Journal of Cancer Research and Clinical Oncology (2022) 148:2313–2322
https://doi.org/10.1007/s00432-022-04020-2

ORIGINAL ARTICLE – CANCER RESEARCH

Single-cell transcriptome highlights a multilayer regulatory network on an invasive trajectory within colorectal cancer progression

Tong Zhou[1] · **Chunhua Li**[1]

Received: 2 March 2022 / Accepted: 9 April 2022 / Published online: 6 May 2022

Abstract

Purpose Colorectal cancer (CRC) is one of the most common and fatal gastrointestinal malignancies, in which cancer stem cells (CSCs) were identified to enable tumor heterogeneity and initiate tumor formation. However, the process from CSCs to invasion cells is unconfirmed.

Methods Several bioinformatics methods, including clustering, pseudotime analysis, gene set variation analysis and gene ontology enrichment, were used to construct a path of gradual transformation of CSCs to invasive cells, called "stem-to-invasion path". A large amount of signaling interactions were collated to build the multilayer regulatory network. Kaplan–Meier curve and time-dependent ROC method were applied to reveal prognostic values.

Results We validated the heterogeneity of cells in the tumor microenvironment and revealed the presence of malignant epithelial cells with high invasive potential within primary colonic carcinomas. Next, the "stem-to-invasion path" was identified through constructing a branching trajectory with cancer cells arranged in order. A multilayer regulatory network considered as the vital factor involved in acquiring invasion characteristics underlying the path was built to elucidate the interactions between tumor cell and tumor-associated microenvironment. Then we further identified a novel combinatorial biomarker that can be used to assess the prognosis for CRC patients, and validated its predictive robustness on the independent dataset.

Conclusion Our work provides new insights into the acquisition of invasive potential in primary tumor cells, as well as potential therapeutic targets for CRC invasiveness, which may be useful for the cancer research and clinical treatment.

Keywords Colorectal cancer · Single-cell RNA transcriptome · Stem-to-invasion path · Multilayer regulatory network · Novel combinatorial biomarker

Introduction

Colorectal cancer (CRC) is one of the most common and highly lethal malignancies of the digestive system, with approximately half of patients succumbing to cancer-related events mainly metastasis (Sung et al. 2020; Yilmaz and Christofori 2010). The CRC cells can metastasize to other sites such as lymph gland, liver, and lung, posing an enormous challenge for treatment. Generally, metastasis occurs in the mid to late stages of cancer development, which is considered to be attributed to a type of tumor cells with stem cell properties known as cancer stem cells (CSCs) (Wang et al. 2021; Liu et al. 2015; Hatano et al. 2017). Moreover, there is a growing evidence that epithelial-mesenchymal transition (EMT), a prerequisite mechanism for invasion and metastasis, is present throughout the cancer process (Mittal 2018; Linde et al. 2018; Hosseini et al. 2016). These suggest that cells in primary tumors may begin to acquire invasive capacity due to activation of the EMT pathway. Therefore, it is essential to investigate the regulatory relationships of genes during tumor development to identify the molecular determinants of aggressiveness.

Additionally, cell function and characteristics are not only determined by the cell itself, but are also influenced by the microenvironment (Boulanger et al. 2007). Several studies have shown that cell–cell interactions in tumor microenvironment play an important role in the growth and progression of cancer (Sistigu et al. 2020; Gu and Mooney 2016). Secretory molecules, a type of intercellular signals, can affect intracellular regulatory networks by binding to the

✉ Chunhua Li
chunhuali@bjut.edu.cn

[1] Department of Biomedical Engineering, Falcuty of Environmental and Life Sciences, Beijing University of Technology, Beijing 100124, China

receptors. Thus, it is significant to elucidate the signaling mechanisms underlying the interaction between the microenvironment and tumor cells with invasive characteristics.

A growing number of studies (Giladi and Amit 2017; Mathys et al. 2019; Li et al. 2017) demonstrates that the traditional bulk transcriptome analysis is a limited approach for dissecting molecular mechanisms because of cellular heterogeneity. Instead, single-cell RNA sequencing (scRNA-seq) technologies generated based on the resolution of individual cells (Tang et al. 2009) can identify distinct cellular expression signals, which provides a wonderful opportunity to determine pivotal molecular determinants for tumor progression and dissect the microenvironment-mediated signaling pathways.

Recently, Pang et al. (2019) revealed an invasion-associated progression path where cells gradually acquire the invasive potential, and identified key factors involved in glioblastoma progression, which provides a marvelous example for exploring key factors in tumor invasiveness. Zhang et al. (2020) constructed a multilayer signaling network that contains pathways from intercellular and intracellular interactions to explain cellular reciprocities in glioblastoma. On this basis, Cheng et al. (2021) updated the interaction data and developed a tool called scMLnet that can be applied to build a multilayer network. They then employed the scMLnet to a scRNA-seq dataset of COVID-19 as an example, to research microenvironmental regulation and reveal several key regulators of ACE2 expression. However, transcription factors (TFs) and their target genes have the cases of negative regulation (Han et al. 2018; Ye et al. 2018), while these studies (Zhang et al. 2020; Cheng et al. 2021) only consider the upregulated intracellular signaling, which is deficient in terms of signaling network mechanisms.

In this study, we used scRNA-seq data to reveal a tumor progression path representing the acquisition of invasive potential and reconstructed a multilayer regulatory network consisting of ligand–receptor interactions, TFs and their target genes, which are the pivotal factors for obtaining invasion characteristics. Moreover, we identified a novel combinatorial biomarker (NCB) that can be used to assess the prognosis for CRC patients and validated its predictive robustness on the independent test set, which can shed light on further individualized treatment (Fig. S1).

Materials and methods

Single-cell RNA-seq data processing and cell type identification

The single-cell transcriptome data in this work were obtained from Gene Expression Omnibus (GEO) database, including two accessions, GSE132465 (named SMC) and GSE144735 (named KUL3) (Lee et al. 2020). To get more objective results, we screened the patients according to the criteria in Text S1A. The information of the screened patients is shown in Table S1 and S2.

Subsequent operations were performed using the R package Seurat (version 4.0) (Satija et al. 2015). Low-quality cells were excluded conforming to the criteria described in Text S1B. After that, the gene expression matrices were normalized to the total UMI counts per cell and converted to the natural logarithmic scale, and the effects of cell cycle and mitochondrial genes were eliminated using the "ScaleData" function. We used the top 2000 highly variable genes to conduct PCA analysis, and clustered the cells with Shared-nearest-neighbor (SNN) algorithm (Jarvis and Patrick 2006). After that, t-distributed stochastic neighbor embedding (t-SNE) method was further implemented for visualization. Finally, cell type annotations were mainly performed by R package singleR (version 1.4) (Aran et al. 2019) and later manually revised with the CellMarker database (Zhang et al. 2019).

Copy number variation (CNV) estimation

We calculated the CNVs for each gene region of each epithelium from tumor samples using the epithelial cells from para-carcinoma tissues as a reference via R package inferCNV (version 1.8) (Patel et al. 2014), and limited the values as -1 to 1. The CNV score of each cell was computed as the quadratic sum of the CNVs in each region. We further calculated the median and standard deviation of CNV scores, which are applied to determine the threshold value k as:

$$k = M - s \tag{1}$$

where M and s are the median and standard deviation of CNV scores. The epithelial cells, whose CNV score $> k$, were defined as malignant ones.

Subclustering and pseudotime analysis of tumor cells

Subclustering and trajectory construction for malignant epithelial cells were performed by R package monocle (version 2.18) (Qiu et al. 2017). Genes with the minimum expression > 0.1 and expressed in $\geq 5\%$ of all cells were selected for subclustering. After subclustering, we selected the genes that were differential expressed between clusters, and adopted DDRTree algorithm (Qi et al. 2015) to reduce the given high-dimensional expression profiles to a low-dimensional space in which single cells are ordered into a trajectory. Then, we found a “stem-to-invasion path” and

extracted the protein-coding genes that changed significantly along the path.

Evaluation of activity for diverse pathways and genes

Gene set variation analysis (GSVA) (Hänzelmann et al. 2013) was implemented to evaluate the relative activation status for pathways. GSVA scores for cancer hallmark and signaling pathways were calculated using predefined gene sets downloaded from MSigDB (Liberzon et al. 2015). For the CSC and invasiveness scores, we averaged the expression levels of the signatures associated with their respective functions, which were manually extracted from previous studies (Tables S3 and S4). The Gene Ontology (GO) enrichment analysis was completed by R package clusterProfiler (version 4.0) (Wu et al, 2021).

Gene expression analysis by hidden Markov model (HMM)

We applied an HMM to predict gene expression states (on or off) throughout pseudotime as described by Shin et al. (2015). Briefly, we divided the pseudotime into 23 units, where cells have the same state for most genes, and the average expression level of each gene in each unit was used as the observed variables for HMM. After that, the Baum–Welch algorithm (Welch 2003) was used to extract the most likely emission probabilities and transition probabilities. Finally, the Viterbi algorithm (Forney 1973) was applied to predict binary gene expression states, using the observed variables along with output from the Baum–Welch algorithm.

Construction of multilayer regulatory network

The intercellular/intracellular signaling interactions (i.e. ligand–receptor pairs, receptor-TF pathways, TF-target gene interaction, TF sub-network and target gene sub-network) were constructed on the SMC and KUL3 datasets, respectively, using the methods described in Text S2. We then took the intersection of the two datasets for each relationship and assembled them to construct a multilayer regulatory network. The software Cytoscape (version 3.9) (Shannon et al. 2003) was used to characterize each relationship and the integrated network.

Identification and evaluation of prognostic significance of the NCB

Given the multilayer regulatory network, we further investigated the signatures from the multilayer regulatory network associated with survival across the CRC samples. In the view of this, we collected bulk RNA-seq data and clinical information of the TCGA-COAD and TCGA-READ datasets from The Cancer Genome Atlas (TCGA) database as the training dataset and those of GSE17536 (Smith et al. 2010) and GSE29621 (Chen et al. 2012) from GEO database as independent testing datasets. The raw data in all these datasets were preprocessed with the criteria mentioned in Text S1C. Finally, after preparing the training set ($N=354$) and the independent test sets, i.e. GSE17536 ($N=177$) and GSE29621 ($N=65$), we normalized the gene expression matrix by log2(TPM + 1) for subsequent survival analysis.

Stepwise regression analysis (Hastie and Pregibon 1992), based on Akaike information criterion (Sakamoto et al. 1986), was performed to identify the prognostic hub genes from the multilayer regulatory network. After that, we built a multivariate Cox regression model (Cox and Oakes 1984) using the hub genes and formulated the following risk score (RS) for predicting patient survival:

$$\mathrm{RS} = \sum_{i=1}^{n} \lambda_i \cdot y_i \tag{2}$$

where y_i is the expression level of gene i, n is the number of genes, and λ_i is the regression coefficient of gene i in the Cox regression model. An NCB risk signature was trained by the training dataset: RS = HSPG2 × (−0.3292) + SERPINE1 × (0.4537) + MMP12 × (−0.1552) + SCARB1 × 0.5605 + CEBPA × (−0.3319) + GPX4 × 0.3904 + SPRR3 × 0.4386 + PLAUR × (−0.4123). The patients in each dataset were divided into two groups, high-risk and low-risk, based on the optimal cut-off value for the maximum sum of sensitivity and specificity of ROC method. We estimated the prognostic performance of the NCB by Kaplan–Meier survival curves with the log-rank test and time-dependent ROC analysis (Heagerty et al. 2000).

Results

Cell type and malignant epithelial cells identification

We obtained single-cell RNA-seq data from the SMC dataset to explore the cellular diversity. After quality control and normalization, we analyzed a total of 34,558 cells from 12 colorectal tumor samples and 6 control paracancerous tissue. Next, SNN algorithm and t-SNE method were conducted to identify 6 main clusters, including epithelial cells, stromal cells, myeloid cells, T cells, B cells and neurons (Fig. 1A–B). From Fig. 1C, we observed that most of epithelial cells, derived from tumor tissues within dramatical variations in proportion among patients, formed patient-specific clusters, suggesting high heterogeneity, shown in Fig. 1C.

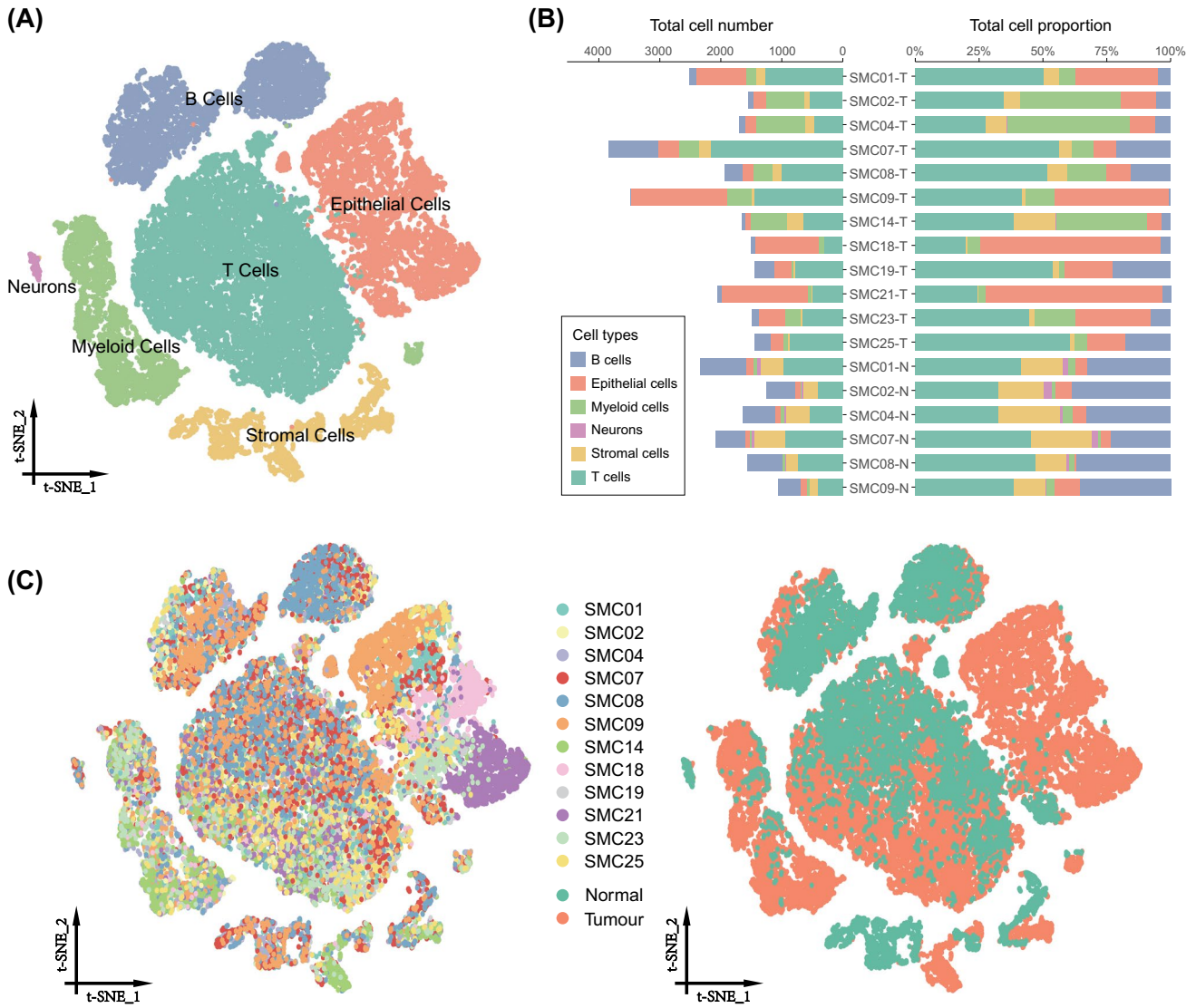


Fig. 1 Cell type identification in the SMC dataset. **A** t-SNE plot of 34,558 cells colored by global cell types. **B** Amounts and proportions of the global cell types in individual CRC samples. **C** t-SNE plots colored by different patients (left) and tumor or normal tissue (right)

This prompted us to investigate their malignant status. We calculated large-scale chromosomal CNVs in the epithelial cells from tumor tissues based on the expression pattern of the ones from tumor-adjacent samples. Next, we excluded the epithelial cells whose CNV score lower than the threshold value (see "Materials and methods" section), and finally obtained 5915 malignant epithelial cells for subsequent analysis.

Heterogeneity of malignant epithelial cells

Tumor heterogeneity is the prime factor leading to cancer progression and therapy failure (Hanahan and Weinberg 2011). We employed subclustering analysis for the malignant epithelial cells and identified 7 subclusters (Fig. 2A). Different expression of genes between subclusters refer to divergent cancer cell characteristics. Therefore, GSVA was used to estimate the functions of cells within each subcluster (Fig. 2B). We chose pathways associated with tumor progression to define the subcluster status and found that Hedgehog and MAPK signaling were enriched in subcluster 6, while mTOR and TGF-β signaling-related genes were highly activated in subcluster 2. Besides, subcluster 1 was expressed MAPK and WNT signaling-related genes, but subcluster 5 was not significantly enriched in genes related to cancer progression compared to other subclusters. In addition, subcluster 3 showed higher expression of G2M checkpoint and DNA repair that control the cell cycle pathway. Subcluster 4 expressed higher stem cell-related genes than others. Subcluster 7 was closely related

to EMT and angiogenesis-associated pathways, implying an invasive and metastatic potential. These results indicate that divergent malignant epithelium subpopulations in CRC have various status, reflecting diverse tumor biology functions.

The "stem-to-invasion path" in the branching structure of malignant colonic epithelia

Based on the GSVA results that subcluster 4 displaying stem cell-like property and subcluster 7 indicating invasive and metastatic potency, we conjectured that single-cell transcriptome might reveal the main variational processes of CSCs during tumor progression. Later, we used trajectory analysis to reconstitute the pseudotime of the malignant epithelia, which contained mainly 7 cell states (Fig. 2C–D). From Fig. 2C, the majority of the cells within state 1 came from subcluster 4, while almost all subcluster 7 cells was converging on state 6, indicating that state 1 and 6 cells might perform possessed the functional annotation and status characterization of subcluster 4 and 7, respectively. Thus, we speculated that there is a path, denoted "stem-to-invasion path", representing the progression from colonic CSCs to invasive cells in pseudotime. To conform the inference, CSC score and invasiveness score were utilized (see "Materials and methods" section) to evaluate the states of the cancer cells on the trajectory. As shown in Fig. 2E–H, we noticed the CSC score was highest in state 1 and the invasiveness score of state 6 was higher than others, which were similar to the dynamic trend of two scores. Hence, we realized that the cells traveled from state 1 through branch point 3, state 2, branch point 1, and finally to state 6, which represent the "stem-to-invasion path".

Then, we sought to decipher the biological functions during the dynamic process. The genes for encoding proteins with significantly changed expression levels (false discovery rate < 1e−4) were captured on the stem-to-invasion path and clustered into two groups (Fig. S2A). GO enrichment analysis was performed and revealed that the genes enriched for biological process terms such as ATP metabolic process, cellular response to hypoxia, regulation of epithelium morphogenesis and stem cell differentiation were upregulated, whereas the genes downregulated were mainly enriched in cell projection assembly and intrinsic apoptosis. We calculated the Spearman correlations between each gene and pseudotime in the two clustered groups (p-value < 0.05), and generated two lists of the top 100 positively/negatively correlated genes to determine their binary on/high or off/low expression state using an HMM (Fig. S2B–C). The results indicate that the stem-to-invasion path can partially reflect the progression of colorectal carcinoma from CSCs to invasive cells.

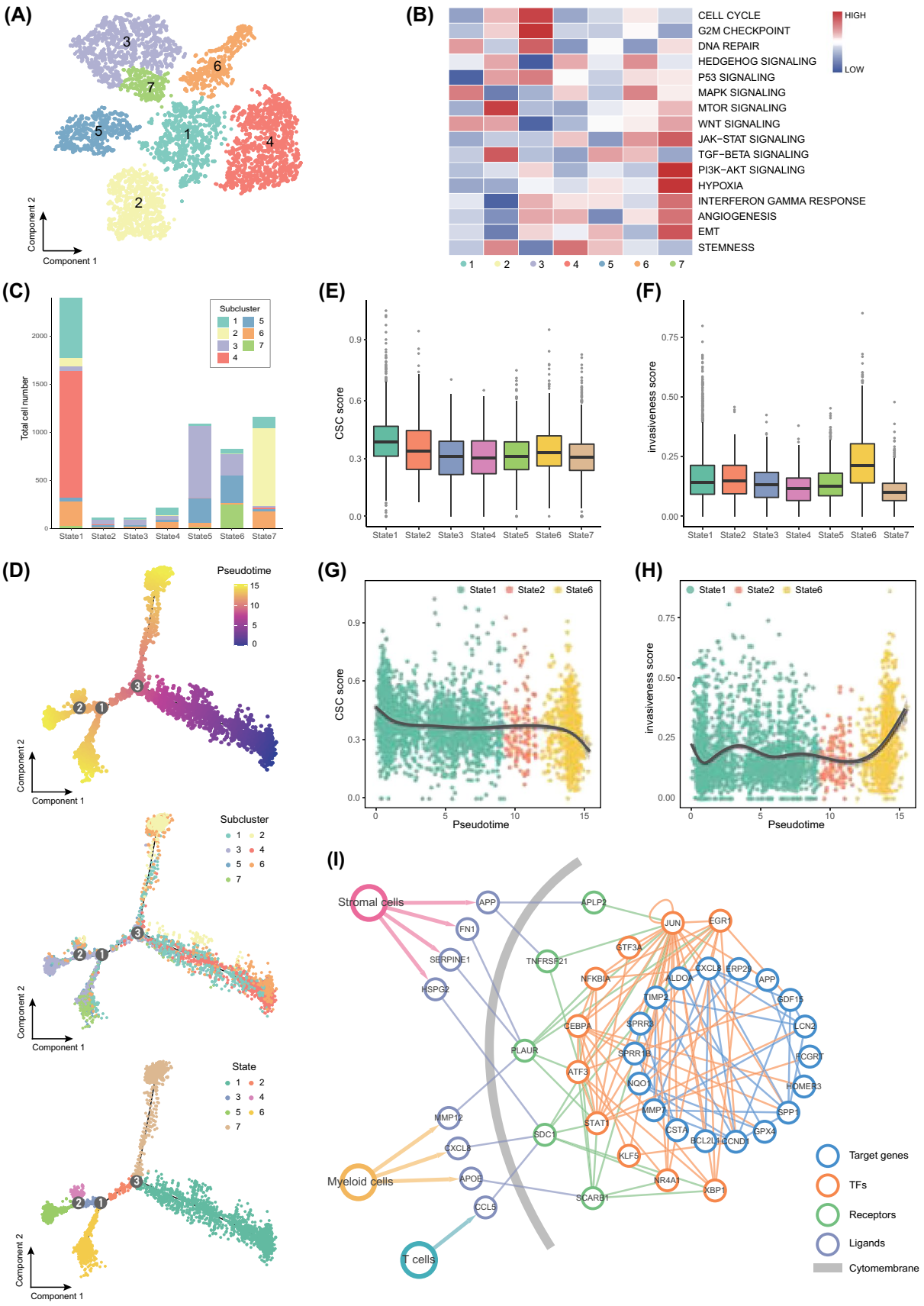
The recurrence of a similar "stem-to-invasion path" in an extra data of colonic cancer epithelial cells

To validate whether the stem-to-invasion path could be reproduced, another single-cell RNA-seq dataset (the KUL3 dataset) was obtained that contains 19,389 cells from five patients. After the same data processing, we finally acquired 2881 epithelial cells (Fig. S3), of which 1828 are malignant, and identified 7 subclusters from these malignant epithelia (Fig. S4A).

Later, we explored the enrichment pathways of the cells in each subcluster by GSVA and reconstructed a trajectory (Fig. S4B–D). We noticed that the cells in subcluster 5, corresponding to state 2, were enriched for EMT and angiogenesis genes, while stem cell-related genes enriched in cells of subcluster 1, which was distributed over state 3. Based on the experience with the SMC dataset, we concluded that the "stem-to-invasion path" was oriented from state 3, through branch point 1, to state 2. Furthermore, similar results were obtained, which confirmed our surmise (Fig. S4E–K). These suggest that we recaptured a similar "stem-to-invasion path" in another dataset of CRC.

Identification of crucial molecules and a multilayer regulatory network in the acquisition of invasive potential

Considering the progressive changes in genes on the trajectories, we tried to identify the molecules that drive the invasive potential. Due to the vital role of lncRNAs in tumor development (Braga et al. 2020), we identified three upregulated (NEAT1, PP7080 and CRNDE) and seven downregulated lncRNAs (RP11-160E2.6, RP11-462G2.1, CH17-373J23.1, EPB41L4A-AS1, SLCO4A1-AS1, THUMPD3-AS1 and SNHG9) shared by both two datasets (Table S5). In addition to most of the lncRNAs have been proven to be involved in invasion and metastasis by many researches, we found three new lncRNAs, RP11-160E2.6, RP11-462G2.1 and CH17-373J23.1, which provided new insights for further experimental studies on the factors affecting tumor invasion. Afterwards, we extracted the differently expressed protein-coding genes from the stem-to-invasion paths to construct an intracellular signaling network. Moreover, cells in the tumor-associated microenvironment interact with each other, we thus investigated the interactions between the cancer cells on the stem-to-invasion path and the other cells by constructing intercellular signaling relationships, and further connected them with the intracellular network to build a multilayer regulatory network, which can reflect the molecular regulatory relationships within carcinoma cells and the signal transmission between cells in the tumor microenvironment during the acquisition of invasive potential.


(A)
3
6
7
5
1
4
2
Component 2
Component 1
(B)
CELL CYCLE
G2M CHECKPOINT
DNA REPAIR
HEDGEHOG SIGNALING
P53 SIGNALING
MAPK SIGNALING
MTOR SIGNALING
WNT SIGNALING
JAK−STAT SIGNALING
TGF−BETA SIGNALING
PI3K−AKT SIGNALING
HYPOXIA
INTERFERON GAMMA RESPONSE
ANGIOGENESIS
EMT
STEMNESS
HIGH
LOW
1 2 3 4 5 6 7
(C)
Subcluster
Total cell number
State1 State2 State3 State4 State5 State6 State7
(D)
Pseudotime
Subcluster
State
Component 2
Component 1
(E)
CSC score
(F)
invasiveness score
(G)
State1 State2 State6
CSC score
Pseudotime
(H)
State1 State2 State6
invasiveness score
Pseudotime
(I)
Stromal cells
Myeloid cells
T cells
APP
APLP2
FN1
SERPINE1
HSPG2
TNFRSF21
JUN
EGR1
GTF3A
NFKBIA
ALDOA
CXCL3
ERP29
TIMP3
GDF15
CEBPA
SPRR3
LCN2
PLAUR
SPRR1B
FCGRT
ATF3
NQO1
HOMER3
MMP12
MMP7
SPP1
STAT1
CSTA
GPX4
SDC1
BCL2L1
CCND1
CXCL8
KLF5
APOE
NR4A1
XBP1
SCARB1
CCL5
Target genes
TFs
Receptors
Ligands
Cytomembrane

◂**Fig. 2** The stem-to-invasion path built by the SMC dataset and multilayer network construction. **A** t-SNE plot of malignant epithelia showing seven subclusters. **B** Heatmap for the status characterization of each subcluster scored by GSVA software. **C** The amount and proportion of cells for each subcluster in seven states. **D** The single-cell trajectory containing seven states. Cells are colored based on pseudotime (top), subcluster (middle) and state (bottom). **E, F** Boxplot for the CSC scores (**E**) and the invasiveness scores (**F**) for each state. **G**, **H** Curve chart for the CSC scores (**G**) and the invasiveness scores (**H**) as a nonlinear function of pseudotime in the path containing states 1, 2 and 6 cells. **I** Multilayer regulatory network of tumor cells communicated with other cells. The nodes with different colors represent ligands (gray), receptors (green), TFs (orange) or target genes (blue)

A total of 1594 upregulated and 1816 downregulated mRNAs were derived from the stem-to-invasion path of the SMC dataset. Then, we gained 37 TF-target gene interactions and 55 receptor-TF links according to the TF-target gene interactions list and the receptor-TF links list. For the KUL3 dataset, after a similar operation, we picked out 2215 upregulated and 628 downregulated mRNAs and sought out 36 TF-target gene interactions and 50 receptor-TF connections. In addition, for each dataset, we calculated the highly expressed ligand genes for other non-epithelial cell types, and selected the upregulated receptor genes from the trajectory to establish the ligand-receptor relations. As a result, we detected 10 ligand-receptor pairs from SMC dataset and 10 from the KUL3 dataset.

Particularly, there was significant overlap for each interaction between the two datasets, including 36 TF-target gene links, 38 receptor-TF connections and 9 ligand-receptor pairs. Furthermore, we obtained 25 TF-TF regulatory relations and a 28-target gene-regulatory sub-network. All these constructed intercellular/intracellular signaling pathways are shown in Fig. S5. Then, a multilayer regulatory network (Fig. 2I) was constructed by integrating all intercellular pathways and intracellular sub-networks, representing the signaling transduction from the non-epithelial secreted ligands, to the tumor cell receptors and then to downstream TFs and target genes. Table S6 lists other research evidences for the functions of the genes on the multilayer network, which were considered to be the key mRNAs implicated in the "stem-to-invasion" progression.

The prognostic significance of the NCB

Based on the multilayer regulatory network, we established an NCB risk signature using stepwise regression and multivariate Cox regression algorithms in the training cohort. The NCB contained 8 genes: HSPG2, SERPINE1, MMP12, SCARB1, PLAUR, CEBPA, GPX4 and SPRR3.

We evaluated the prognostic significance of the NCB on the training set and the two independent testing sets. As shown in Fig. 3A, Kaplan–Meier analysis indicated revealed significantly worse overall survival (OS) outcomes ($p < 0.001$) for the patients of the high-risk group in all datasets. Next, we calculated the AUC of the time-dependent ROC with the 1-year, 3-year and 5-year survival of the patients to access the prognostic accuracy of the NCB (Fig. 3D). We found the NCB had a good performance with a 1-year AUC of 0.718, a 3-year AUC of 0.696 and a 5-year AUC of 0.768. In parallel, similar statistical results were observed in the GSE17536 and GSE29621 cohorts (Fig. 3B, C, E, F).

We performed univariate and multivariate Cox regression analyses on the NCB and other clinicopathologic factors including age, gender and stage, and found that the NCB was an independent prognostic signature for the OS of CRC patients in all datasets (Tables 1, S7 and S8). Later, time-dependent ROC curves were adopted to compare the prognostic accuracy of all clinical signatures. The AUCs of ROC curves (Fig. 3G–I) showed that the NCB signature were superior to other risk signatures in predicting the 3-year survival rates. These results confirm that the NCB possess good prognostic power and is almost as accurate as the other common clinicopathologic factors.

Discussion

Most deaths from solid tumors, including CRC, are caused by invasion and metastasis of carcinoma cells. In the view of this, we sought to explore the significant molecules driving the acquisition of invasive potential and their regulatory relationships. Based on the phenotypic diversity of cells in colon cancer microenvironment and the fact that EMT occurs early in many cancer types, we speculated there might be a type of malignant epithelial cells owning high invasive potential. Thus, we first identified the cell types and detach the data of malignant epithelial cells. Next, subclustering analysis and GSVA was performed to recognize the different functions of the cells. Moreover, considering that cancer initiation, proliferation, invasion and metastasis occur as a continuous process, we introduced the pseudotime analysis to construct a trajectory revealing tumor progression, and captured the significantly expressed genes on the trajectory at the transcriptome level. Combining the results of GSVA, we realized that the root of the trajectory showed a high CSC character, and another branch was enriched with the cells of high invasive potential. According to this, we identified a cell-traveling path ("stem-to-invasion path") from the root to the invasive branch, which represents the transition process from CSCs to invasive cells. The same procedures were run in the two scRNA-seq datasets (SMC and KUL3) and the similar observations were revealed, which we think could be used to explore the molecular events that promote the acquisition of invasive potential in CRC progression. Therefore, we further utilized the differentially expressed

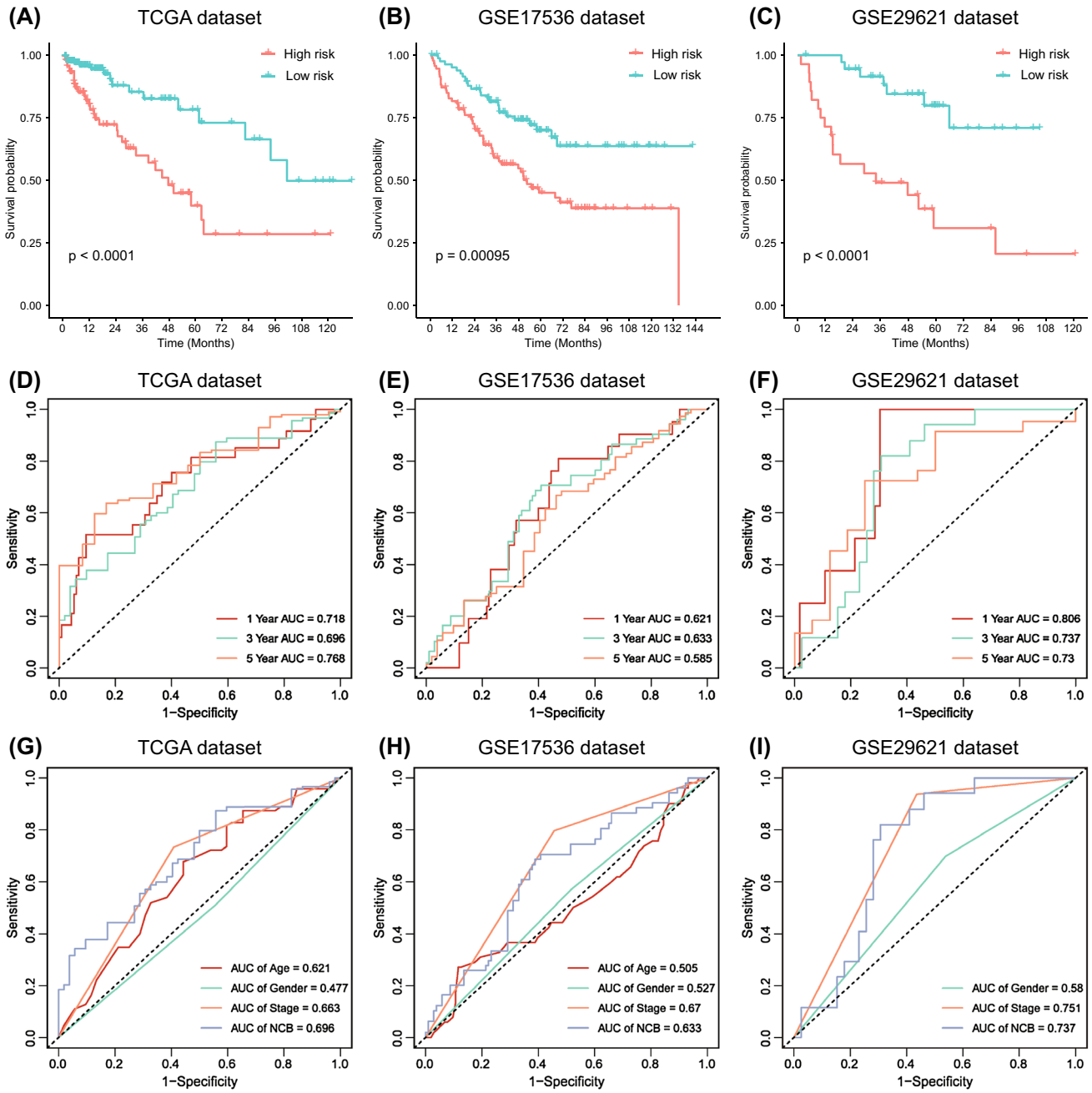


Fig. 3 The prognostic significance and accuracy of the NCB. **A–C** The Kaplan–Meier survival curves for prognostic significance assessed by the TCGA (**A**), the GSE17536 (**B**) and the GSE29621 dataset (**C**). **D–F** The time-dependent ROC curves for prognostic accuracy evaluated by the AUC with respect to 1-year, 3-year and 5-year survival of CRC patients on the TCGA (**D**), the GSE17536 (**E**) and the GSE29621 dataset (**F**). **G–I** Time-dependent ROC curves comparing the prognostic accuracy by age, gender, stage and the NCB with respect to 3-year survival on the TCGA (**G**), the GSE17536 (**H**) and the GSE29621 dataset (**I**)

Table 1 Univariate and multivariate Cox regression analysis of clinicopathologic factors (age, gender and stage) and the NCB signature for OS prediction in the TCGA dataset

Variable	Univariate cox		Multivariate cox	
	p-value	HR (95% CI)	*p*-value	HR (95% CI)
Age	0.025	1.029 (1.004–1.056)	0.003	1.040 (1.013–1.068)
Gender (male vs. female)	0.305	1.344 (0.764–2.367)	0.527	1.213 (0.667–2.203)
Stage (III & IV vs. I & II)	0.004	2.481 (1.340–4.594)	0.009	2.329 (1.236–4.367)
NCB	<0.001	2.718 (1.906–3.876)	<0.001	2.723 (1.896–3.910)

genes on the stem-to-invasion path and the ligand genes expressed by other cells to reconstruct a multilayer regulatory network, which consists of the pathways from intercellular ligand-receptor pairs to intracellular TF-target gene interactions. Notably, based on the network, we identified an NCB for tumor prognostic prediction, which had a good performance on the training set and two testing sets. In addition, we proved that the NCB possessed better prognostic accuracy and robustness compared to other traditional clinical signatures. This good prognostic value of the invasion-related genes confirms that invasion and metastasis of cancer cells are important causes of patient death.

We acknowledge that the above work is based on some assumptions and simplifications. The path for carcinoma cells gaining invasive potential was simulated by sorting the cells according to the gene expressions using transcriptome data from individual time points rather than multiple time points, due to the difficulties of data acquisition and processing. Furthermore, intercellular and intracellular signal transduction pathways involve complex post-translational modifications as well as protein-nucleic acid interactions. Nonetheless, due to the hardship for collecting high throughput single-cell proteomic data, the use of scRNA-seq to estimate protein activity in cells is an alternative approach, as presented in the work of Zhang et al. (2020) In the future, we will collect more information, especially proteomic and time-course transcriptomic data, to update our work.

Additionally, there are still other deficiencies for further improvement. First, it is well known that there is a close relationship between mRNA and ncRNA (e.g. miRNA and lncRNA), nonetheless, an appropriate method is currently pending to investigate the regulation mode between ncRNA and mRNA in the malignant cells during acquiring the invasive potential. Moreover, though the essential genes in the multilayer regulatory network had been identified and their prognostic effect had been demonstrated, complementary basic experiments are still necessary to reveal the specific mechanisms of NCB in the promotion pf tumor development.

In summary, this study used CRC single-cell RNA-seq data to reveal the process by which tumor cells acquire invasive characteristics and establish a regulatory network using significant factors in the process, from which the NCB was shown to be of good prognostic value for CRC patients. These may provide a new perspective for characterizing the invasion in CRC tumor microenvironment, which may be useful for the cancer research and clinical treatment.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00432-022-04020-2.

Acknowledgements We thank Dr. Jianming Zeng (University of Macau), and all the members of his bioinformatics team, biotrainee, for generously sharing their experience. This work was supported by the National Natural Science Foundation of China (31971180).

Author contributions Tong Zhou: designed the project, performed analysis, made figures, and wrote the manuscript. Chunhua Li: designed the project, revised the manuscript, and supplied the funding. All authors discussed results and commented on the manuscript.

Funding This study was funded by the National Natural Science Foundation of China (31971180).

Declarations

Conflict of interest The authors state no conflict of interest.

References

Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, Butte AJ, Bhattacharya M (2019) Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol 20:163–172

Boulanger CA, Mack DL, Booth BW, Smith GH (2007) Interaction with the mammary microenvironment redirects spermatogenic cell fate in vivo. Proc Natl Acad Sci USA 104:3871–3876

Braga EA, Fridman MV, Moscovtsev AA, Filippova EA, Dmitriev AA, Kushlinskii NE (2020) LncRNAs in ovarian cancer progression, metastasis, and main pathways: ceRNA and alternative mechanisms. Int J Mol Sci 21:8855

Chen DT, Hernandez JM, Shibata D, McCarthy SM, Humphries LA, Clark W, Elahi A, Gruidl M, Coppola D, Yeatman T (2012) Complementary strand microRNAs mediate acquisition of metastatic potential in colonic adenocarcinoma. J Gastrointest Surg 16:905–912

Cheng J, Zhang J, Wu Z, Sun X (2021) Inferring microenvironmental regulation of gene expression from single-cell RNA sequencing data using scMLnet with an application to COVID-19. Brief Bioinform 22:988–1005

Cox DR, Oakes D (1984) Analysis of survival data. Chapman & Hall, London

Forney GD (1973) The Viterbi Algorithm. Proc IEEE 61:268–278

Giladi A, Amit I (2017) Immunology, one cell at a time. Nature 547:27–29

Gu L, Mooney DJ (2016) Biomaterials and emerging anticancer therapeutics: engineering the microenvironment. Nat Rev Cancer 16:56–66

Han H, Cho JW, Lee S, Yun A, Kim H, Bae D, Yang S, Kim CY, Lee M, Kim E, Lee S, Kang B, Jeong D, Kim Y, Jeon HN, Jung H, Nam S, Chung M, Kim JH, Lee I (2018) TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res 46:D380–D386

Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144:646–674

Hänzelmann S, Castelo R, Guinney J (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinform 14:7

Hastie TJ, Pregibon D (1992) Statistical models in S. In: Chambers JM, Hastie TJ (eds) Generalized linear models. Wadsworth & Brooks/Cole, Pacific Grove, pp 195–248

Hatano Y, Fukuda S, Hisamatsu K, Hirata A, Hara A, Tomita H (2017) Multifaceted interpretation of colon cancer stem cells. Int J Mol Sci 18:1446

Heagerty PJ, Lumley T, Pepe MS (2000) Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 56:337–344

Hosseini H, Obradović MMS, Hoffmann M, Harper KL, Sosa MS, Werner-Klein M, Nanduri LK, Werno C, Ehrl C, Maneck M, Patwary N, Haunschild G, Gužvić M, Reimelt C, Grauvogl M, Eichner N, Weber F, Hartkopf AD, Taran FA, Brucker SY, Fehm T, Rack B, Buchholz S, Spang R, Meister G, Aguirre-Ghiso JA, Klein CA (2016) Early dissemination seeds metastasis in breast cancer. Nature 540:552–558

Jarvis RA, Patrick EA (2006) Clustering using a similarity measure based on shared near neighbors. IEEE Trans Comput 22:1025–1034

Lee HO, Hong Y, Etlioglu HE, Cho YB, Pomella V, Van den Bosch B, Vanhecke J, Verbandt S, Hong H, Min JW, Kim N, Eum HH, Qian J, Boeckx B, Lambrechts D, Tsantoulis P, De Hertogh G, Chung W, Lee T, An M, Shin HT, Joung JG, Jung MH, Ko G, Wirapati P, Kim SH, Kim HC, Yun SH, Tan IBH, Ranjan B, Lee WY, Kim TY, Choi JK, Kim YJ, Prabhakar S, Tejpar S, Park WY (2020) Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. Nat Genet 52:594–603

Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJL, Kong SL, Chua C, Hon LK, Tan WS, Wong M, Choi PJ, Wee LJK, Hillmer AM, Tan IB, Robson P, Prabhakar S (2017) Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat Genet 49:708–718

Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P (2015) The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 1:417–425

Linde N, Casanova-Acebes M, Sosa MS, Mortha A, Rahman A, Farias E, Harper K, Tardio E, Reyes Torres I, Jones J, Condeelis J, Merad M, Aguirre-Ghiso JA (2018) Macrophages orchestrate breast cancer early dissemination and metastasis. Nat Commun 9:21

Liu WT, Jing YY, Yu GF, Han ZP, Yu DD, Fan QM, Ye F, Li R, Gao L, Zhao QD, Wu MC, Wei LX (2015) Toll like receptor 4 facilitates invasion and migration as a cancer stem cell marker in hepatocellular carcinoma. Cancer Lett 358:136–143

Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, Menon M, He L, Abdurrob F, Jiang X, Martorell AJ, Ransohoff RM, Hafler BP, Bennett DA, Kellis M, Tsai LH (2019) Single-cell transcriptomic analysis of Alzheimer's disease. Nature 570:332–337

Mittal V (2018) Epithelial mesenchymal transition in tumor metastasis. Annu Rev Pathol 13:395–412

Pang B, Xu J, Hu J, Guo F, Wan L, Cheng M, Pang L (2019) Single-cell RNA-seq reveals the invasive trajectory and molecular cascades underlying glioblastoma progression. Mol Oncol 13:2588–2603

Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL, Louis DN, Rozenblatt-Rosen O, Suvà ML, Regev A, Bernstein BE (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 344:1396–1401

Qi, M, Li W, Goodison S, Sun Y (2015) Dimensionality reduction via graph structure learning. In: The 21th ACM SIGKDD International Conference.

Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C (2017) Reversed graph embedding resolves complex single-cell trajectories. Nat Methods 14:979–982

Sakamoto Y, Ishiguro M, Kitagawa G (1986) Akaike Information criterion statistics. D. Reidel Publishing Company, Dordrecht

Satija R, Farrell JA, Gennert D, Schier AF, Regev A (2015) Spatial reconstruction of single-cell gene expression data. Nat Biotechnol 33:495–502

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504

Shin J, Berg DA, Zhu Y, Shin JY, Song J, Bonaguidi MA, Enikolopov G, Nauen DW, Christian KM, Ming GL, Song H (2015) Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. Cell Stem Cell 17:360–372

Sistigu A, Musella M, Galassi C, Vitale I, De Maria R (2020) Tuning cancer fate: tumor microenvironment's role in cancer stem cell quiescence and reawakening. Front Immunol 11:2166

Smith JJ, Deane NG, Wu F, Merchant NB, Zhang B, Jiang A, Lu P, Johnson JC, Schmidt C, Bailey CE, Eschrich S, Kis C, Levy S, Washington MK, Heslin MJ, Coffey RJ, Yeatman TJ, Shyr Y, Beauchamp RD (2010) Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. Gastroenterology 38:958–968

Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F (2020) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 71:209–249

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA (2009) mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6:377–382

Wang H, Gong P, Chen T, Gao S, Wu Z, Wang X, Li J, Marjani SL, Costa J, Weissman SM, Qi F, Pan X, Liu L (2021) Colorectal cancer stem cell states uncovered by simultaneous single-cell analysis of transcriptome and telomeres. Adv Sci (Weinh) 8:2004320

Welch LR (2003) Hidden Markov Models and the Baum-Welch Algorithm. IEEE Inf Theory Soc Newslett 53:194–211

Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, Fu X, Liu S, Bo X, Yu G (2021) clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation (N Y) 2:100141

Ye Y, Li SL, Wang SY (2018) Construction and analysis of mRNA, miRNA, lncRNA, and TF regulatory networks reveal the key genes associated with prostate cancer. PLoS One 13:e0198055

Yilmaz M, Christofori G (2010) Mechanisms of motility in metastasizing cells. Mol Can Res 8:629–642

Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, Luo T, Xu L, Liao G, Yan M, Ping Y, Li F, Shi A, Bai J, Zhao T, Li X, Xiao Y (2019) Cell Marker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res 47:D721–D728

Zhang J, Guan M, Wang Q, Zhang J, Zhou T, Sun X (2020) Single-cell transcriptome-based multilayer network biomarker for predicting prognosis and therapeutic response of gliomas. Brief Bioinform 21:1080–1097