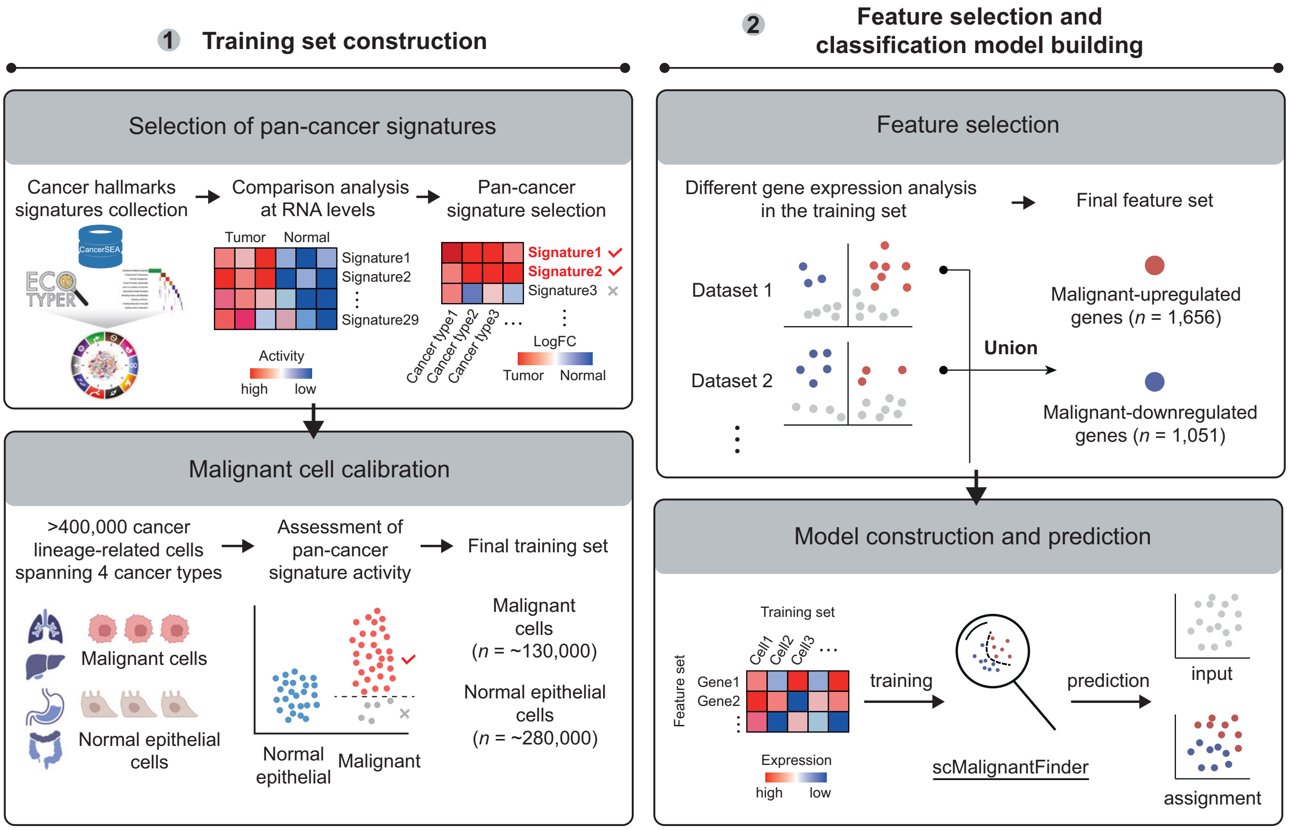
1 Training set construction
2 Feature selection and classification model building
Selection of pan-cancer signatures
Cancer hallmarks signatures collection
Comparison analysis at RNA levels
Pan-cancer signature selection
CancerSEA
ECO TYPER
Tumor
Normal
Signature1
Signature2
Signature29
Activity
high
low
Signature1 ✓
Signature2 ✓
Signature3 ✕
Cancer type1
Cancer type2
Cancer type3
LogFC
Tumor
Normal
Malignant cell calibration
>400,000 cancer lineage-related cells spanning 4 cancer types
Assessment of pan-cancer signature activity
Final training set
Malignant cells
Normal epithelial cells
Normal epithelial
Malignant
Malignant cells (n = ~130,000)
Normal epithelial cells (n = ~280,000)
Feature selection
Different gene expression analysis in the training set
Final feature set
Dataset 1
Dataset 2
Union
Malignant-upregulated genes (n = 1,656)
Malignant-downregulated genes (n = 1,051)
Model construction and prediction
Training set
Cell1
Cell2
Cell3
Feature set
Gene1
Gene2
Expression
high
low
training
scMalignantFinder
prediction
input
assignment

Fig. 1 | Workflow of scMalignantFinder. scMalignantFinder identifies malignant cells through two steps: (1) constructing a training set by calibrating malignant cells using nine curated cancer gene signatures across over 400,000 epithelial cells from four cancer types, and (2) performing feature selection by taking the union of differentially expressed genes (DEGs) across datasets, which captures both common DEGs and those unique to specific datasets. These refined steps provide the foundation for building a logistic regression model that classifies cells as malignant or normal based on single-cell RNA-seq (scRNA-seq) data.

normal specimens[21], was consistently downregulated. Using the activity of these nine cancer gene signatures in normal epithelial cells as a baseline, we filtered out 2.5% of the original malignant cells (Supplementary Fig. 2a) and obtained the final training set consisting of 416,774 cells, including 134,053 malignant cells and 282,721 normal epithelial cells.

Next, we conducted differential gene expression analysis between malignant cells and normal epithelial cells within each dataset in the training set, and identified a range of 327 to 2226 DEGs per dataset (Fig. 1). Compared to downregulated DEGs, upregulated DEGs in malignant cells tended to be more common across different cancer types (Supplementary Fig. 2b). Notably, 68.3% of upregulated DEGs and 87.4% of downregulated DEGs were found to be specific to individual datasets. Considering that tumors exhibit both shared functional characteristics[19] and high heterogeneity[23], we incorporated features that capture both shared and distinct characteristics of malignant cells. Specifically, we retained DEGs that are common across datasets, representing universal features of malignant cells, as well as dataset-specific DEGs to account for unique characteristics of individual tumors. Finally, a set of 2707 DEGs, comprising 1656 upregulated genes and 1051 downregulated genes in malignant cells (Supplementary Data 3), were used to construct the input expression matrix for the logistic regression classifier-scMalignantFinder (Fig. 1).

Performance validation by diverse scRNA-seq datasets

To evaluate the performance of scMalignantFinder, we collected multiple independent scRNA-seq datasets from various sample sources, including cancer cell lines, tumor biopsies, pre-cancerous tissues, tumor-adjacent tissues, and deceased non-cancer donor specimens. The collected test set used for validation consisted of 607,109 cells, including 260,734 malignant cells and 346,375 normal tissue cells, from 633 samples spanning 22 cancer types. We compared scMalignantFinder with four other tools: PreCanCell[6], ikarus[10], Cancer-Finder[11], and CopyKAT[8]. and evaluated the performance using seven metrics, including the area under the receiver operating characteristic curve (AUROC), accuracy, balanced accuracy (the average of sensitivity and specificity), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

We first evaluated scMalignantFinder using two expression datasets including 53,513 malignant cells from 198 cancer cell lines[24] and 104,148 normal epithelial cells from 18 tissues from non-cancer donors[25], which served as gold standard datasets due to their evident sample sources (Supplementary Data 5 and 9). When pooling all cells from each dataset together, scMalignantFinder achieved a sensitivity of 1.000 in identifying malignant cells and a specificity of 0.786 in identifying normal epithelial cells, outperforming other methods (PreCanCell: 0.996 and 0.503; ikarus: 0.834 and 0.642; Cancer-Finder: 1.000 and 0.022; CopyKAT: 0.594 and 0.397) (Fig. 2a). When evaluating individual cancer cell line, scMalignantFinder demonstrated high and consistent sensitivity in detecting malignant cells, with a minimum sensitivity of 0.98 (Fig. 2b), indicating its robustness across diverse cell lines and cancer types.

Furthermore, we applied scMalignantFinder to 11 additional scRNA-seq datasets from cancer patients[16,17,23,26–28] (see Supplementary Data 1), involving nine cancer types, three of which were included in the training set (lung, colorectal, and liver cancer), and six not included (head and neck, kidney, prostate, breast, pancreatic, and ovarian cancer). The original cell annotations provided in the respective studies were used as true labels for validation. Across all 13 datasets, scMalignantFinder achieved an average accuracy of 0.824 and an average balanced accuracy of 0.732, surpassing other methods (PreCanCell: 0.713 and 0.613; ikarus: 0.446 and 0.533; Cancer-Finder: 0.654 and 0.531; CopyKAT: 0.427 and 0.571) (Fig. 2c; Supplementary Fig. 3a). Notably, scMalignantFinder exhibited the highest accuracy in 6 of 13 datasets and the highest balanced accuracy in 7 out of 11 datasets (Fig. 2a; Supplementary Fig. 3b).

While PreCanCell and Cancer-Finder demonstrated comparable or superior sensitivity compared to scMalignantFinder, they displayed lower specificity, and higher false positive rate (Fig. 2b, c). Conversely, ikarus and CopyKAT exhibited the highest specificity but the lowest sensitivity (Fig. 2b, c). The combined high sensitivity and PPV provided by scMalignantFinder highlight its ability to accurately identify malignant cells while maintaining a low false positive rate. This is a fundamental prerequisite for gaining an in-depth understanding of tumor heterogeneity through single-cell data analysis. Overall, scMalignantFinder outperformed the other four methods in distinguishing malignant cells from their normal counterparts.

Next, we retrained Cancer-Finder and ikarus using our calibrated training set to explore the effects of training set design strategy on model performance (Fig. 2c; Supplementary Fig. 3a, b). After retraining, Cancer-Finder exhibited a noticeable improvement, with its average accuracy increasing from 0.654 to 0.828 and average balanced accuracy rising from 0.531 to 0.669. For ikarus, we also replaced its gene signatures with the DEGs identified from our training set. Retraining enhanced ikarus's performance, with its average accuracy increasing from 0.446 to 0.488 and average balanced accuracy rising from 0.533 to 0.612. These performance improvements were primarily reflected in enhanced specificity for identifying normal cells, underscoring the importance of training data calibration and rational feature selection.

To assess whether the performance of scMalignantFinder is influenced by sequencing depth, we reanalyzed its prediction on samples with different median gene counts across multiple test sets. We measured Spearman correlation between median gene count and balanced accuracy within each dataset, and did not find any significant correlation in all the datasets (Supplementary Fig. 4), to some extent, demonstrating the robustness of scMalignantFinder.

Transcriptional characteristics of misclassified cell populations

Given that the prediction by scMalignantFinder was not entirely consistent with the original label, we investigated the transcriptional characteristics of the misclassified cells. The cells in the test set were classified into four categories: (1) true malignant, labeled and correctly predicted as malignant, (2) predicted malignant, labeled as normal but wrongly predicted as malignant, (3) predicted normal, labeled as malignant but wrongly predicted as normal, and (4) true normal, labeled and correctly predicted as normal (Supplementary Data 6 and 10).

We first examined the expression profiles of the curated cancer gene signatures for these four cell categories (Supplementary Data 2). As expected, in the liver dataset[28], a substantial increase in the activity of multiple upregulated signatures was observed in predicted malignant cells compared to true normal cells; conversely, predicted normal cells exhibited an enrichment of downregulated signature, surpassing the activity of true malignant cells (Fig. 3a). This trend was consistent across several other datasets (Supplementary Fig. 5a). Statistical analysis across the test sets indicated that predicted malignant cells ranked second in activity levels among seven out of nine cancer gene signatures, closely following true malignant cells. Similarly, predicted normal cells demonstrated activity levels in seven signatures that closely resembled those of true normal cells (Fig. 3b). Furthermore, Copy Number Variation (CNV) has been closely associated with the development and progression of various cancers due to its impact on gene expression[29]. By performing single-cell CNV inference[7], we discovered that predicted malignant cells exhibited CNV profiles similar to true malignant cells, with significantly higher CNV levels compared to true normal cells (Supplementary Fig. 5b, c).

To further characterize the transcriptional features of predicted malignant cells, we conducted differential gene expression analysis between the two types of malignant cells and true normal cells. Our analysis revealed that among the identified DEGs, 20% of the genes upregulated in true malignant cells also exhibited elevated expression in predicted malignant cells. Similarly, 42% of the downregulated genes in true malignant cells showed consistent downregulation in predicted malignant cells (Fig. 3c). Pathway enrichment analysis revealed that the commonly upregulated DEGs were mainly involved in angiogenesis and metabolic reprogramming (Fig. 3d), which are typical hallmarks of cancer[19]. Furthermore, we predicted the upstream regulators of the upregulated DEGs and identified the highly ranked transcription factors (TFs), such as HIF1A[30] and TWIST1[31] (Fig. 3e).

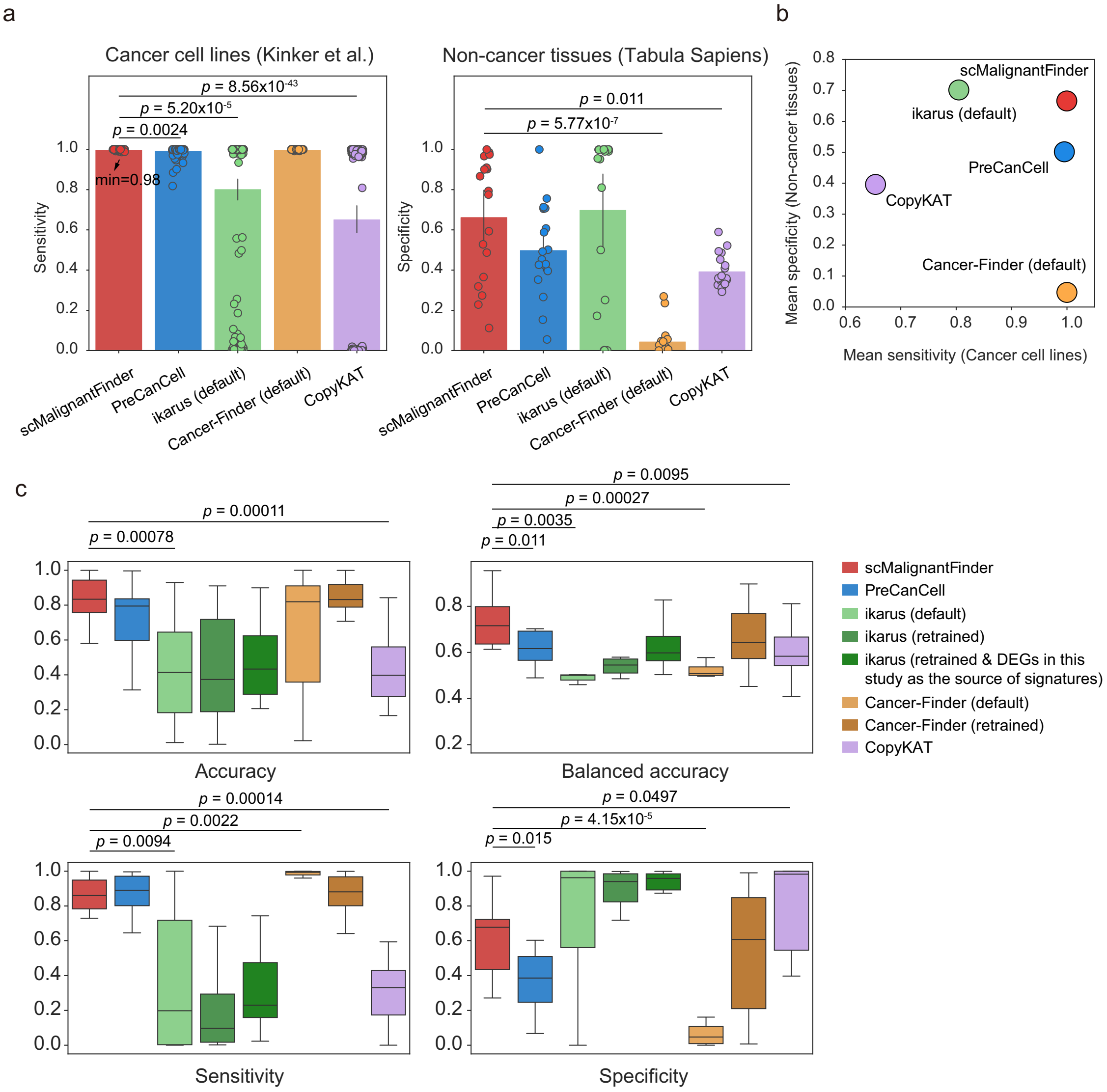

Fig. 2 | Performance evaluation of scMalignantFinder. a Barplot showing the sensitivity of each method applied to the cancer cell line scRNA-seq dataset (left) and the specificity of each method applied to the non-cancer tissue scRNA-seq dataset (right). Each point represented a cell line (left) or a tissue (right). Statistical significance was determined by a two-sided Wilcoxon rank-sum test. n = 198 cell lines (left) or 18 tissues (right); bar height represents the mean value; whiskers extend to the mean ± 95% confidence intervals. **b** Scatterplot showing the mean sensitivity (calculated from the cancer cell line scRNA-seq dataset) and mean specificity (calculated from the non-cancer tissue scRNA-seq dataset) for each method presented in **a**. **c** Boxplot showing seven metrics of each method on 13 scRNA-seq datasets (Supplementary Data 1). "Retrained" indicates the models were re-trained using the calibrated training set from this study, while "DEGs in this study as the source of signatures" refers to ikarus being retrained with gene signatures derived from the DEGs identified in this study. Statistical significance was determined by a two-sided Wilcoxon rank-sum test. n = 13 datasets; center line represents the median value; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range. Source data are provided as a Source Data file (Supplementary Data 5 and 9).

These TFs have been reported to be upregulated in various cancers and play roles in downstream cancer-promoting processes through transcriptional regulation such as altered substrate metabolism, angiogenesis, tumor invasion, and metastasis.

Overall, these results indicate that predicted malignant cells exhibit transcriptional characteristics of malignant cells. The presence of false positives (predicted malignant cells) and false negatives (predicted normal cells) are probably due to the continuously transitional states of cells between normal and malignancy at transcriptional profiles.

Application of scMalignantFinder in elucidating cell states during carcinogenesis

Since scMalignantFinder can also report the malignancy probability of each cell, we set out to analyze the malignant state during cancer progression. To this end, we applied scMalignantFinder to colorectal cancer datasets[16,17], which comprises epithelial cell expression profiles from 36 normal mucosal tissues, 23 colorectal polyps (precursors of colorectal carcinoma), and 6 tumors, and predicted the malignancy probability and classification for each epithelial cell (Supplementary Fig. 6a; Supplementary Data 11). It was shown that the percentage of predicted malignant cells was nearly 0% in normal mucosa, significantly increased in colorectal polyps, and peaked in tumors. Similarly, the malignancy probability exhibited a clear increasing trend from normal mucosa to colorectal polyps, and to tumors.

In addition, we applied scMalignantFinder to an scRNA-seq dataset representing different stages of gastric cancer progression[32] (Supplementary Fig. 6b), which included samples from three non-atrophic gastritis (NAG) biopsies as normal controls, three chronic atrophic gastritis (CAG) biopsies and six intestinal metaplasia (IM) biopsies both representing precancerous lesions, and one early gastric cancer (EGC) biopsy. The analysis revealed that the percentage of predicted malignant cells was 0% in NAG, increased

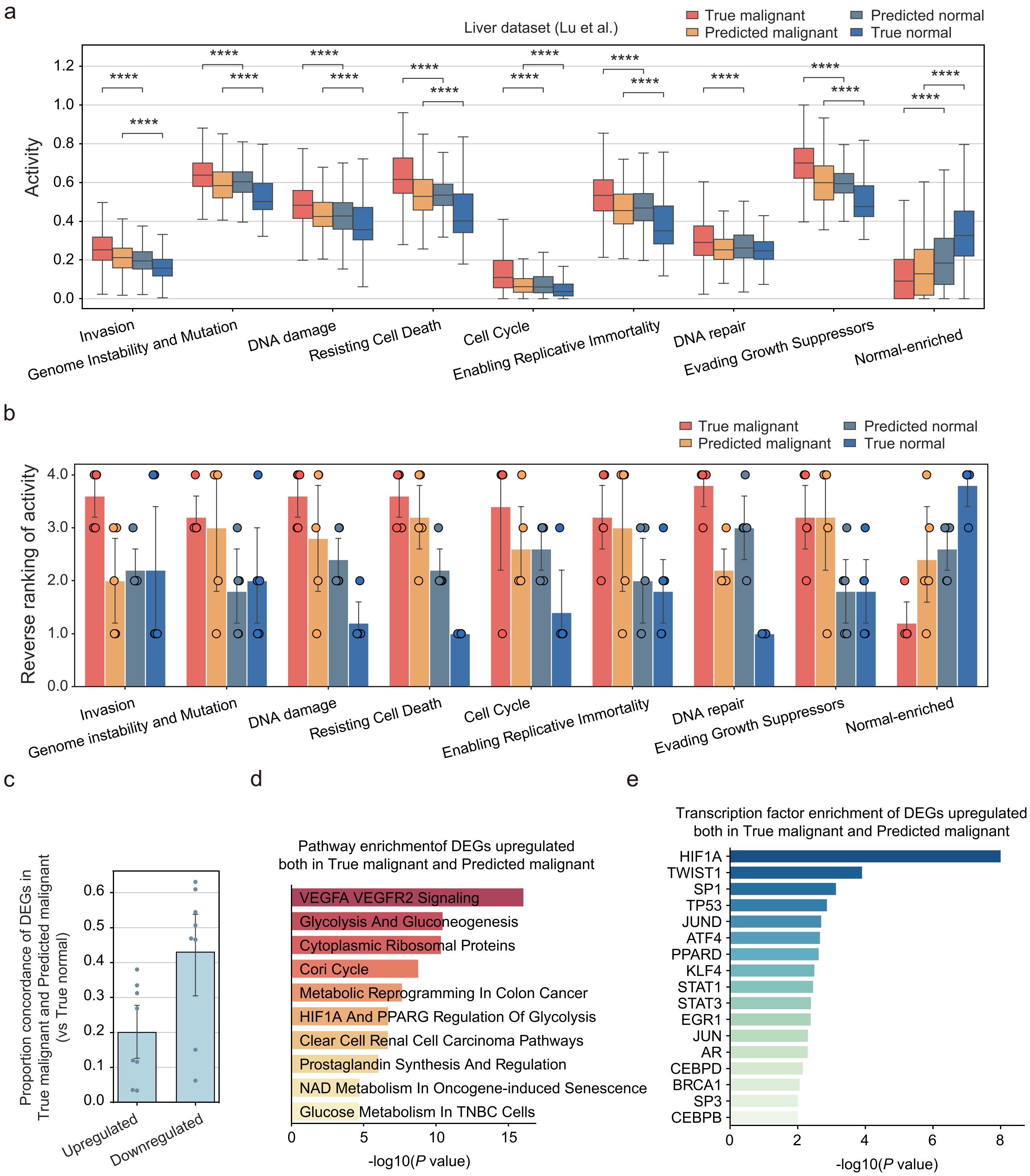

Fig. 3 | Characterization of transcriptional features in misclassified cells. **a** Boxplot showing activity of nine curated cancer gene signatures in four groups of cells (true malignant, predicted malignant, predicted normal, true normal) in the liver dataset. Statistical significance was determined by a two-sided Student's t test (****$P < 0.0001$). Center line represents the median value; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range. **b** Barplot showing reverse ranking of activity of nine curated cancer gene signatures in four groups of cells across five test sets including lung, liver, and colorectal cancer. Reverse ranking is based on the mean activity among the four groups. For a given dataset and a group of cells, if the mean of their activities is the highest among the four groups, their reverse ranking is 4; if the mean activity is the second highest, the reverse ranking is 3, and so on. n = 5 datasets; bar height represents the mean value; whiskers extend to the mean ± 95% confidence intervals. **c** Boxplot showing proportion concordance of differentially expressed genes (DEGs) in true malignant and predicted malignant cells (vs true normal cells). **d**, **e** Pathway (**d**) and transcription factor (**e**) enrichment of DEGs upregulated in both true malignant and predicted malignant cells. Source data are provided as a Source Data file (Supplementary Data 6, 10, and 11).

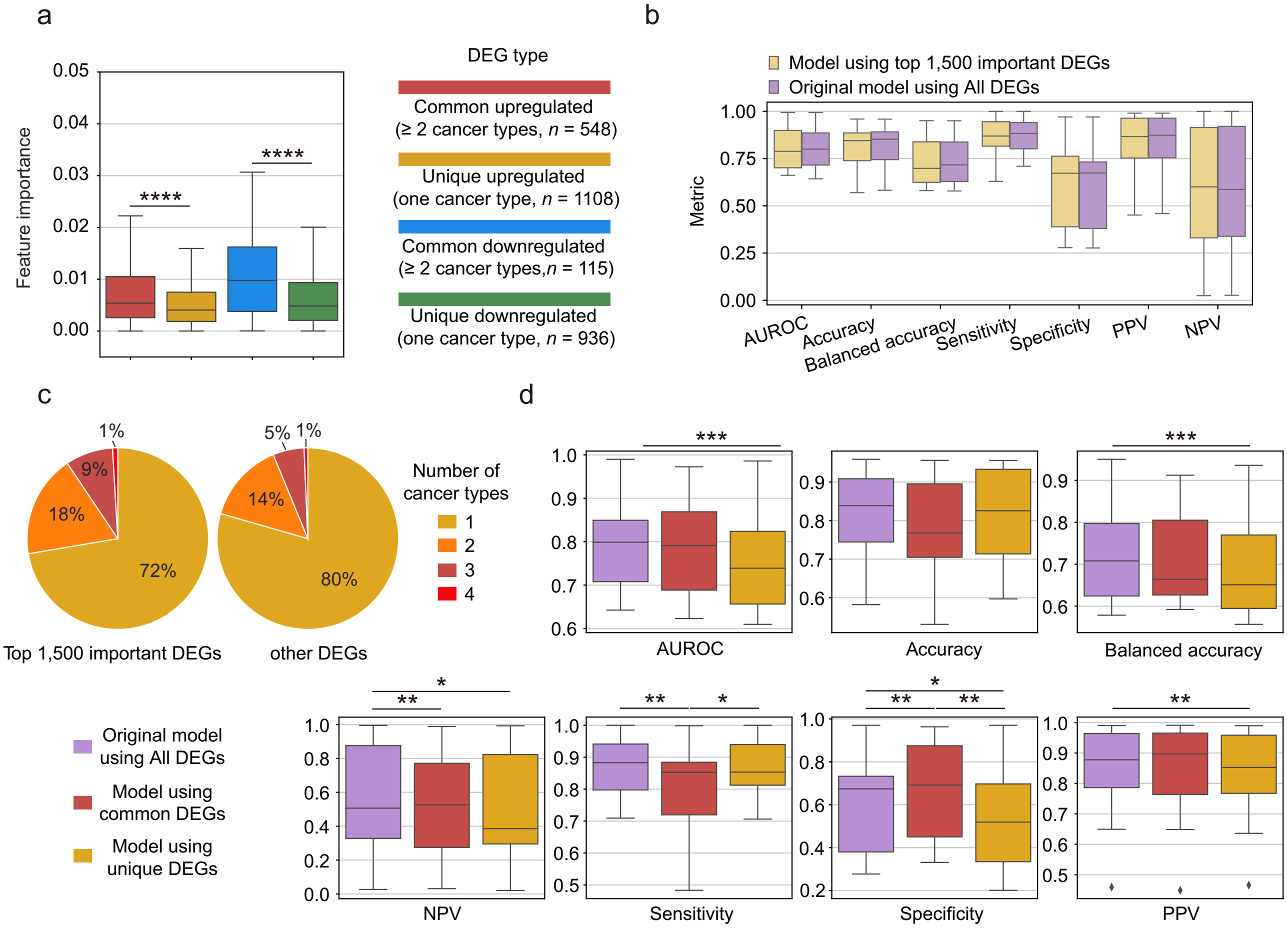

Fig. 4 | Contribution of each DEG to model training and prediction. a Boxplot showing feature importance of two types of DEGs (common and unique). Unique DEGs are exclusive to a specific cancer type, while common DEGs are found in at least two cancer types (Supplementary Data 3). The importance of each DEG is determined by the absolute value of the model coefficient (Supplementary Data 4). Statistical significance was determined by a two-sided Student's t test (****: $P < 0.0001$). Center line represents the median value; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range. **b** Comparison of the original model and a model using the top 1500 important DEGs on seven prediction metrics. n = 13 datasets; center line represents the median value; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range. **c** Pie chart illustrating the statistics of how many cancer types in the training set the top 1500 important DEGs (left) and the remaining DEGs (right) were found in. **d** Comparison among the original model and models using either common DEGs or unique DEGs as features. Statistical significance was determined by a two-sided Wilcoxon rank-sum test (*: $P < 0.05$, **: $P < 0.01$, ***: $P < 0.001$). n = 13 datasets; center line represents the median value; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range. Source data are provided as a Source Data file (Supplementary Data 7 and 12).

in CAG and IM, and reached its highest level in EGC. Notably, both the percentage of malignant cells and the malignancy probability in IM were significantly higher than in CAG, consistent with the established fact that IM carries a higher risk of gastric cancer compared to CAG[33].

The above results suggested that scMalignantFinder can shed lights on the dynamics underlying the malignant transformation of epithelial cells.

The roles of different types of DEGs in identifying malignant cells

We conducted feature importance assessment to quantify the contribution of each feature used in scMalignantFinder (Supplementary Fig. 7a; Supplementary Data 4, 7, and 12). The 2707 DEGs were categorized into common DEGs and unique DEGs based on whether they were differentially expressed across at least two cancer types in the training set (Supplementary Data 3). It was found that the commonly upregulated or downregulated DEGs in malignant cells exhibited significantly higher feature importance than the unique DEGs (Fig. 4a; Supplementary Fig. 7b), supporting the notion that there are conserved mechanisms underlying carcinogenesis across cancer types[19].

Subsequently, we investigated whether the classification performance of the original model could be replicated using a subset of DEGs with top-ranked importance. As expected, the classification performance improved as the number of important genes increased (Supplementary Fig. 7c); A model with approximately 1500 input genes achieved performance comparable to the original 2707-feature model across all metrics (Fig. 4b; Supplementary Fig. 7c, d). These top 1500 DEGs included a larger proportion of common DEGs (28%) compared to the remaining DEGs (20%) (Fig. 4c).

To further explore the role of common and unique DEGs in identifying malignant cells, we constructed separate models using each type of DEGs. The model based on the common DEGs performed comparably to the original model, showing similar AUROC and balanced accuracy (Fig. 4d). However, the model based on the unique DEGs exhibited significantly poorer performance (Fig. 4d). These findings suggest that the common transcriptional signatures among malignant cells of different cancer types are more crucial for the performance of scMalignantFinder compared to those that are specific to individual cancer types.

Interestingly, models constructed solely using common DEGs demonstrated better specificity but sacrifice some sensitivity, while models constructed using cancer-specific DEGs exhibited better sensitivity (Fig. 4d). Unlike previous methods that exclusively retain common DEGs[6,10], scMalignantFinder integrates both types of DEGs, resulting in a more comprehensive discriminatory performance, as evidenced by its simultaneous reduction in both false positive and false negative rates.

We also explored the association between upregulated DEGs and clinical features. Following a previous study[23], we computed the activity of DEGs in bulk TCGA samples and discovered their associations with worse prognosis across multiple cancer types (Supplementary Fig. 7e), suggesting that the transcriptional signatures upregulated in malignant cells may involve in cancer progression. Furthermore, we analyzed the overlap of DEGs with 704 targets of approved drugs with known mechanism in the PHAROS database[34]. Compared to unique DEGs (OR = 1.0), common DEGs (OR = 1.9) and the DEGs that contributed most to the model predictions (OR = 1.5) exhibited significant enrichment among approved drug targets (Supplementary Fig. 7e).

Application of scMalignantFinder in tumor spatial transcriptomics

The presence of spatial heterogeneity within tumors, resulting from clonal expansion and the intricate interplay between cancer cells and the surrounding microenvironment, underscores the importance of analyzing the transcriptome in a spatial context to understand key oncological processes such as tumor progression and metastasis[35]. In recent years, the rapidly advancing technology of spatial transcriptomics (ST) sequencing has emerged as a powerful tool for unraveling the complex spatial architecture of the tumor microenvironment[36]. Identifying malignant regions within this context is crucial yet challenging, as ST sequencing techniques capture barcode spots with diameters ranging from 55–100 μm, potentially measuring a mixture of signals from multiple cells belonging to different lineages[37].

We explored the potential of scMalignantFinder to identify malignant spots in ST data (Fig. 5a). First, the pretrained model was used to calculate two key features for each spot in the ST dataset: malignancy probability and malignant signature activity, the latter representing the overall expression of upregulated DEGs (Supplementary Data 3). Additionally, an image score was derived for each spot based on the corresponding histological image. These three features - malignancy probability, malignant signature activity, and image score - formed a feature matrix. hierarchical clustering was then conducted on the feature matrix to group the spots into three clusters, considering that carcinomas are typically composed of malignant, normal epithelial, and non-epithelial regions. To determine the assignment of a region cluster, we employed a rank aggregation method. First, we calculated the average score for each feature within each cluster. Next, we ranked the clusters for each feature based on these average scores. The ranks of the three features were then averaged for each cluster. The cluster with the top average rank was preliminarily identified as the malignant region, while the remaining two clusters were provisionally designated as normal regions. Finally, spatial neighborhood relationships were incorporated to refine these classifications, yielding the final region predictions (Fig. 5a).

We applied scMalignantFinder to predict malignant regions in eight tumor ST datasets[38–41], encompassing five cancer types: breast cancer, oral squamous cell carcinoma, renal cell carcinoma, prostate cancer, and squamous cell carcinoma. In these datasets, the malignant regions had been annotated by pathologists in prior studies[11,37,41] (Supplementary Data 8 and 13). Our findings revealed that the two malignant features predicted by scMalignantFinder - malignancy probability and signature activity - were highly enriched in the annotated malignant spots in seven out of the eight ST slices (Fig. 5b, c; Supplementary Fig. 8a, b). Furthermore, the predicted results showed strong concordance with expert annotations, achieving an average accuracy of 0.783 and an average balanced accuracy of 0.800 (Fig. 5b, e; Supplementary Fig. 8). Compared to a prior logistic regression model relying solely on malignancy probability to identify malignant cells, the current approach, which integrates multi-feature clustering and incorporates spatial neighborhood relationships, significantly improved the performance (Supplementary Fig. 9a; Supplementary Data 14).

Next, we benchmarked scMalignantFinder against two recently developed methods for tumor ST analysis. The first one is Cancer-Finder[11], specifically designed to identify malignant regions in ST data (Supplementary Data 14). scMalignantFinder demonstrated comparable performance across the eight ST slices, with average accuracy and balanced accuracy exceeding Cancer-Finder by 0.067 and 0.062, respectively (Supplementary Fig. 9b, c). The second method, Cottrazm[42], is known for its capability to delineate tumor boundaries. When applied to three ST slices (Supplementary Fig. 9d), 87% ~ 99% of the Cottrazm inferred tumor boundary spots were adjacent to a mix of scMalignantFinder predicted malignant and normal spots (Supplementary Fig. 9e, f), aligning with the definition of tumor boundaries connecting malignant and normal regions.

To further validate the reliability of scMalignantFinder identified malignant regions, we performed multidimensional functional analyses, beginning with CNV profiling (Supplementary Data 8 and 15). Malignant regions are expected to exhibit elevated CNV levels. By using inferCNV[7] to calculate the CNV scores for each spot in the ST datasets, it was shown that spots classified as malignant exhibited significantly higher CNV scores than those as normal (Fig. 5f, g). For instance, in a breast cancer ST slice, the predicted malignant regions exhibited gains in chromosomes 1q and 8q, as well as a loss in chromosome 1p, consistent with the notion that breast cancer are commonly associated with these CNV alterations[8,43] (Supplementary Fig. 10a). Similarly, deconvolution analysis of cell type composition[37] revealed a significant enrichment of malignant cells in regions identified as malignant, whereas stromal and immune cells were predominantly localized in areas classified as normal (Supplementary Fig. 10b, c; Supplementary Data 15). Furthermore, we assessed tumor signature activity in ST slices of breast cancer and renal cell carcinoma using previously published signature genes for these cancer types[42], and found that the tumor signature activity was notably higher in the predicted malignant regions compared to the normal regions (Fig. 5h; Supplementary Fig. 10d, e; Supplementary Data 15).

Overall, these results demonstrate the potential of scMalignantFinder in annotating spatial transcriptomic data. Further improvements can be achieved by including a wider range of cancer types and utilizing more tailored training data.

Discussion

Over the past decade, scRNA-seq technology has been widely applied in oncology research[3,23,24,44]. However, the lack of reliable methods for identifying malignant cells has been a major obstacle to accurately studying tumor heterogeneity. Here, we have developed scMalignantFinder to annotate malignant cells effectively. Built upon meticulously calibrated training datasets and robust feature selection strategies, scMalignantFinder outperforms existing automated methods, including ikarus, PreCanCell, Cancer-Finder, and CopyKAT, across multiple datasets from cancer cell lines, non-cancer tissues, and various patient-derived cancer types. Additionally, scMalignantFinder shows promising potential in identifying malignant regions in ST data and tracking carcinogenic processes, aiding researchers in characterizing the spatiotemporal dynamics of tumor evolution[45].

Current automated tools[6,10,11] such as ikarus, PreCanCell, and Cancer-Finder were designed to classify malignant cells and all other cells in the tumor microenvironment. In contrast, scMalignantFinder specifically distinguishes malignant cells from their originating normal counterparts within the cancer lineage. This strategy aligns with the logic employed in cell subtype or sublineage analysis within hierarchical cell classification approaches, which have recently proven effective for cell type annotation in the tumor microenvironment[5,37,46].

scMalignantFinder has significantly advanced both the scale and quality of training data for malignant cell identification, guided by well-established knowledge. It was built on over 400,000 single-cell transcriptomes, representing the largest reference dataset for this purpose to date. The initially annotated malignant cells, derived from various methodologies in the original studies, were meticulously refined using curated cancer gene signatures, resulting in a cancer signature-calibrated training set where the malignant cells exhibit pan-cancer characteristics. As expected, scMalignantFinder assigned higher malignancy probability to the calibrated malignant cells in the training set compared to the 2.5% of cells that were excluded (Supplementary Fig. 2a). Consistently, the excluded cells showed a loss of either upregulated or downregulated cancer signatures. Retraining

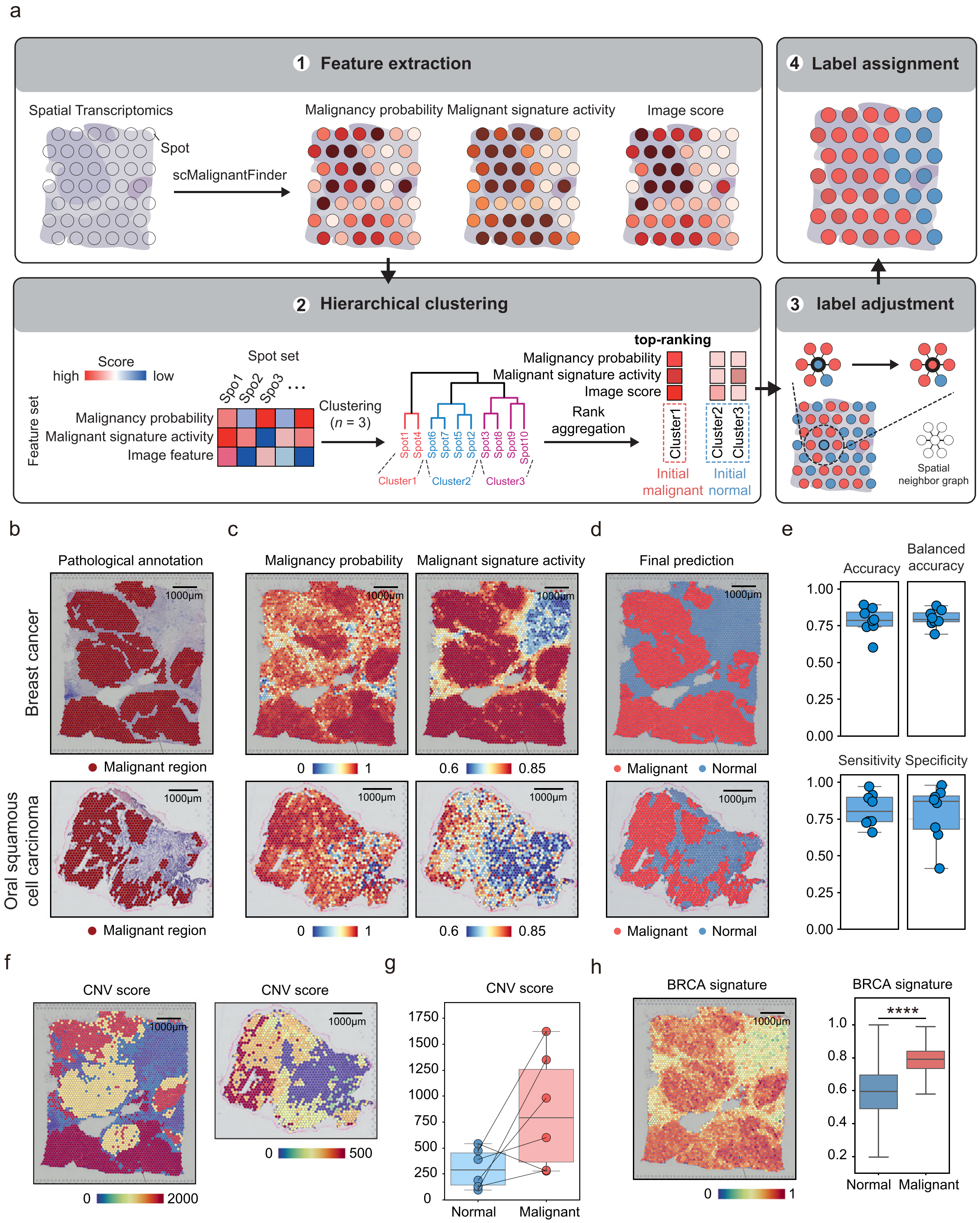

Fig. 5 | Workflow for identifying malignant regions in tumor spatial transcriptomics (ST) data. **a** The pretrained scMalignantFinder model calculates the malignant probability and malignant signature activity for each spot. Additionally, image scores derived from histological images are included as the third feature. The three features (malignant probability, malignant signature activity, and image score) are combined into a feature matrix for hierarchical clustering, grouping spots into three clusters. A rank aggregation method was employed to rank clusters based on their average ranks across three features, enabling their preliminary classification as malignant or normal. Spatial neighborhood relationships are then incorporated to refine region annotations, resulting in the final classification of malignant and normal spots. **b–d** Annotations of ST spots by different methods, including annotation by pathologists (**b**), determination of malignancy probability and signature activity (**c**), and final classification (**d**). **e** Boxplot showing four metrics of scMalignantFinder on eight ST datasets (Supplementary Data 1). n = 8 datasets; center line represents the median value; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range. **f** Spatial visualization of copy number variation (CNV) scores in ST slices of breast cancer (left) and oral squamous cell carcinoma (right). **g** Boxplot comparison of CNV scores between predicted malignant and normal regions in eight ST datasets. n = 6 datasets; center line represents the median value; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range. **h** Tumor signature activity (left) and boxplot comparison between predicted malignant and normal regions (right) in an ST slice of breast cancer. Statistical significance was determined by a two-sided Student's t test (****: $P < 0.0001$). Center line represents the median value; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range. Source data are provided as a Source Data file (Supplementary Data 8, 13–15).

existing models, such as Cancer-Finder and ikarus, on this well-calibrated dataset led to modest performance improvements (Fig. 2c; Supplementary Fig. 3a), highlighting the importance of large-scale and high-quality training data in enhancing the accuracy of malignant cell identification.

Based on the cancer signature-calibrated malignant cells, we adopted the DEGs between malignant cells and normal epithelial cells as features for our model. Due to the extensive molecular heterogeneity among tumors, we utilized the union of the DEG set in each dataset, differing from previous methods that used intersections[6,10]. This difference at least partly accounts for the lower rates of false positives and false negatives observed in our model. Attempts to directly adopt curated cancer gene signatures as model features were less effective (Fig. 2c; Supplementary Fig. 3a), probably because DEGs derived from malignant cells are more closely associated with malignancy at the transcriptional level. Notably, 75% of the identified DEGs are not encompassed in the curated cancer gene signatures, which hold the promise to offer additional insights into carcinogenesis. We found that the overexpressed genes across different cancer types, rather than specific to a particular type, primarily contributed to the model and were significantly enriched among approved drug targets, supporting the notion that pan-cancer hallmarks may be promising therapeutic targets due to their association with fundamental principles of carcinogenesis[19].

Tracking the malignant transformation of cells and unraveling the biological process underlying carcinogenesis are crucial for identifying molecular targets at key stages and enabling early tumor detection and precise diagnosis[47]. In this context, we displayed the potential of scMalignantFinder to investigate cellular and transitional dynamics during tumor progression by distinguishing malignant cells from their normal epithelial counterparts. Moving forward, we aim to further our initial findings in gastric cancer and colorectal cancer by delving into the molecular mechanisms underlying these intermediate states.

Within the recent advancements in ST technology in cancer biology[48], we have adapted the pre-trained model to identify malignant regions, demonstrating excellent scalability in this field. Through a multi-feature clustering framework, incorporating malignant features, image-based characteristics, and spatial neighborhood networks, scMalignantFinder demonstrated notable improvements in accurately identifying malignant regions. Future enhancements could focus on the continuous accumulation of expert-curated ST data and the incorporation of additional spatial features, such as the spatial distribution probability of malignant cells[37], to enhance the identification of malignant regions.

Despite these advantages, scMalignantFinder still has limitations and room for improvement. It is designed to identify malignant cells from cancerous lineages, relying on the prerequisite identification of basic cell types. To address this, we have integrated scATOMIC[5], a recently developed automated cell type annotator for pan-cancer analysis, into the current version. This integration enables users to rapidly identify malignant cells, normal cells, and other cell types within the tumor microenvironment using scRNA-seq data. Additionally, the capabilities of scMalignantFinder require further testing, particularly for cancers with unknown or controversial origins, such as synovial sarcoma[49] and certain brain tumors[50].

In summary, with a machine learning framework driven by both data and knowledge, scMalignantFinder achieves a significant advancement in malignant cell identification, offering a scalable and versatile tool for both single-cell and spatial transcriptomic analyses.

Methods

scMalignantFinder design

Data collection and processing. We collected five single-cell RNA-seq datasets from various cancer types (lung, colorectal, gastric, and liver cancers)[14–18] as detailed in Supplementary Data 1. Cell type annotations were obtained from the original studies, and we retained gene expression matrices for malignant and normal epithelial cells. Additionally, bulk RNA-seq data for 18 cancer types were sourced from TCGA, with 16 datasets used after excluding cancer types with fewer than five samples (Supplementary Data 1).

Selection of cancer gene signatures. We compiled 29 gene signatures related to cancer hallmarks from the following sources:

(1) 14 cancer functional states (Stemness, Invasion, Metastasis, Proliferation, EMT, Angiogenesis, Apoptosis, Cell cycle, Differentiation, DNA damage, DNA repair, Hypoxia, Inflammation, and Quiescence) from the cancerSEA database[20].

(2) 10 cancer hallmarks (Sustaining Proliferative Signaling, Evading Growth Suppressors, Avoiding Immune Destruction, Enabling Replicative Immortality, Tumour-Promoting Inflammation, Activating Invasion and Metastasis, Inducing Angiogenesis, Genome Instability and Mutation, Resisting Cell Death, Deregulating Cellular Energetics) from manually curated pathway gene sets[22].

(3) 5 epithelial cell states (Basal-like, Normal-enriched, Pro-angiogenic, Pro-inflammatory, Metabolic) from the Ecotyper framework[21].

Each cell in the scRNA-seq datasets and each sample in the bulk RNA-seq datasets were scored using these gene signatures. For scRNA-seq data, normalization was performed using Scanpy's "pp.normalize_total" function (version 1.9.3) (https://scanpy.readthedocs.io/en/stable/), followed by scoring with the AUCell function in the pySCENIC package (0.12.1) (https://github.com/aertslab/pySCENIC). For bulk RNA-seq data, the ssGSEA function from the GSVA R package (1.46.0) was used.

Log2 fold changes in gene activities between malignant and normal cells were calculated, and significance was assessed using the Wilcoxon rank-sum test. Gene signatures with a P-value $\le 1\times10^{-10}$ in scRNA-seq and $\le 1\times10^{-5}$ in bulk RNA-seq were considered significant.

Upregulated and downregulated gene signatures were filtered based on the following criteria, resulting in nine consistent cancer gene signatures (eight upregulated and one downregulated) identified in both datasets (Supplementary Data 2):

(1) Gene signatures that were significantly upregulated in ≥ 4 scRNA-seq datasets and significantly downregulated in ≤ 1 scRNA-seq dataset were considered upregulated in malignant cells;

(2) Gene signatures that were significantly downregulated in ≥4 scRNA-seq datasets and significantly upregulated in ≤ 1 scRNA-seq dataset were considered downregulated in malignant cells;

(3) Gene signatures that were significantly upregulated in ≥12 bulk RNA-seq datasets and significantly downregulated in ≤ 2 bulk RNA-seq datasets (or significantly upregulated in ≥6 bulk RNA-seq datasets and significantly downregulated in ≤1 bulk RNA-seq dataset) were considered upregulated in tumor samples;

(4) Gene signatures that were significantly downregulated in ≥12 bulk RNA-seq datasets and significantly upregulated in ≤ 2 bulk RNA-seq datasets (or significantly downregulated in ≥6 bulk RNA-seq datasets and significantly upregulated in ≤ 1 bulk RNA-seq dataset) were considered downregulated in tumor samples.

Training set calibration with curated cancer gene signatures. We used a gene signature scoring-based approach to filter originally labeled malignant cells. Cells were retained if they had higher upregulated signature activity or lower downregulated signature activity compared to normal cells:

(1) malignant cells with upregulated signature activity higher than the average activity of normal cells.

(2) malignant cells with downregulated signature activity lower than the average activity of normal cells.

Raw gene count matrices were filtered to exclude cells with fewer than 200 or more than 8000 gene counts or with more than 50% mitochondrial reads. This resulted in 416,774 cells (134,053 malignant and 282,721 normal epithelial cells) in the training set.

Feature selection and logistic model training. Differentially expressed genes (DEGs) were identified using Seurat's FindAllMarkers function (v4.3.0)[51]. Genes with a $|\text{Log2FC}| \ge 0.5$, P-value < 0.01, and an expression proportion (pct.1) ≥ 0.2 were selected as DEGs. DEGs were determined separately for each dataset, and the union of upregulated and downregulated genes across datasets was used as features, resulting in 2707 DEGs (1656 upregulated and 1051 DEGs as features). A logistic regression model was constructed using these DEGs as input features and trained with Python's 'sklearn.linear_model.LogisticRegression' (v1.2.2). The expression profiles from the calibrated training dataset were used for model training.

Performance evaluation of scMalignantFinder

We validated the performance of scMalignantFinder using test sets from multiple independent datasets[16,17,23–28] (Supplementary Data 1). The dataset including cells originating only from cancer cell lines was used as the positive gold standard test set, while the dataset including cells from non-cancer donor tissues served as the negative gold standard test set. The cell type annotations for the datasets were sourced from the original studies. (Supplementary Data 1). Following the same normalization method as the training set, all datasets were used as inputs for scMalignantFinder. The predictive performance was evaluated using seven metrics: area under receiver operating characteristic (AUROC), accuracy, balanced accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), calculated as follows:

$$accuracy_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{1}$$

$$sensitivity_i = \frac{TP_i}{TP_i + FN_i} \tag{2}$$

$$specificity_i = \frac{TN_i}{TN_i + FP_i} \tag{3}$$

$$balanced\ accuracy_i = \frac{sensitivity_i + specificity_i}{2} \tag{4}$$

$$PPV_i = \frac{TP_i}{TP_i + FP_i} \tag{5}$$

$$NPV_i = \frac{TN_i}{TN_i + FN_i} \tag{6}$$

where TP_i, TN_i, FP_i, FN_i represent the true positives, true negatives, false positives, and false negatives on dataset i.

We compared scMalignantFinder with four automated tools: ikarus[10] (0.0.3) (https://github.com/BIMSBbioinfo/ikarus), PreCanCell[6] (1.1.0) (https://github.com/WangX-Lab/PreCanCell), Cancer-Finder[11] (https://github.com/Patchouli-M/SequencingCancerFinder), and CopyKAT[8] (1.1.0) (https://github.com/navinlabcode/copykat). Each tool was used with its respective preprocessing methods and default parameters, except for ikarus, where 'adapt_signatures' was set to True, and CopyKAT, where 'ngene.chr' was set to 1.

We also retrained ikarus and Cancer-Finder using the training set and DEGs constructed in this study, as detailed under the sections "Training set calibration" and "Feature selection and logistic model training." For ikarus, its gene signatures were replaced with DEGs identified from the calibrated training dataset. The retrained models were subsequently assessed for their predictive performance using the same independent test sets and evaluation metrics.

Characterization of "misclassified" cells in the test sets

Cells were categorized into four groups based on initial annotations and scMalignantFinder predictions: (1) true malignant, labeled as malignant and correctly predicted as malignant, (2) predicted malignant, predicted as malignant but labeled as normal, (3) predicted normal, predicted as normal but labeled as malignant, and (4) true normal, labeled as normal and correctly predicted as normal. Cells whose prediction did not match the original labels were considered "misclassified". These cells were characterized using functional annotation and CNV inference. Functional annotation of selected genes was analyzed using Enrichr (https://maayanlab.cloud/Enrichr/)[52] against the WikiPathways and TRRUST datasets. Pathways and transcription factors displaying a P-value < 0.05 were deemed significantly enriched. The CNV profiles were inferred using the 'infercnpy' package (0.4.2) available at https://github.com/icbi-lab/infercnvpy. During the inference process, we utilized the following parameters: a window_size of 250 and excluded two chromosomes (chrX and chrY). The normal reference cells were defined as cells annotated as normal based on the original study. The calculation of the CNV score for each cell was performed using the 'infercnvpy.tl.cnv_score' function.

Application of scMalignantFinder to single-cell data spanning multiple stages of carcinogenesis

We analyzed scRNA-seq datasets representing different stages of carcinogenesis in colorectal[16,17] and gastric cancer[32]. For colorectal cancer, epithelial cell profiles from normal mucosal tissues, colorectal polyps, and tumors were processed. For gastric cancer, samples included non-atrophic gastritis (NAG), chronic atrophic gastritis (CAG), intestinal metaplasia (IM), and early gastric cancer (EGC). Using scMalignantFinder, malignancy probability and classification were predicted for each cell. Percentages of predicted malignant cells and malignant probabilities were computed and compared across the progression stages of both cancers.

Feature importance analysis

The 'coef_' attribute of the scMalignantFinder model represents the coefficients assigned to each feature within the model. These coefficients serve as indicators of feature importance and are determined by their absolute values. Subsequently, all features are sorted in descending order based on their importance, and the highest-ranked features are selected to construct a classification model. The significance of each performance metric, when compared to the original model, is evaluated using a paired-sample t-test. Furthermore, the features of the model were categorized as unique or common based on their presence across cancer types (Supplementary Data 3). Unique DEGs refer to genes that are present in only one specific cancer type, while common DEGs are genes that are found in two or more cancer types.

Association of DEGs with clinical features and drug targets

To analyze the association between DEG gene sets and overall survival, we followed the methodology of a previous study[23]. In this approach, the calculated effect size was defined as the hazard ratio, computed through Cox regression, with the P value obtained via an overall likelihood test[23]. To test the association between DEG gene sets and drug targets, we collected 704 targets with approved drugs from the PHAROS database (labeled as Tclin)[34] and computed the odds ratio following the methods of a recent study[53].

scMalignantFinder workflow for identifying malignant regions

For the ST data obtained from previous studies[38–41] (Supplementary Data 1), we applied scMalignantFinder to identify malignant regions through a multi-step process that integrates malignancy probability, gene signature activity and image-based features. Each spot in the ST data was assigned two attributes: malignancy probability, calculated using the logistic regression classifier trained on single-cell transcriptomics data, and malignant signature activity, which quantifies the expression of malignant-upregulated DEGs using the AUCell function from the pySCENIC package. For datasets with paired histological images, such as hematoxylin and eosin (H&E) staining or immunofluorescence, an additional image score was computed using 'calculate_image_features' function in Squidpy[54] (1.3.1) (https://squidpy.readthedocs.io/en/stable/index.html). These three features—malignancy probability, malignant signature activity, and image score—were combined into a feature matrix.

Next, hierarchical clustering was performed on this feature matrix using the Ward method to classify spots into three clusters, considering that solid tumors typically consist of malignant regions, normal epithelial regions, and non-epithelial regions. To determine the identity of each cluster, a rank-based vote aggregation method was used. The average score for each feature was first calculated within each cluster. Clusters were then ranked for each feature based on these average scores. The ranks across the three features were averaged for each cluster. The cluster with the top average rank was preliminarily identified as the malignant region, while the other two clusters were provisionally designated as normal regions. If two or more clusters shared the highest average rank, the cluster with the highest malignancy probability was assigned as the malignant region.

To refine the preliminary classifications, a spatial neighbor graph was constructed using Squidpy's 'spatial_neighbors' function. This graph enabled the reassignment of spot classifications based on spatial neighborhood relationships. Specifically, if more than half of a spot's neighbors within the closest distance belonged to a particular cluster, the spot was reassigned to that cluster. If not, the original classification was retained. The images with pathological annotations were extracted from previous studies[11,37,41].

Benchmarking against tumor spatial transcriptomics analysis methods

We benchmarked scMalignantFinder against two recently developed methods for tumor ST analysis. The first method, Cancer-Finder[11], utilizes a domain generalization-based deep learning approach specifically designed for identifying malignant regions in ST data. We employed the pretrained Cancer-Finder model along with its associated checkpoints optimized for ST datasets (https://github.com/Patchouli-M/SequencingCancerFinder). Both scMalignantFinder and Cancer-Finder were applied to the eight ST datasets, and performance metrics, including accuracy and balanced accuracy, were calculated for comparison. The second method, Cottrazm[42] (0.1.1) (https://github.com/Yelab2020/Cottrazm), was used to delineate tumor boundaries in three ST slices. Cottrazm classifies spots into malignant, normal, or boundary regions. To evaluate the consistency between scMalignantFinder and Cottrazm, we compared the predictions of malignant and normal spots between the two methods. Additionally, we analyzed the spatial relationships between tumor boundary spots identified by Cottrazm and neighboring malignant or normal spots predicted by scMalignantFinder, assessing proximity within one or two spot-size distances.

Functional analysis of malignant regions

To validate the identified malignant regions, we performed functional analyses, including CNV profiling, cell type deconvolution, and tumor signature activity assessment. CNV scores for each spot in the ST datasets were calculated using the STCNV and STCNVScore function integrated within the Cottrazm pipeline, which calls inferCNV (https://github.com/broadinstitute/inferCNV). Cell type deconvolution was conducted using SpaCET[37] (1.2.0) (https://github.com/data2intelligence/SpaCET) to estimate the proportion of malignant cells, stromal cells, and immune cells within the predicted malignant and normal regions. Additionally, tumor signature activity was assessed using gene sets previously published for breast cancer and renal cell carcinoma[42]. The AUCell function from the pySCENIC package was used to calculate signature activity for each spot, quantifying the expression of tumor-associated genes in malignant and normal regions.

Statistics and reproducibility

All statistical analyses were performed using Python (3.10) with appropriate libraries for data processing, statistical testing, and visualization. Statistical significance was assessed using two-sided Wilcoxon rank-sum tests, two-side Student's t tests, or Mann-Whitney U tests where applicable, as indicated in figure legends. For survival analyses, Cox regression and likelihood ratio tests were employed to determine statistical significance. Significance levels are denoted as *P < 0.05, **P < 0.01, ***P < 0.001, and ****P < 0.0001.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All the bulk RNA-seq, single-cell RNA-seq (scRNA-seq), and spatial transcriptomics (ST) datasets used in this study are publicly available (Supplementary Data 1). The TCGA bulk RNA-seq datasets, covering sixteen cancer types, were collected from UCSC Xena Browser [https://xenabrowser.net/datapages/?dataset=tcga_RSEM_Hugo_norm_count&host=https%3A%2F%2Ftoil.xenahubs.net&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443]. The scRNA-seq datasets retrieved from the NCBI Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) can be downloaded using the following accession numbers: GSE132465[17], GSE144735[17], GSE151530[18], GSE157220[24], GSE134520[32]. Additionally, scRNA-seq datasets covering six cancer types (pancreas, kidney, prostate, breast, ovarian, and head and neck cancer)[23] were collected from the website https://www.weizmann.ac.il/sites/3CA/. The remaining scRNA-seq datasets were obtained through the links provided in CZ CELLxGENE Data Explorer[55]: Salcher et al.[14] [https://cellxgene.cziscience.com/collections/edb893ee-4066-4128-9aec-5eb2b03f8287], Nowicki-Osuch et al.[15] [https://cellxgene.cziscience.com/collections/a18474f4-ff1e-4864-af69-270b956cee5b], Chen et al.[16] [https://cellxgene.cziscience.com/collections/a48f5033-3438-4550-8574-cdff3263fdfd], Tabula Sapiens[25] [https://cellxgene.cziscience.com/collections/e5f58829-1a66-40b5-a624-9046778e74f5], Sikkema et al.[26] [https://cellxgene.cziscience.com/collections/6f6d381a-7701-4781-935c-db10d30de293], Chan et al.[27] [https://cellxgene.cziscience.com/collections/62e8f058-9c37-48bc-9200-e767f318a8ec]. The ST data were acquired from the following hyperlinks or accession numbers: Prostate cancer [https://www.10xgenomics.com/datasets/human-prostate-cancer-adenocarcinoma-with-invasive-carcinoma-ffpe-1-standard-1-3-0], Breast Cancer ([https://www.10xgenomics.com/datasets/human-breast-cancer-block-a-section-1-1-standard-1-1-0] and [https://www.10xgenomics.com/datasets/human-breast-cancer-ductal-carcinoma-in-situ-invasive-carcinoma-ffpe-1-standard-1-3-0], Renal cell carcinoma (GSE175540[40] and https://data.mendeley.com/datasets/g67bkbnhhg/1[39]), Squamous cell carcinoma (GSE144240[38]), Oral squamous cell carcinoma (GSE208253[41]). The cancer gene signatures used in this study were obtained from the following hyperlinks: 14 cancer functional states from CancerSEA[20] [http://biocc.hrbmu.edu.cn/CancerSEA/goDownload], 10 cancer hallmarks[22] [https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-018-25076-6/MediaObjects/41598_2018_25076_MOESM10_ESM.xlsx], 5 epithelial cell states from EcoTyper[21] [https://ars.els-cdn.com/content/image/1-s2.0-S0092867421010618-mmc4.xlsx]. List of known targets with approved drugs was obtained from PHAROS database[34] [https://pharos.nih.gov/]. Source data for the figures can be found in the Supplementary Data 5–15.

Code availability

The open-source package of scMalignantFinder is available on GitHub (https://github.com/Jonyyqn/scMalignantFinder). Additional scripts to reproduce the figures in this study are deposited in Zenodo (https://doi.org/10.5281/zenodo.12194623)[56].

Received: 30 June 2024; Accepted: 17 March 2025;
Published online: 27 March 2025

References

1. Fan, J., Slowikowski, K. & Zhang, F. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Exp. Mol. Med.* **52**, 1452–1465 (2020).
2. Dagogo-Jack, I. & Shaw, A.T. Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* **15**, 81–94 (2018).
3. Aran, D. Single-cell RNA sequencing for studying human cancers. *Annu Rev. Biomed. Data Sci.* **6**, 1–22 (2023).
4. Zhang, Y. et al. Single-cell RNA sequencing in cancer research. *J. Exp. Clin. Cancer Res.* **40**, 81 (2021).

5. Nofech-Mozes, I., Soave, D., Awadalla, P. & Abelson, S. Pan-cancer classification of single cells in the tumour microenvironment. *Nat. Commun.* **14**, 1615 (2023).
6. Yang, T., Yan, Q., Long, R., Liu, Z. & Wang, X. PreCanCell: an ensemble learning algorithm for predicting cancer and non-cancer cells from single-cell transcriptomes. *Comput. Struct. Biotechnol. J.* **21**, 3604–3614 (2023).
7. Patel, A.P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
8. Gao, R. et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat. Biotechnol.* **39**, 599–608 (2021).
9. Taylor, A.M. et al. Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell* **33**, 676–689.e673 (2018).
10. Dohmen, J., et al. Identifying tumor cells at the single-cell level using machine learning. *Genome Biol.* **23**, 123 (2022).
11. Zhong, Z., et al. Domain generalization enables general cancer cell annotation in single-cell and spatial transcriptomics. *Nat. Commun.* **15**, 1929 (2024).
12. Luecken, M.D. & Theis, F.J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
13. Pasquini, G., Rojo Arias, J.E., Schafer, P. & Busskamp, V. Automated methods for cell type annotation on scRNA-seq data. *Comput Struct. Biotechnol. J.* **19**, 961–969 (2021).
14. Salcher, S. et al. High-resolution single-cell atlas reveals diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer. *Cancer Cell* **40**, 1503–1520.e1508 (2022).
15. Nowicki-Osuch, K. et al. Single-cell RNA sequencing unifies developmental programs of esophageal and gastric intestinal metaplasia. *Cancer Discov.* **13**, 1346–1363 (2023).
16. Chen, B. et al. Differential pre-malignant programs and microenvironment chart distinct paths to malignancy in human colorectal polyps. *Cell* **184**, 6262–6280.e6226 (2021).
17. Lee, H.O. et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat. Genet.* **52**, 594–603 (2020).
18. Ma, L. et al. Single-cell atlas of tumor cell evolution in response to therapy in hepatocellular carcinoma and intrahepatic cholangiocarcinoma. *J. Hepatol.* **75**, 1397–1408 (2021).
19. Hanahan, D. Hallmarks of cancer: new dimensions. *Cancer Discov.* **12**, 31–46 (2022).
20. Yuan, H. et al. CancerSEA: a cancer single-cell state atlas. *Nucleic Acids Res.* **47**, D900–D908 (2019).
21. Luca, B.A. et al. Atlas of clinically distinct cell states and ecosystems across human solid tumors. *Cell* **184**, 5482–5496.e5428 (2021).
22. Iorio, F., et al. Pathway-based dissection of the genomic heterogeneity of cancer hallmarks' acquisition with SLAPenrich. *Sci. Rep.* **8**, 6713 (2018).
23. Gavish, A. et al. Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. *Nature* **618**, 598–606 (2023).
24. Kinker, G.S. et al. Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nat. Genet.* **52**, 1208–1218 (2020).
25. Tabula Sapiens, C. et al. The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
26. Sikkema, L. et al. An integrated cell atlas of the lung in health and disease. *Nat. Med.* **29**, 1563–1577 (2023).
27. Chan, J.M. et al. Signatures of plasticity, metastasis, and immunosuppression in an atlas of human small cell lung cancer. *Cancer Cell* **39**, 1479–1496.e1418 (2021).
28. Lu, Y., et al. A single-cell atlas of the multicellular ecosystem of primary and metastatic hepatocellular carcinoma. *Nat. Commun.* **13**, 4594 (2022).
29. Shao, X. et al. Copy number variation is highly correlated with differential gene expression: a pan-cancer study. *BMC Med. Genet.* **20**, 1–14 (2019).
30. Jun, J.C., Rathore, A., Younas, H., Gilkes, D. & Polotsky, V.Y. Hypoxia-inducible factors and cancer. *Curr. Sleep. Med. Rep.* **3**, 1–10 (2017).
31. Zhao, Z., Rahman, M.A., Chen, Z.G. & Shin, D.M. Multiple biological functions of Twist1 in various cancers. *Oncotarget* **8**, 20380–20393 (2017).
32. Zhang, P. et al. Dissecting the single-cell transcriptome network underlying gastric premalignant lesions and early gastric cancer. *Cell Rep.* **27**, 1934–1947.e1935 (2019).
33. de Vries, A.C. et al. Gastric cancer risk in patients with premalignant gastric lesions: a nationwide cohort study in the Netherlands. *Gastroenterology* **134**, 945–952 (2008).
34. Sheils, T.K. et al. TCRD and Pharos 2021: mining the human proteome for disease biology. *Nucleic Acids Res.* **49**, D1334–D1346 (2021).
35. Maniatis, S., Petrescu, J. & Phatnani, H. Spatially resolved transcriptomics and its applications in cancer. *Curr. Opin. Genet. Dev.* **66**, 70–77 (2021).
36. Longo, S.K., Guo, M.G., Ji, A.L. & Khavari, P.A. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat. Rev. Genet.* **22**, 627–644 (2021).
37. Ru, B., Huang, J., Zhang, Y., Aldape, K. & Jiang, P. Estimation of cell lineages in tumors from spatial transcriptomics data. *Nat. Commun.* **14**, 568 (2023).
38. Ji, A.L. et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinome. *Cell* **182**, 1661–1662 (2020).
39. Li, R. et al. Mapping single-cell transcriptomes in the intra-tumoral and associated territories of kidney cancer. *Cancer Cell* **40**, 1583–1599.e1510 (2022).
40. Meylan, M. et al. Tertiary lymphoid structures generate and propagate anti-tumor antibody-producing plasma cells in renal cell cancer. *Immunity* **55**, 527–541.e525 (2022).
41. Arora, R., et al. Spatial transcriptomics reveals distinct and conserved tumor core and edge architectures that predict survival and targeted therapy response. *Nat. Commun.* **14**, 5029 (2023).
42. Xun, Z., et al. Reconstruction of the tumor spatial microenvironment along the malignant-boundary-nonmalignant axis. *Nat. Commun.* **14**, 933 (2023).
43. Baslan, T. et al. Novel insights into breast cancer copy number genetic heterogeneity revealed by single-cell genome sequencing. *Elife* **9**, e51480 (2020).
44. Barkley, D. et al. Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. *Nat. Genet.* **54**, 1192–1201 (2022).
45. Swanton, C. Intratumor heterogeneity: evolution through space and time. *Cancer Res.* **72**, 4875–4882 (2012).
46. Ghaddar, B. & De, S. Hierarchical and automated cell-type annotation and inference of cancer cell of origin with census. *Bioinformatics* **39**, btad714 (2023).
47. Chang, J., Zheng, T. & Wu, C. Early cancer detection through comprehensive mapping of dynamic tumorigenesis. *Cancer Discov.* **14**, 2037–2040 (2024).
48. Rao, A., Barkley, D., Franca, G.S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).
49. Fiore, M. et al. The biology of synovial sarcoma: state-of-the-art and future perspectives. *Curr. Treat. Options Oncol.* **22**, 109 (2021).
50. Kim, H.J., Park, J.W. & Lee, J.H. Genetic architectures and cell-of-origin in glioblastoma. *Front. Oncol.* **10**, 615400 (2020).
51. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e1821 (2019).
52. Kuleshov, M.V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).

53. Dann, E. et al. Single-cell RNA sequencing of human tissue supports successful drug targets. *medRxiv*, http://www.medrxiv.org/content/10.1101/2024.04.04.24305313v1 (2024).
54. Palla, G. et al. Squidpy: a scalable framework for spatial omics analysis. *Nat. Methods* **19**, 171–178 (2022).
55. Megill, C. et al. Cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. http://www.biorxiv.org/content/10.1101/2021.04.05.438318v1 (2021).
56. Yu, Q., Li, Y. & Chen, Y. scMalignantFinder distinguishes malignant cells in single-cell and spatial transcriptomics by leveraging cancer signatures. *Zenodo*, https://doi.org/10.5281/zenodo.12194623 (2025).

Acknowledgements

This work was supported by a grant from the National Natural Science Foundation of China (32100531).

Author contributions

Conceptualization: Y.Y.L. and Y.C.; investigation: Y.C., Q.Y., and Y.Y.L.; writing: Q.Y., Y.Y.L., and Y.C.; supervision: Y.Y.L. and Y.C.

Competing interests

The authors declare the following competing interests: Qiaoni Yu and Yunqin Chen are employees of Genebase. The remaining authors declare no competing interest.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-025-07942-y.

Correspondence and requests for materials should be addressed to Yuan-Yuan Li or Yunqin Chen.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Debarka Sengupta and Johannes Stortz. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.