OncoTargets and Therapy

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Investigating the Cell Origin and Liver Metastasis Factors of Colorectal Cancer by Single-Cell Transcriptome Analysis

Zhilin Sha[1],*, Qingxiang Gao[1],*, Lei Wang[2],*, Ni An[3], Yingjun Wu[1], Dong Wei[4], Tong Wang[5], Chen Liu[1], Yang Shen[1]

[1]Department 1 of Biliary Tract Surgery, Eastern Hepatobiliary Surgery Hospital, Naval Medical University, Shanghai, People's Republic of China; [2]Department of General Surgery, Yancheng Hospital of Traditional Chinese Medicine, Yancheng, Jiang Su, People's Republic of China; [3]Department of Anesthesiology, the Eighth Medical Center of Chinese PLA General Hospital, Beijing, People's Republic of China; [4]Department of General Surgery (Second Ward), the No.1 People's Hospital of Pinghu, Pinghu, Zhe Jiang, People's Republic of China; [5]Department of Anesthesiology, No.32295 Troop of Chinese PLA, Liaoyang, People's Republic of China

*These authors contributed equally to this work

Correspondence: Yang Shen, Eastern Hepatobiliary Surgery Hospital, Naval Medical University, No. 700 of North Moyu Road, Jia Ding District, Shanghai, 201823, People's Republic of China, Email drbruceshen@163.com; Chen Liu, Eastern Hepatobiliary Surgery Hospital, Naval Medical University, No. 700 of North Moyu Road, Jia Ding District, Shanghai, 201823, People's Republic of China, Email dinoliu@126.com

Background: Colorectal cancer (CRC) is one of the deadliest causes of death by cancer worldwide. Liver metastasis (LM) is the main cause of death in patients with CRC. Therefore, identification of patients with the greatest risk of liver metastasis is critical for early treatment and reduces the mortality of patients with colorectal cancer liver metastases.

Methods: Initially, we characterized cell composition through single-cell transcriptome analysis. Subsequently, we employed copy number variation (CNV) and pseudotime analysis to delineate the cellular origins of LM and identify LM-related epithelial cells (LMECs). The LM-index was constructed using machine learning algorithms to forecast the relative abundance of LMECs, reflecting the risk of LM. Furthermore, we analyzed drug sensitivity and drug targeted gene expression in LMECs and patients with a high risk of LM. Finally, functional experiments were conducted to determine the biological roles of metastasis-related gene in vitro.

Results: Single-cell RNA sequencing analysis revealed different immune landscapes between primary CRC and LM tumor. LM originated from chromosomal variants with copy number loss of chr1 and chr6p and copy number gain of chr7 and chr20q. We identified the LMECs cluster and found LM-associated pathways such as Wnt/beta-catenin signaling and KRAS signaling. Subsequently, we identified ten metastasis-associated genes, including SOX4, and established the LM-index, which correlates with poorer prognosis, higher stage, and advanced age. Furthermore, we screened two drugs as potential candidates for treating LM, including Linsitinib_1510, Lapatinib_1558. Immunohistochemistry results demonstrated significantly elevated SOX4 expression in tumor samples compared to normal samples. Finally, in vitro experiments verified that silencing SOX4 significantly inhibited tumor cell migration and invasion.

Conclusion: This study reveals the possible cellular origin and driving factors of LM in CRC at the single cell level, and provides a reference for early detection of CRC patients with a high risk of LM.

Keywords: colorectal cancer, liver metastasis, single-cell sequencing, SOX4, prognostic

Introduction

Colorectal cancer (CRC) is one of the most common malignant tumors in the digestive tract, with the second highest incidence and mortality rate of cancer.[1,2] Early clinical symptoms are not obvious, but symptoms such as blood in stool, constipation and abdominal pain appear with increasing tumor extent, which seriously affects the quality of life of patients.[3,4] Since CRC is highly heterogeneous, clinical indicators and prognosis are quite different, especially in the stage II and III stages of CRC.[5] Currently, the preferred treatments are radiotherapy and surgical resection of the lesion site.[6] However, the rapid disease

Received: 18 January 2024
Accepted: 6 April 2024
Published: 17 April 2024

progression of CRC, with the majority of patients presenting with liver metastasis, makes it particularly important to assess risk stratification and identify patients with early metastases.[7]

The liver is the main target organ of CRC metastasis. About a quarter of patients with CRC have the symptom of liver metastasis at the time of diagnosis, and more patients get liver metastasis after surgical eradication of the primary tumor.[8] Nearly 90% of patients cannot be completely eradicated after the occurrence of liver metastasis, which is also one of the most important causes of death in CRC.[9] According to the Chinese guidelines for the diagnosis and comprehensive treatment of colorectal liver metastasis (2023), the 5-year survival rate of patients with unresectable liver metastasis is extremely low.[10] However, the 5-year survival rate of patients with complete resection of liver metastasis increased to more than 30%. Therefore, research for molecular features and molecular biomarkers for the prediction of tumor metastasis has attracted more attention and has been extremely required. In recent years, many predictors focused on the composition and origin of cells in the tumor immune microenvironment (TIME) and their influence on the site of metastasis.[11] Studies indicated the important role of different TIMEs caused by different cells, which is highly related to malignant tumor progression and immunotherapy.[12] However, the TIMEs from a liver metastasis perspective and ideal metastasis-related predictors have been less studied with comprehensive analysis and convinced experimental validation.

Our study focused on liver metastasis from colorectal cancer. By analyzing single-cell transcriptome data from primary and liver metastasis (LM) samples, we discerned disparities in TIME and identified a malignant epithelial subpopulation termed liver metastasis related epithelial cells (LMECs). Based on single-cell RNA sequencing (scRNA-seq) data and liver metastasis related epithelial cells abundance in bulk RNA data, LM-index represented liver metastasis risk was constructed and validated. Additionally, by GSVA enrichment analysis we found the top activated pathways, including Wnt/beta-catenin signaling pathways, G2M checkpoint, and KRAS signaling in LMECs. The expression difference of marker gene SOX4 was confirmed by IHC. The migration and invasion ability influenced by SOX4 also been verified by in vitro experiments. In brief, this study provides essential insights for the early identification of high-risk patients, with liver metastasis resulting from colorectal cancer.

Materials and Methods

CRC Primary and LM Data Acquisition

The scRNA-seq matrix data of CRC and LM were collected from GEO, NCBI, and Stanford (GSE178318, PRJNA748525 and https://dna-discovery.stanford.edu/research/datasets/). Based on the paired data of the primary tumor site and LM, three paired samples were extracted from GSE178318. Thirteen primary tumor samples were extracted from PRJNA748525, and seven LM samples were extracted from Stanford dataset, resulting in a total of 16 CRC in primary samples and 10 LM samples for scRNA data (Table S1). The bulk mRNA of colorectal cancer (CRC) sample data was obtained from the TCGA official website (https://portal.gdc.cancer.gov/), including colon adenocarcinoma (TCGA_COAD) and rectum adenocarcinoma (TCGA_READ) with a total of 650 tumor samples. After the quality control of bulk mRNA data, 647 colorectal tumors were finally retained with complete prognostic information (survival time and survival status) (Table S2). To validate the protein expression of SOX4, immunohistochemistry (IHC) data was obtained from the HPA database (The Human Protein Atlas, https://www.proteinatlas.org). The study was approved by the ethics committee of Eastern Hepatobiliary Surgery Hospital, Naval Medical University.

Analysis and Cluster Annotation of scRNA-Seq Data

For primary and metastatic tumor samples, we analyzed single-cell data using Seurat v4.1.1, filtering out cells with more than 15% mitochondrial content, more than 5% hemoglobin content, and less than 200 or more than 6000 gene expression. Data normalization, dimensionality reduction and cell clustering were performed using the Seurat R package. The FindVariableFeatures function was used to screen out 2000 highly variable genes from the filtered expression matrices, and then principal component analysis was performed using the RunPCA function, retaining the top 20 principal components for further analysis. Batch effects were removed with RunHarmony of R package harmony. The FindClusters function was employed for cell clustering, and the RunUMAP function was used for nonlinear dimensionality reduction with 0.3 resolution. Cell clusters were manually annotated based on the CellMarker 2.0 database and collected cell-specific markers.

Immune subgroup analysis extracted and merged all cells annotated as immune types and removed batch effects again by using RunHarmony of R package harmony. Cell clustering has been performed with FindClusters and nonlinear dimensionality reduction using the RunUMAP function in Seurat.

Gene Set Variation Analysis

HALLMARK gene set was downloaded from MSigDB (https://www.gsea-msigdb.org, v2023.1), GSVA for gene expression enrichment analysis in primary and metastasis tumor in single cell level using the R package GSVA.[13] Intergroup difference analysis was performed using the limma package, and significantly enriched pathways were screened with the cutoff as $|t|>4$, padj < 0.05, which indicates the greatest differences between the two groups for each cell type.[13]

Copy Number Variation Analysis

Based on scRNA expression data in chromosome ordering, we used the R package inferCNV to detect copy number variation (CNV) and selected differentiate malignant epithelial cells from non-malignant epithelial cells by calculating the CNV score with HMM set as TRUE. After the CNV score sorting, the top 5 percentage observation cells with high CNV scores were employed as malignant baseline cells. Correlations between reference cells, observation cells and malignant baseline cells were calculated, respectively. Malignant epithelial cells were picked out from the observation, whose correlation was higher than 95% of reference.[14,15] Uphyloplot was used to visualize the CNV evolutionary branching diagram in the phylogenetic tree diagram.[16]

Pseudotime of Malignant Epithelial Cells

Pseudotime trajectory analysis based on gene expression profiles of malignant epithelial cells was reported using the default parameters of the R package monocle 2.[17]

LM-Index of Bulk mRNA Data

Malignant cell expression profiles of risk genes were fitted using gelnet (v1.2.1, https://CRAN.R-project.org/package=gelnet) from the R package. The LM-index was generated through the one-class logistic regression (OCLR) algorithm.[18] Subsequently, we calculated the Spearman correlation between the weight vectors of risk genes and the mRNA expression of samples. Finally, using a linear transformation, the LM-index was mapped to a range of 0 to 1 by subtracting the minimum value and dividing by the maximum Spearman correlation coefficient value. The complete.obs was selected for missing values, and rows containing missing values were ignored. LM-index was computed for each TCGA sample for scoring. In addition, we further estimated the relationships among the LM-index, clinical characteristics, and survival information.

Identification of Therapeutic Agents for Patients with LMECs and High LM Index

Using the R package oncoPredict, chemotherapy drug sensitivity was predicted according to the GDSC2 database. Predicted half inhibitory concentration (IC50) values were calculated. Different drug responses were detected between LMECs and other malignant epithelial cells and between high and low LM index samples in scRNA data and bulk mRNA data, respectively. The rank sum test was used to compare the predicted IC50 differences (log2fc > 1, pval < 0.05), and the correlation between IC50 and LM-index was analyzed using Spearman correlation analysis (cor < -0.05, pval < 0.05).

Cell Culture and qRT-PCR Analysis

The human colon cancer cell line HCT116 (BNCC287750, Bena Culture Collection, China) and the normal intestinal epithelial cells NCM-460 (BNCC339288, Bena Culture Collection, China) were cultured in DMEM (Gibco, Grand Island, NY, USA) containing 1% penicillin/streptomycin and 10% fetal bovine serum (Gibco). The cells were placed in a humidified incubator containing 5% CO2 maintained at 37 ° C.

Total RNA was extracted from the cells using TRIzol universal total RNA extraction reagent (Invitrogen, Carlsbad, CA, USA). The quality and concentration of the extracted RNA were assessed using a UV spectrophotometer. Once qualified, reverse transcription was done using Transcriptor First Strand cDNA Synthesis Kit (GenStar, Beijing, China).

Subsequently, qPCR assay was conducted employing LightCycler 480 Fluorescence Quantitative System (Roche, Basel, Switzerland). The reference gene was β-actin, and the primer sequences were listed in Table S3. The mRNA expression levels were calculated according to the $2^{-\Delta\Delta CT}$ method (three repeats).

Cell Transfection

SOX4 knockdown was generated using small interfering RNAs (siRNAs). The SOX4 siRNA sequences were included in Table S4. Briefly, cells were seeded at 50% confluence in a 6-well plate and infected with negative control (NC), and knockdown (si-SOX4). All transfections were carried out with Lipofectamine 3000 (Invitrogen, Carlsbad, CA, USA).

Scratch Assay

A total of 7×10^5 cells were seeded in each well of a 6-well plate for 24 hours. A line was drawn in the middle of the well with a 10 µL pipette tip. After washing with phosphate buffered saline (PBS) twice, cells were cultured for 24 hours in a 37°C incubator. Then, wounds were photographed by microscope at different time intervals. The distances of the wounds were measured by ImageJ.

Invasion Assays

The invasion capacities of si-control and si-SOX4 cells were analyzed by polycarbonate membranes (8 µm pore) in 24-well transwell chambers (Corning, NY, USA). About 1×10^4 cells in serum-free medium containing 0.1% BSA were added to the upper chamber. For invasion assay, transwell chambers were coated with prediluted extracellular matrix (3 mg/mL, Merck, Darmstadt, Germany) for 1 hour before adding cells to the upper chamber. The medium supplemented with 0.1% BSA and EGF (50 ng/mL, MCE, NJ, USA) was added into the down chamber. After 24 hours incubation, cells in the upper chamber were completely scraped and trans to the lower membrane. The polycarbonate membranes were fixed and stained with Giemsa solution (Solarbio, Beijing, China) and photographed by microscope.

Statistical Analysis

Wilcoxon tests were employed for two-group comparisons, and Kruskal–Wallis (K-W) tests were employed for multiple group comparisons. The Kaplan–Meier (K-M) method was applied to compare the survival curves of various groups, and the Log rank test was utilized for comparing the survival curves. Spearman correlation analysis was conducted to assess the correlations. P value <0.05 was taken as statistically significant.

Results

Evaluation of Cellular Component and Tumor Immune Environments

We conducted scRNA-seq analyses to investigate the TIME, metastasis origins and related factors. TCGA bulk RNA-seq was used to validate the performance of LM-index. Furthermore, potential therapy drugs were screened. Finally, in vitro experiments verified the ability of marker gene for inhibiting tumor cell migration and invasion.

After performing quality control and data filtering on the obtained single-cell data, a total of 104,484 primary CRC tumor (pCRC) cells and 31,270 LM cells were obtained. Then these cells were then clustered into 24 and 20 clusters based on data normalization, high-variable gene selection, principal component analysis, dimensionality reduction, and UMAP clustering (Figure S1). Subsequently, the clusters in the pCRC and LM were manually annotated for known cell types, respectively (Figure 1A and B). For pCRC clusters, mast cell, fibroblast cell, epithelial cell, dendritic cell, monocyte cell, macrophage cell, neutrophil cell, CD4 T cell, CD8 T cell, plasma cell, and B cell were identified by canonical markers (Figure 1C). Meanwhile, LM clusters contained hepatocyte cell, fibroblast cell, epithelial cell, endothelial cell, dendritic cell, monocyte cell, macrophage cell, CD4 T cell, CD8 T cell, plasma cell, and B cell (Figure 1D). The proportion of each cell types showed different tumor immune environment between pCRC and LM (Figure 1E–G). Among them, CD4 T cell and CD8 T cell, as well as the whole immune cells, occupied a considerable proportion. In addition, epithelial cells had the third higher proportions in both pCRC and LM groups. Similarly, dendritic cell, monocyte cell, and macrophage cell had a comparable proportion. However, the abundances of plasma cell, B cell, and fibroblast cell were higher in pCRC, whose proportion was even higher than that in

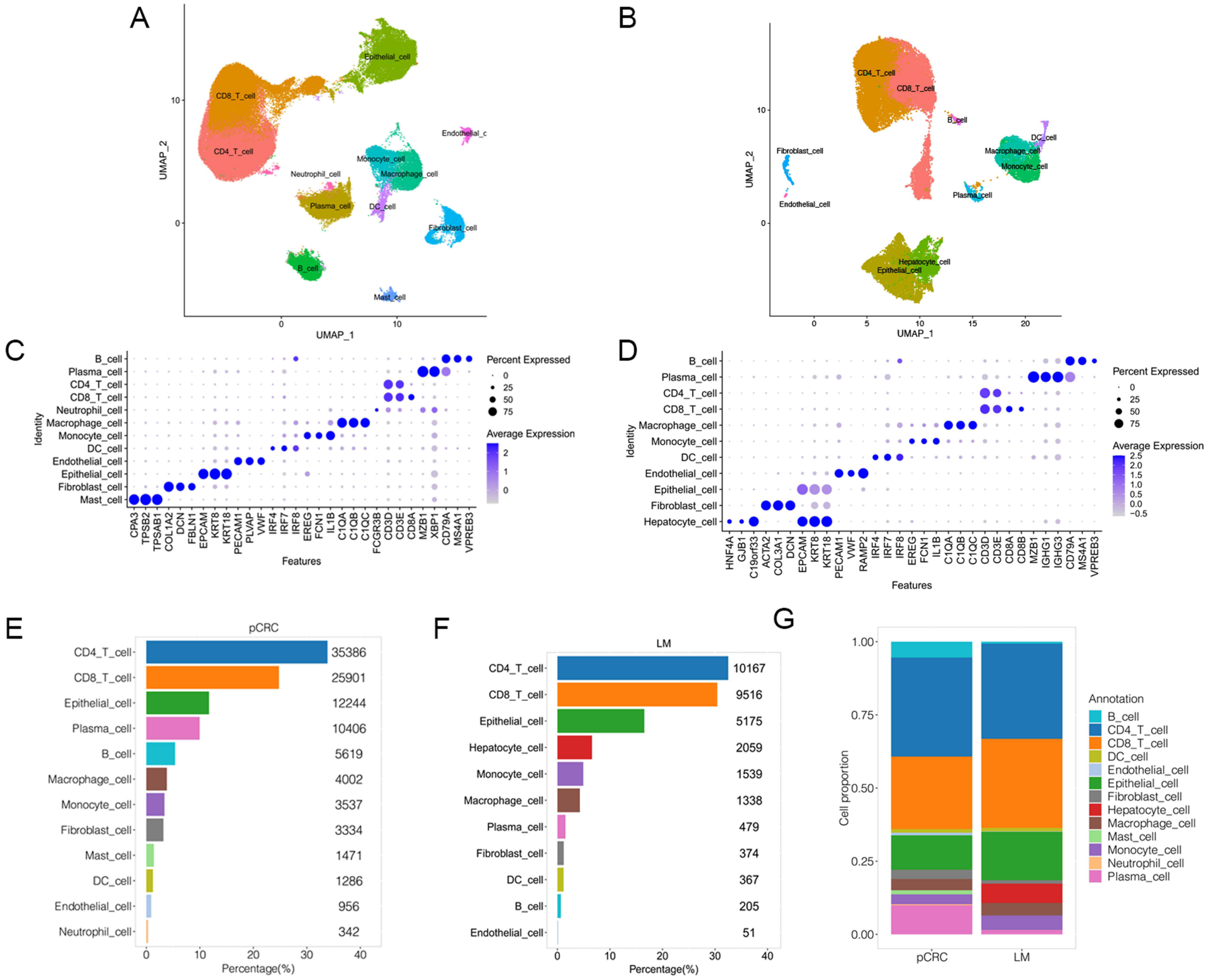

Figure 1 Different cellular composition of pCRC and LM cells. (**A**) Cellular composition in pCRC. (**B**) Cellular composition in LM. (**C**) Dotplot of cell markers expression in pCRC. (**D**) Dotplot of cell markers expression in LM. (**E**) Proportion of each cell types in pCRC. (**F**) Proportion of each cell types in LM. (**G**) Comparison of the cell proportion of pCRC and LM.

myeloid cells. Mast cell and neutrophil cell were only detected in pCRC group, while hepatocyte cell was only annotated in the LM group.

To elucidate the specific cell components of the immune cells in pCRC and LM, we extracted and reclustered all the immune cells in both groups. Then, nineteen distinct clusters were identified and grouped into eleven subtypes (Figure 2A and B). Among them, CD4 T cell, CD8 T cell, and naive T cell were the top three most abundant immune cell types and were distributed in both pCRC and LM (Figure 2C). Then GSVA enrichment analysis was performed in the above three types of immune cells, and more cancer progression related hallmark pathways were activated in the LM group. (Figure 2D–F). Furthermore, the GSVA was also performed on other immune cells, and the result shown in Figure S2.

Identification of Malignant Cells and Liver Metastasis Origins

To analyze the clonal structure and LM origin of malignant cells, a total of 17,419 epithelial cells, including 12,244 in pCRC and 5175 in LM, were scored as observation by inferCNV (Figure 3A and B). CD4 T cell and CD8 T cell were employed as reference. After the CNV score sorting, the top 5 percentage observation cells with high CNV scores were employed as malignant baseline cells. Correlations between reference cells, observation cells, and malignant baseline cells were calculated, respectively. Malignant epithelial cells were picked out from the observation, whose correlation

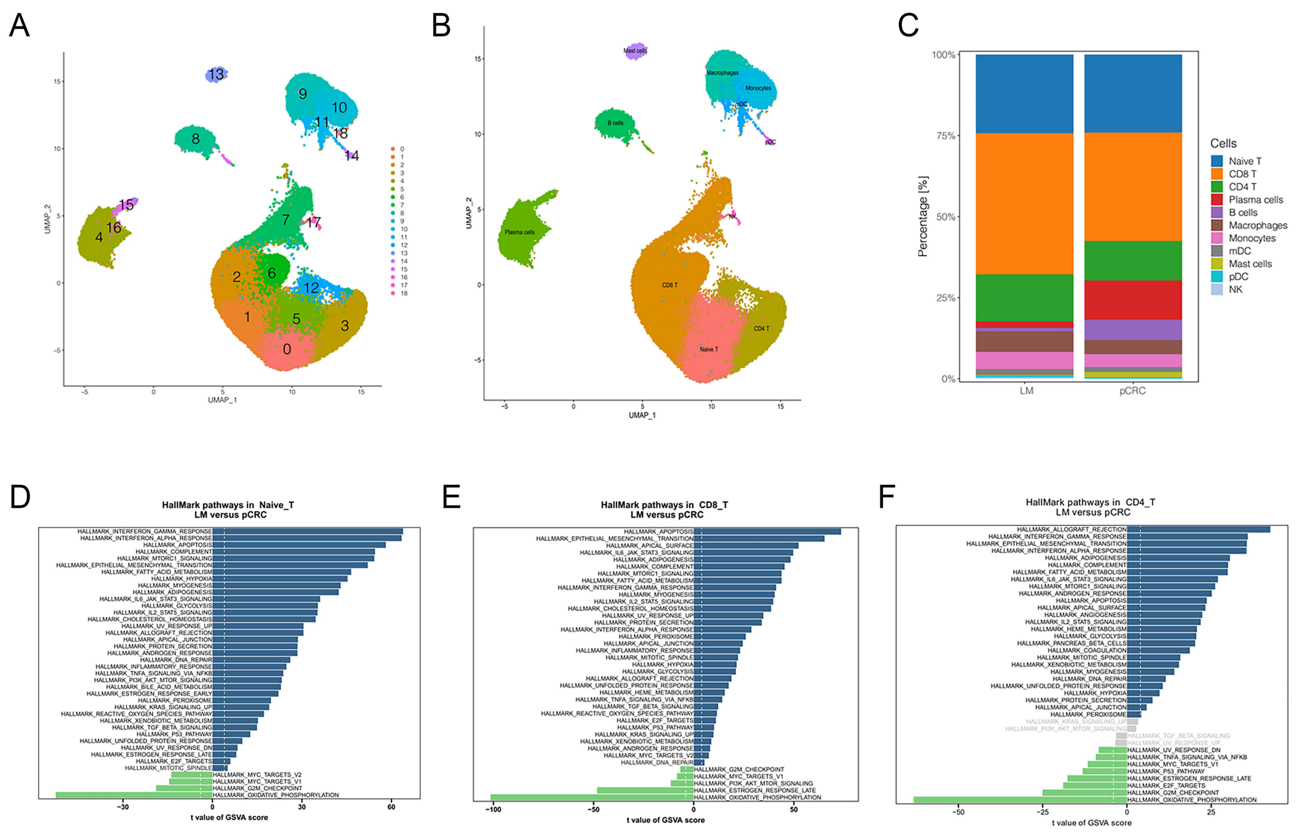

Figure 2 Cellular composition of immune cell subpopulation. (**A**) UMAP cluster diagram of immune cells. (**B**) Annotation of immune cells divided into subsets. (**C**) Proportion of each cell types in immune cells. (**D**)Native T cell LM vs pCRC GSVA enrichment. (**E**)CD8 T cell LM vs pCRC GSVA enrichment. (**F**)CD4 T cell LM vs pCRC GSVA enrichment.

was higher than 95% of reference. Finally, in pCRC, 9128 malignant cells were identified, while 3252 malignant cells were detected in LM (Figure 3C).

Based on the CNV score, evolutionary phylogenetic tree was used to analyze the evolutionary relationship between pCRC and LM. We expected to filter out similar CNV events in the root of pCRC and LM to find the malignant subclone, which may be highly in connection with LM. Notably, loss of chr1 and chr6p and gain of chr7 and chr20q were observed in malignant cells from the root of LM, which also be observed in the root of pCRC (Figure 3D). These results suggest that pCRC subclones with loss of chr1 and chr6p and gain of chr7 and chr20q may be the origin of LM. In addition, we extracted all genes with CNV events involved in the chromosome described above in all branched genes in pCRC and LM. The upset plot shows 768 genes in pCRC with the aforementioned chromosome were shared by all subclones, while 330 genes were shared in LM (Figure 3E and F). The genes from both groups were intersected, and 275 common genes were obtained (Figure 3G).

Furthermore, GSVA analysis was performed to compare the pathway enrichment difference between malignant cells in pCRC and LM. G2M checkpoint, E2F targets, and KRAS signaling were the top 3 activated pathways in pCRC, and complement, IFN gamma response and estrogen response early were the top 3 activated pathways in LM (Figure 3H).

Identification of Liver Metastasis Related Malignant Epithelial Cell

To further elucidate malignant epithelial cell subsets and their role in LM, 12,380 malignant epithelial cells from pCRC and LM were reclustered. Malignant epithelial cells showed heterogeneity, resulting in a total of 19 subclusters (Figure 4A). Among these subclusters, cluster 0, cluster 2, cluster 3, cluster 4, and cluster 8 contained malignant epithelial cells in both pCRC and LM groups (Figure 4B), which were defined as LM-related malignant epithelial cells (LMECs). To further explore whether there is a more obvious relationship between pCRC and LM, pseudotime analysis was used to generate a pseudotime

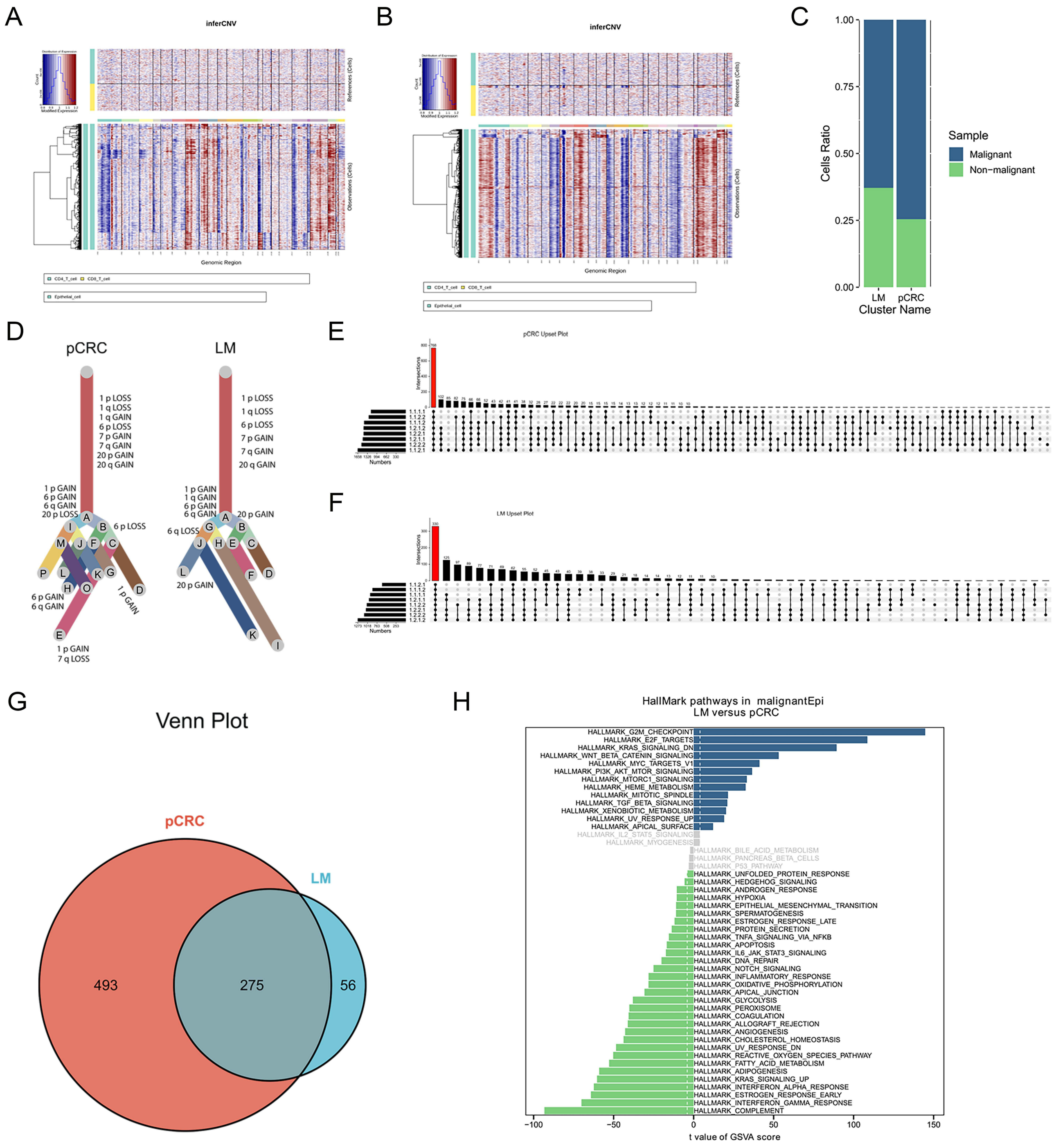

Figure 3 CNV and clonality analysis of malignant cells. (**A**) InferCNV scoring plot of pCRC epithelial cells. (**B**) InferCNV scoring plot of LM epithelial cells. (**C**) Malignant cell ratio plot. (**D**) InferCNV clones evolutionary trees. (**E**) Upset plot of genes in metastasis related chromosome in pCRC. (**F**) Upset plot of genes in metastasis related chromosome in LM. (**G**) Venn plot of shared genes. (**H**) GSVA enrichment plot of malignant cells.

trajectory based on the above LMECs (Figure 4C and D). Branch point 1 showed differentiation of primary and metastatic malignant epithelial cells, in which metastatic malignant epithelial cells entered cell fate 1 and primary malignant epithelial cells entered cell fate 2. Subsequently, further branch-related gene expression analysis was performed to identify the branch key genes that determine cell fate. We identified 50 genes at branch point 1 that regulate cell differentiation (Figure 4E). Meanwhile, further GSVA enrichment analysis showed that Wnt/beta-catenin signaling pathways, G2M checkpoint, and KRAS signaling were the top 3 activated pathways in LMECs based on LM versus pCRC (Figure 4F).

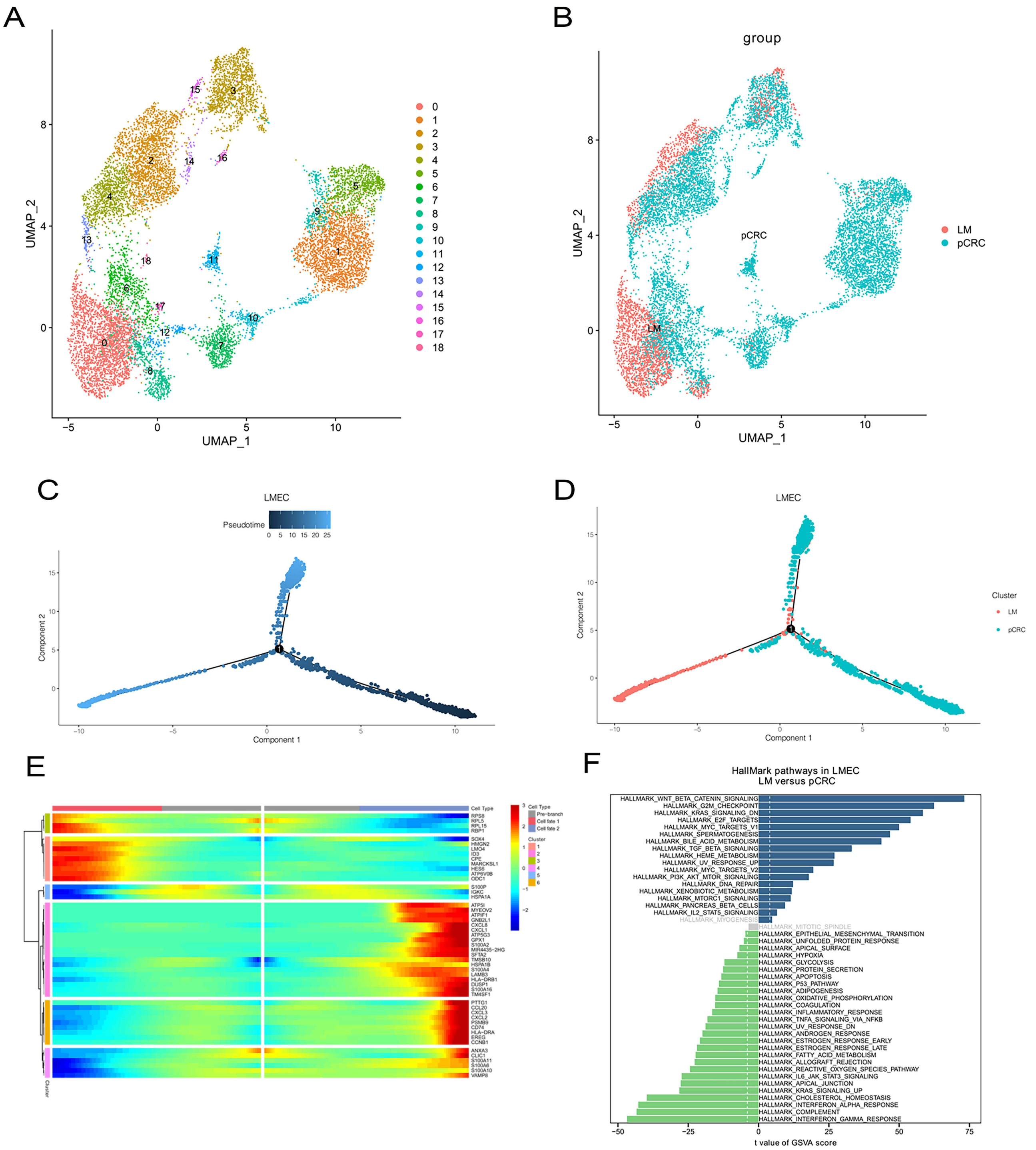

Figure 4 Identification of LMECs by pseudotime analysis. (**A**) UMAP diagram of malignant epithelial cells in cluster. (**B**) UMAP diagram of malignant epithelial cells in group. (**C-D**) Pseudotime diagram of shared subgroups of in pCRC and LM. (**E**) Heatmap of branch related genes. (**F**) LM vs pCRC GSVA enrichment of LMECs.

Construction and Validation of LM-Index with Metastasis-Associated Factors

After scRNA clustering in the first place, 1798 and 1785 differentially expressed genes of epithelial cells were detected in pCRC and LM, respectively. And 859 genes with the same direction were obtained after the intersection of the two parts. Then, a total of 271 genes were obtained after joining LMECs differentially expressed genes. Further, the list of the above genes was introduced into 50 branch-related genes in the pseudotime analysis. Finally, 10 metastasis-associated factors, SOX4, MARCKSL1, HES6, RPS8, DUSP1, HLA-DRB1, HLA-DRA, CD74, HSPA1A and HSPA1B, were obtained (Figure 5A). Among them, SOX4 was widely studied in the migration and invasion,[19–21] and finally used as a biomarker for LMECs (Table S5).

Furthermore, R package gelnet was used to construct the LM-index through machine learning algorithms to predict the relative abundance of the LMECs, which represents the risk of LM. Then the LM-index was used to calculate the abundance of LMECs on TCGA bulk mRNA samples (Figure 5B), and the samples were divided into the high LM-index and low LM-index groups. Notable, the high LM-index group had a significantly worse prognosis ($P < 0.05$, Figure 5C, Figure S3). In addition, we found that the LM-index had significant differences in age, node, and metastasis in TNM (Figure 5D), which indicated that LMECs abundance was positively correlated with TNM stage and negatively correlated with survival time in CRC patients.

Drug Identification Analysis

Based on the GDSC2 database, we further screened candidate targeted therapy drugs for inhibiting LM of colorectal cancer. Highly sensitive drugs targeting LMECs and patients with high LM-index were detected, respectively. One hundred and twenty-eight GDSC2 compounds for targeting LMECs were obtained, as well as 4 GDSC2 compounds for patients with high LM-index. Two drugs from the GDSC2 database (Linsitinib_1510, Lapatinib_1558) were finally

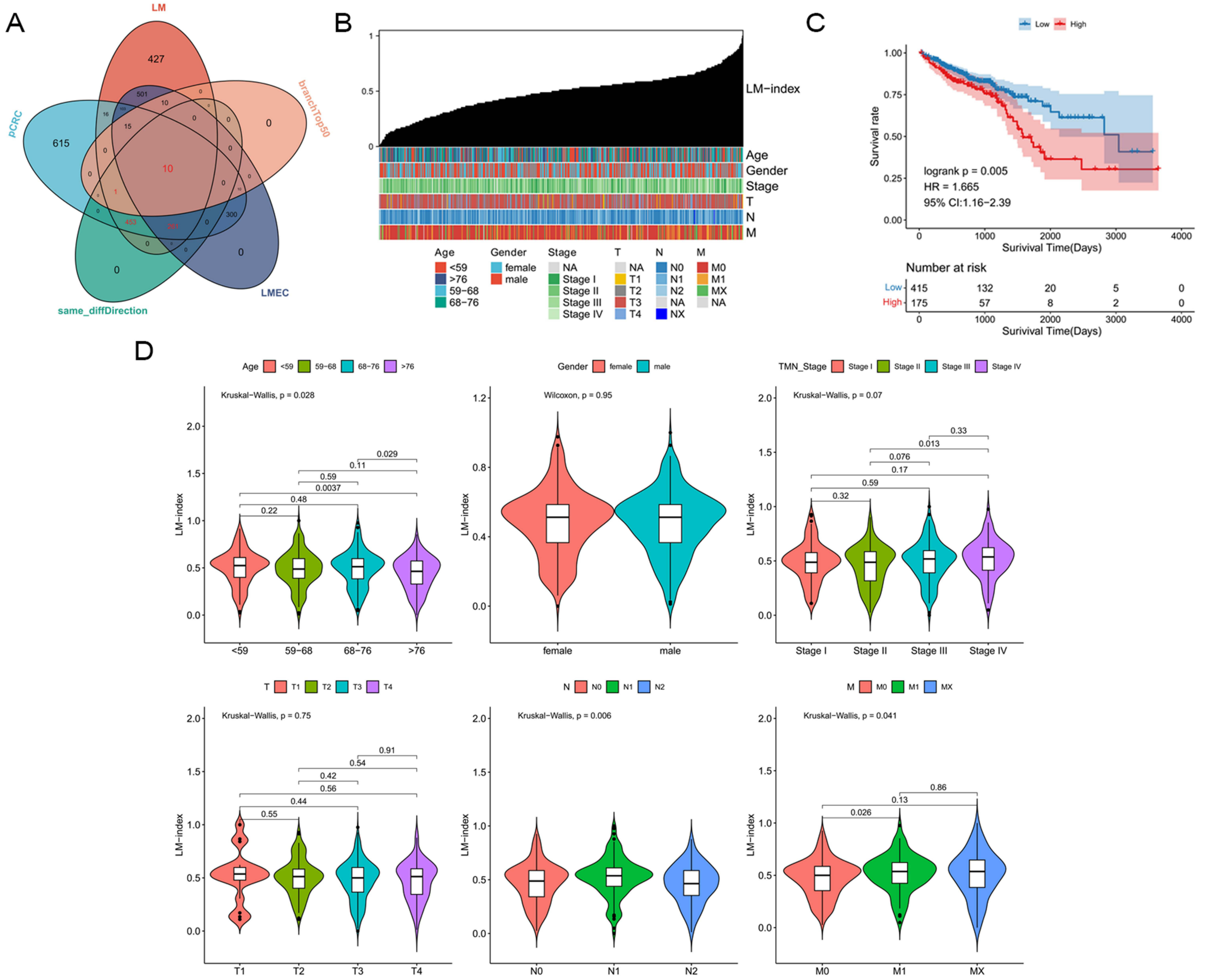

Figure 5 LM-index with metastasis-associated factors and clinical indicators. (**A**)Venn diagram showing signature gene selection. (**B**) Risk gene LM-index plotted against different clinical indicators. (**C**) Survival curves of LM-index corresponding patients. (**D**) Difference analysis between clinical indicators and LM-index.

identified as candidates for the treatment of LM by intersecting two parts of compounds with a negative correlation between predicted IC50 and LM-index (Figure 6A–C, Figure S4). Among them, targeted genes were obtained from the Drugbank database, and the gene expression in scRNA data and TCGA bulk mRNA data were analyzed. The result showed that targeted gene expression was higher in LMECs and high LM-index groups (Figure 6D–E).

External Experimental Validation

To explore the SOX4 potential influence in LM of CRC, we first used qRT-PCR to evaluate its expression in human colon cancer cell line HCT116 and the normal intestinal epithelial cells NCM-460. The expression level of SOX4 was significantly increased in the cancer cell line compared to normal cells (Figure 7A). In addition, IHC staining showed that the expression of SOX4 in tumor samples was higher than that in normal samples (Figure 7B).

Then, HCT-116 cells were transfected by the interference RNA and control. The silencing of SOX4 was confirmed by RT-qPCR in the cancer cell line (Figure 7C). After that, the scratch assay indicated that the knockdown of SOX4 significantly suppressed the HCT-116 cells' migration capacity compared to control cells (Figure 7D). Finally, the results of transwell invasion assays suggested that the invasion ability of HCT-116 cells was significantly inhibited by si-SOX4 (Figure 7E). So, the in vitro experiments indicated that si-SOX4 could inhibit the cell migration and invasion ability of CRC, which might be involved in the metastasis process.

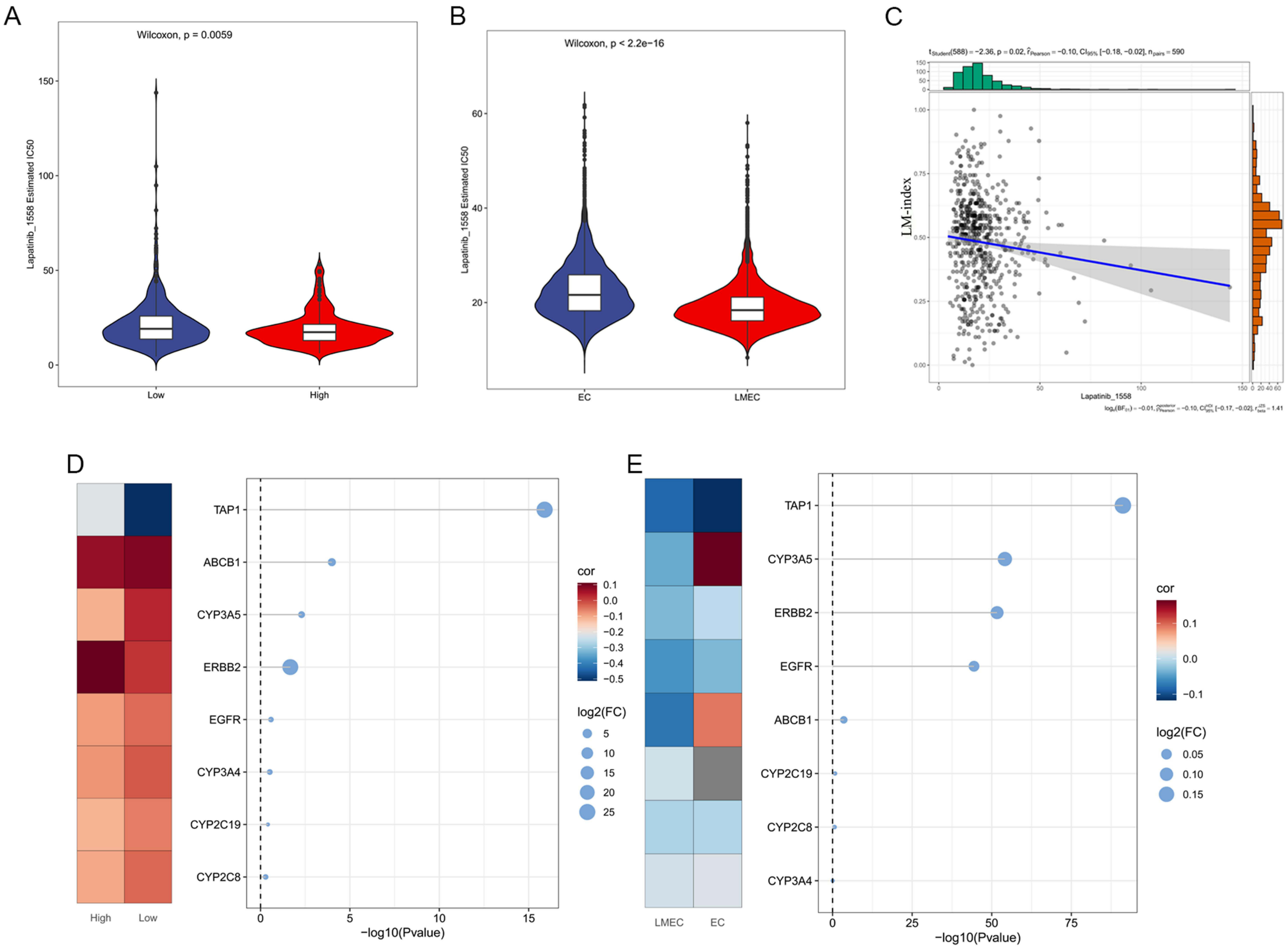

Figure 6 Drug characterization and analysis. (**A**) Violin plot of predicted IC50 in Lapatinib_1558 with TCGA bulk mRNA data. (**B**) Violin plot of predicted IC50 in Lapatinib_1558 with scRNA data. (**C**) Spearman correlation analysis of lapatinib between predicted IC50 and LM-index. (**D**) Expression of lapatinib targeted gene in TCGA bulk mRNA data. (**E**) Expression of lapatinib targeted gene in scRNA data.

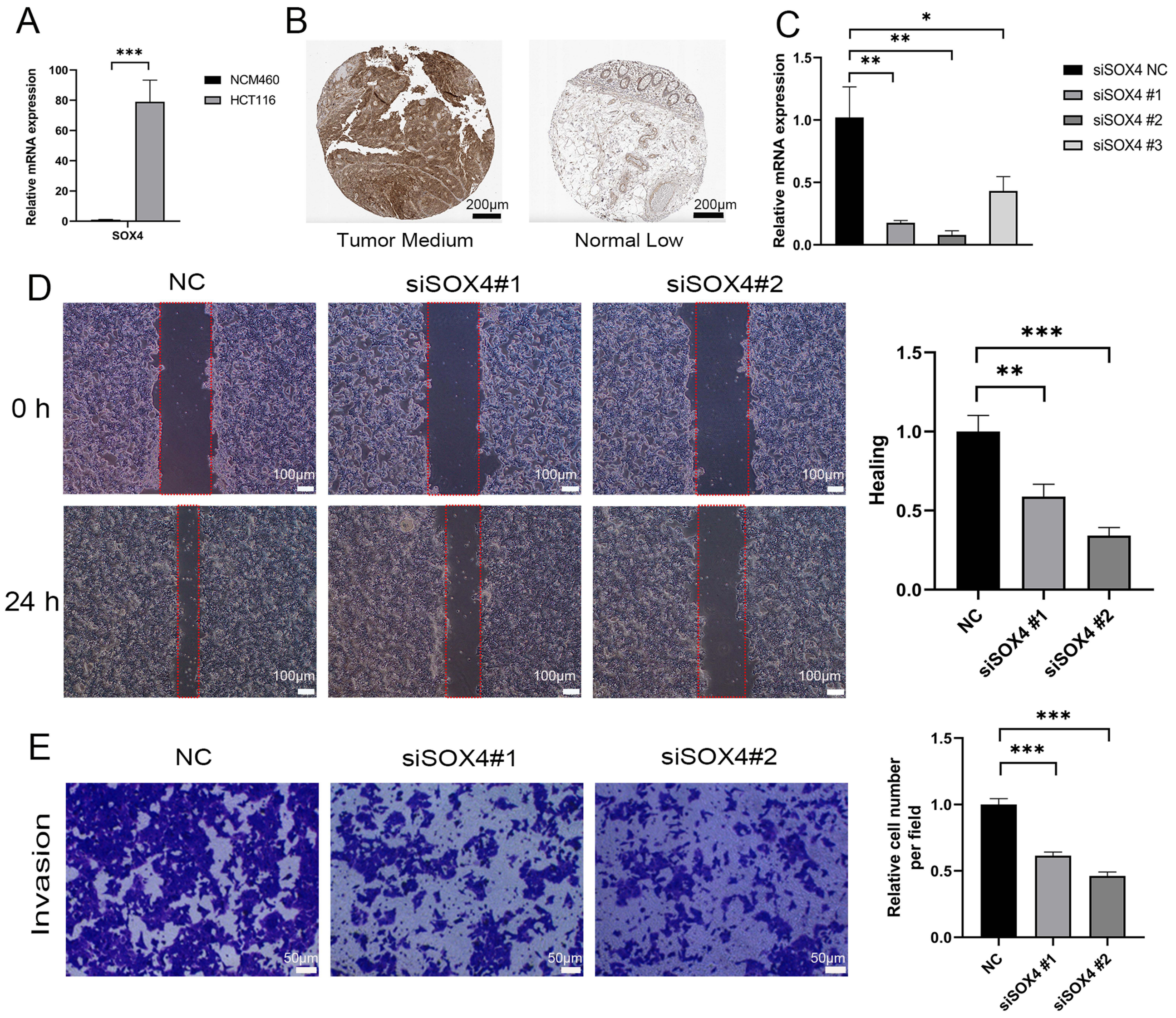

Figure 7 External experimental verification. (**A**) qPCR validation of SOX4 expression levels in tumor and normal cell lines. (**B**) IHC staining of protein expression levels of SOX4 in tumor and normal samples (The scale bar is 200 μm). (**C**) SOX4 silencing in HCT116 cells. (**D**) Scratch experiments to assess the migratory capacity of siSOX4 (The scale bar is 100 μm). (**E**) Transwell experiment to assess the invasive capacity of siSOX4. (#1, #2, #3 are three different siSOX4 sequences, p values were shown as: *$p<0.05$; **$p<0.01$; ***$p<0.001$.The scale bar is 50 μm).

Discussion

CRC remains a prevalent malignancy within the gastrointestinal tract, predominantly affecting men over 50 years old.[22] Salty dietary habits, smoking, excessive alcohol consumption, and prolonged unhealthy lifestyles constitute primary risk factors associated with CRC development.[23] Due to the subtle early symptoms of colorectal cancer, current detection methods often fail to achieve timely screening, resulting in diagnoses frequently accompanied by liver metastasis. This occurrence is primarily attributed to the high heterogeneity of cancer cells, facilitating their proliferation, migration, and invasion.[24] The detection of early liver metastases and subsequent lesion resection have emerged as potential treatment strategies capable of improving cancer patients' outcomes. Previous findings have spurred our interest in exploring the predictive and prognostic value of genes associated with liver metastasis and their immunological relevance to CRC, utilizing a bioinformatics approach.

The emergence of liver metastasis in CRC typically signifies a reshaping of the TIME. The interstitial tissue surrounding the tumor interacts dynamically with immune cells to evade immune surveillance, infiltrate normal colorectal tissue, and metastasize to the liver, resulting in the reprogramming of cells within the liver's TIME to mimic the original tumor niche. This process leads to clinically detectable liver metastasis.[25–27] In the study, we identified the differences in cell composition between LM and CRC.

In addition, CD4 T cell and CD8 T cell were found enriched in both pCRC and LM groups, which suggested that the T cell may main participant in TIME. Further enrichment analysis of GSVA in immune cells suggested that LM of colorectal cancer was associated with the activation of cancer progression-related pathways such as HALLMARK EPITHELIAL MESENCHYMAL TRANSITION, HALLMARK APOPTOSIS, and HALLMARK MTORC1 SIGNALING.[28,29]

Further exploration of the cell origin of pCRC and LM groups through CNV and pseudotime analyses revealed distinct differentiation patterns between primary tumors and liver metastases. Additionally, substantial heterogeneity was observed among tumor samples, highlighting the importance of identifying shifting risk factors. Understanding genes associated with the risk of liver metastases are pivotal as they may serve as prognostic indicators. In this study, we defined a subcluster of malignant epithelial cells that contained both pCRC and LM groups as LMECs and found that Wnt/beta-catenin signaling pathways, G2M checkpoint, and KRAS signaling were the top 3 activated pathways in LMECs, suggesting that these pathways closely associated with LM. The important roles of pathway Wnt/beta-catenin signaling and KRAS signaling in CRC liver metastasis have been reported.[30,31] For G2M checkpoint pathways, it needs to be further proved in vivo and in vitro.

Differentially expressed genes in the pCRC and LM samples of single-cell epithelial cells overlapped in the same direction. Intersection analysis of LMECs and temporal branching revealed genes such as SOX4 associated with the risk of CRC liver metastasis. SOX4, a protein-coding gene primarily involved in cell development and fate determination, has garnered attention in numerous malignancies, including breast cancer, osteosarcoma, neuroblastoma, and colorectal cancer. Multiple studies have positioned SOX4 as an oncogene.[32] It functions as a critical resistance mechanism against T-cell-mediated cytotoxicity in triple-negative breast cancer.[33] Mechanistic studies have revealed that inactivation of SOX4 in tumor cells enhances gene expression in numerous innate and adaptive immune pathways crucial for tumor immunity.[34,35] Similarly, SOX4 regulation in highly aggressive osteosarcoma significantly inhibits cell migration and invasion.[36] Additionally, in neuroblastoma, under the combined effects of SOX4 and p53, the expression of pro-apoptotic proteins is upregulated.[37] In this study, the in vitro experiments confirmed the important role of SOX4 in cell migration and invasion, which indicated that SOX4 may be an important marker for LM in CRC.

Based on the above risk genes, we constructed the LM-index and evaluated its correlation with patients' clinical indicators. The findings indicated a positive correlation between LM-index and TNM stage, with lower survival rates observed among patients with a high LM-index. Significant correlations were also noted between age, tumor spread, tumor metastasis, and LM-index, underscoring the role of age and tumor progression as determinants in CRC occurrence. Linsitinib_1510 and Lapatinib_1558 were unearthed, which may be beneficial to the patients with high LM-index. However, our research possesses certain limitations. Although we have confirmed SOX4 as a CRC biomarker, its role in distinct cell subsets remains unclear. Additionally, further investigations are warranted to explore the drug effect within the subgroup and its significance in preventing metastasis.

Conclusions

In this study, we performed a single-cell profiling of primary CRC and liver metastases with 135,754 cells, providing a fundamental and comprehensive understanding of cellular composition in TIME of primary tumors and liver metastases of CRC. The LM-index was constructed, and the relationship between the LM-index and the prognosis and characteristics of the CRLM patients were identified. Meanwhile, identified candidate drugs to which high LM-index patients may be susceptible. Furthermore, we validate the cell origin representing LMAECs and identify SOX4 as a potential biomarker. These findings lay a fundamental guide for the early identification and management of high-risk patients in LM of colorectal cancer.

Ethical Approval and Consent to Participate

The study was approved by the ethics committee of Eastern Hepatobiliary Surgery Hospital, Naval Medical University.

The ethical considerations related to the dataset used in this study are described below. GSE178318: All patients who provided specimens signed an informed consent form and agreed to the specimens being used for scientific research. PRJNA748525: The written informed consent was received from each patient prior to participation. The study was approved by the Ethics Committee of Zhongshan Hospital and recorded by Ministry of Science and Technology of the

People's Republic of China. PMC9811165: This study was conducted in compliance with the Helsinki Declaration. All patients were enrolled according to a study protocol approved by the Stanford University School of Medicine Institutional Review Board. Written informed consent was obtained from all patients.

Acknowledgments

This study did not receive any specific funding.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Disclosure

The authors affirm that there are no potential conflicts of interest that could be perceived in terms of business or financial relationships in the conduct of this study.

References

1. Moullet M, Funston G, Mounce LT, et al. Pre-diagnostic clinical features and blood tests in patients with colorectal cancer: a retrospective linked-data study. *Br J General Practice*. 2022;72(721):e556–e63. doi:10.3399/bjgp.2021.0563
2. Yang J, Peng JY, Chen W. Synchronous colorectal cancers: a review of clinical features, diagnosis, treatment, and prognosis. *Digestive Surgery*. 2011;28(5–6):379–385. doi:10.1159/000334073
3. Cervantes A, Adam R, Roselló S, et al. Metastatic colorectal cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Ann Oncol*. 2023;34(1):10–32. doi:10.1016/j.annonc.2022.10.003
4. Chen K, Collins G, Wang H, Toh JWT. Pathological Features and Prognostication in Colorectal Cancer. *Current Oncol*. 2021;28(6):5356–5383. doi:10.3390/curroncol28060447
5. Zhang C, Yin S, Tan Y, et al. Patient Selection for Adjuvant Chemotherapy in High-Risk Stage II Colon Cancer: a Systematic Review and Meta-Analysis. *Am j Clin Oncol*. 2020;43(4):279–287. doi:10.1097/coc.0000000000000663
6. Schneider NI, Langner C. Prognostic stratification of colorectal cancer patients: current perspectives. *Cancer Manage Res*. 2014;6:291–300. doi:10.2147/cmar.S38827
7. Akgül Ö, Çetinkaya E, Ersöz Ş, Tez M. Role of surgery in colorectal cancer liver metastases. *World J Gastroenterol*. 2014;20(20):6113–6122. doi:10.3748/wjg.v20.i20.6113
8. Engstrand J, Nilsson H, Strömberg C, Jonas E, Freedman J. Colorectal cancer liver metastases - a population-based study on incidence, management and survival. *BMC Cancer*. 2018;18(1):78. doi:10.1186/s12885-017-3925-x
9. Dendy MS, Ludwig JM, Kim HS. Predictors and prognosticators for survival with Yttrium-90 radioembolization therapy for unresectable colorectal cancer liver metastasis. *Oncotarget*. 2017;8(23):37912–37922. doi:10.18632/oncotarget.16007
10. Zhu D, Ren L, Xu J. Interpretation of guidelines for the diagnosis and comprehensive treatment of colorectal cancer liver metastases in China (v2013). *Chine j Gastrointestinal Surgery*. 2014;17(6):525–529.
11. Zhang Y, Song J, Zhao Z, et al. Single-cell transcriptome analysis reveals tumor immune microenvironment heterogenicity and granulocytes enrichment in colorectal cancer liver metastases. *Cancer Lett*. 2020;470:84–94. doi:10.1016/j.canlet.2019.10.016
12. Zheng X, Ma Y, Bai Y, et al. Identification and validation of immunotherapy for four novel clusters of colorectal cancer based on the tumor microenvironment. *Front Immunol*. 2022;13:984480. doi:10.3389/fimmu.2022.984480
13. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinf*. 2013;14:7. doi:10.1186/1471-2105-14-7
14. Durante MA, Rodriguez DA, Kurtenbach S, et al. Single-cell analysis reveals new evolutionary complexity in uveal melanoma. *Nat Commun*. 2020;11(1):496. doi:10.1038/s41467-019-14256-1
15. Suphavilai C, Chia S, Sharma A, et al. Predicting heterogeneity in clone-specific therapeutic vulnerabilities using single-cell transcriptomic signatures. *Genome med*. 2021;13(1):189. doi:10.1186/s13073-021-01000-y
16. Kurtenbach S, Cruz AM, Rodriguez DA, Durante MA, Harbour JW. Uphyloplot2: visualizing phylogenetic trees from single-cell RNA-seq data. *BMC Genomics*. 2021;22(1):419. doi:10.1186/s12864-021-07739-3
17. Qiu X, Mao Q, Tang Y, et al. Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*. 2017;14(10):979–982. doi:10.1038/nmeth.4402
18. Wang Z, Wang Y, Yang T, et al. Machine learning revealed stemness features and a novel stemness-based classification with appealing implications in discriminating the prognosis, immunotherapy and temozolomide responses of 906 glioblastoma patients. *Brief Bioinform*. 2021;22(5):32. doi:10.1093/bib/bbab032
19. Jafarnejad SM, Wani AA, Martinka M, Li G. Prognostic significance of Sox4 expression in human cutaneous melanoma and its role in cell migration and invasion. *Am J Pathol*. 2010;177(6):2741–2752. doi:10.2353/ajpath.2010.100377
20. Liu Y, Zeng S, Jiang X, Lai D, Su Z. SOX4 induces tumor invasion by targeting EMT-related pathway in prostate cancer. *Tumour Biol*. 2017;39(5):1010428317694539. doi:10.1177/1010428317694539

21. Ruan H, Yang H, Wei H, et al. Overexpression of SOX4 promotes cell migration and invasion of renal cell carcinoma by inducing epithelial-mesenchymal transition. *Int j Oncol*. 2017;51(1):336–346. doi:10.3892/ijo.2017.4010
22. Patel SG, May FP, Anderson JC, et al. Updates on Age to Start and Stop Colorectal Cancer Screening: recommendations From the U.S. Multi-Society Task Force on Colorectal Cancer. *Gastroenterology*. 2022;162(1):285–299. doi:10.1053/j.gastro.2021.10.007
23. Masdor NA, Mohammed Nawi A, Hod R, Wong Z, Makpol S, Chin SF. The Link between Food Environment and Colorectal Cancer: a Systematic Review. *Nutrients*. 2022;14(19). doi:10.3390/nu14193954
24. Liu Y, Zhang Q, Xing B, et al. Immune phenotypic linkage between colorectal cancer and liver metastasis. *Cancer Cell*. 2022;40(4):424–37.e5. doi:10.1016/j.ccell.2022.02.013
25. Zhang Q, Liu S, Liu Y, et al. Liver Metastasis Modulate Responses of Suppressive Macrophages and Exhausted T Cells to Immunotherapy Revealed by Single Cell Sequencing. *Adv Genet*. 2022;3(4):2200002. doi:10.1002/ggn2.202200002
26. Zhao S, Mi Y, Guan B, et al. Tumor-derived exosomal miR-934 induces macrophage M2 polarization to promote liver metastasis of colorectal cancer. *J Hematol Oncol*. 2020;13(1):156. doi:10.1186/s13045-020-00991-2
27. Zhou H, Zhu L, Song J, et al. Liquid biopsy at the frontier of detection, prognosis and progression monitoring in colorectal cancer. *Mol Cancer*. 2022;21(1):86. doi:10.1186/s12943-022-01556-2
28. Li W, Chang J, Wang S, et al. miRNA-99b-5p suppresses liver metastasis of colorectal cancer by down-regulating mTOR. *Oncotarget*. 2015;6 (27):24448–24462. doi:10.18632/oncotarget.4423
29. Gulhati P, Bowen KA, Liu J, et al. mTORC1 and mTORC2 regulate EMT, motility, and metastasis of colorectal cancer via RhoA and Rac1 signaling pathways. *Cancer Res*. 2011;71(9):3246–3256. doi:10.1158/0008-5472.can-10-4058
30. Nash GM, Gimbel M, Shia J, et al. KRAS mutation correlates with accelerated metastatic progression in patients with colorectal liver metastases. *Ann Surg Oncol*. 2010;17(2):572–578. doi:10.1245/s10434-009-0605-3
31. Zhu Y, Li X. Advances of Wnt Signalling Pathway in Colorectal Cancer. *Cells*. 2023;12(3). doi:10.3390/cells12030447published
32. Moreno CS. SOX4: the unappreciated oncogene. *Semin Cancer Biol*. 2020;67(Pt 1):57–64. doi:10.1016/j.semcancer.2019.08.027
33. Bagati A, Kumar S, Jiang P, et al. Integrin αvβ6-TGFβ-SOX4 Pathway Drives Immune Evasion in Triple-Negative Breast Cancer. *Cancer Cell*. 2021;39(1):54–67.e9. doi:10.1016/j.ccell.2020.12.001
34. Qiu Z, Khairallah C, Chu TH, et al. Retinoic acid signaling during priming licenses intestinal CD103+ CD8 TRM cell differentiation. *J Exp Med*. 2023;220(5):923. doi:10.1084/jem.20210923
35. Zhang J, Xiao C, Feng Z, et al. SOX4 promotes the growth and metastasis of breast cancer. *Cancer Cell Int*. 2020;20:468. doi:10.1186/s12935-020-01568-2
36. Bai CJ, Gao T, Liu JY, Li S, Wang XY, Fan ZF. SNHG9/miR-214-5p/SOX4 feedback loop regulates osteosarcoma progression. *Neoplasma*. 2022;69(5):1175–1184. doi:10.4149/neo_2022_220228N218
37. Miyazaki M, Otomo R, Matsushima-Hibiya Y, et al. The p53 activator overcomes resistance to ALK inhibitors by regulating p53-target selectivity in ALK-driven neuroblastomas. *Cell Death Discov*. 2018;4:56. doi:10.1038/s41420-018-0059-0