IVYSPRING
INTERNATIONAL PUBLISHER

Journal of Cancer

2020; 11(23): 6861-6873. doi: 10.7150/jca.49262

Research Paper

Pathologic evolution-related Gene Analysis based on both single-cell and bulk transcriptomics in Colorectal Cancer

Jiali Li[1], Zihang Zeng[1], Jiarui Chen[1], Xingyu Liu[1], Xueping Jiang[1], Wenjie Sun[1], Yuan Luo[1], Jiangbo Ren[2], Yan Gong[2,3]✉ and Conghua Xie[1,4,5]✉

1. Department of Radiation and Medical Oncology, Zhongnan Hospital of Wuhan University, Wuhan, China.
2. Department of Biological Repositories, Zhongnan Hospital of Wuhan University, Wuhan, China.
3. Human Genetics Resource Preservation Center of Hubei Province, Zhongnan Hospital of Wuhan University, Wuhan, China.
4. Hubei Key Laboratory of Tumor Biological Behaviors, Zhongnan Hospital of Wuhan University, Wuhan, China.
5. Hubei Cancer Clinical Study Center, Zhongnan Hospital of Wuhan University, Wuhan, China.

✉ Corresponding authors: Conghua Xie, Department of Radiation and Medical Oncology, Zhongnan Hospital of Wuhan University, 169 Donghu Road, Wuhan, Hubei 430071, China. Tel.: +86-27-67812607; E-mail: chxie_65@whu.edu.cn; Yan Gong, Department of Biological Repositories, Zhongnan Hospital of Wuhan University, 169 Donghu Road, Wuhan, Hubei 430071, China. Tel.: +86-27-67811461; E-mail yan.gong@whu.edu.cn.

Received: 2020.06.09; Accepted: 2020.10.11; Published: 2020.10.22

Abstract

Purpose: The patients diagnosed with colorectal cancer (CRC) are likely to undergo differential outcomes in clinical survival owing to different pathologic stages. However, signatures in association with pathologic evolution and CRC prognosis are not clearly defined. This study aimed to identify pathologic evolution-related genes in CRC based on both single-cell and bulk transcriptomics.

Patients and methods: The CRC single-cell transcriptomic dataset (GSE81861, n=590) with clinical information and tumor microenvironmental tissues was collected to identify the pathologic evolution-related genes. The colonic adenocarcinoma and rectum adenocarcinoma transcriptomics from The Cancer Genome Atlas were obtained as the training dataset (n=363) and 5 other CRC transcriptomics cohorts from Gene Expression Omnibus (n=1031) were acquired as validation data. Graph-based clustering analysis algorithm was applied to identify pathologic evolution-related cell populations. Pseudotime analysis was performed to construct the trajectory plot of pathologic evolution and to define hub genes in the evolution process. Cell-type identification by estimating relative subsets of RNA transcripts was then executed to build a novel cell infiltration classifier. The prediction efficacy of this classifier was validated in bulk transcriptomic datasets.

Results: Epithelial and T cells were elucidated to be related to the pathologic stages in CRC tissues. Pseudotime analysis and survival analysis indicated that HOXC5, HOXC8 and BMP5 were the marker genes in pathologic evolution process. Our cell infiltration classifier exhibited excellent forecast efficacy in predicting pathologic stages and prognosis of CRC patients.

Conclusion: We identified pathologic evolution-related genes in single-cell transcriptomic and proposed a novel specific cell infiltration classifier to forecast the prognosis of CRC patients based on pathologic stage-related hub genes HOXC6, HOXC8 and BMP5.

Key words: single-cell sequencing; colorectal cancer; pathologic stage; prognosis; TCGA; GEO

Introduction

Colorectal cancer (CRC) results in as many as 900,000 deaths each year, accounting for approximately 9.2% cancer mortality worldwide [1]. It is expected to increase by more than 20% to 1,100,000 in 2030 [2]. In addition, statistical analysis indicated that its incidence substantially increased in patients younger than 40s [3, 4]. Although the developments of early diagnosis, immunotherapy and chemotherapy remarkably facilitate the detection and curation of CRC, the overall survival remains less than 60% in developed countries [5]. During the last few decades, numerous researches focused on molecular heterogeneity in colorectal oncocytes [6, 7], which probably contributed to totally divergent clinical endings. Explorations on the underlying genetic alteration are highly required to benefit CRC patients.

The pathologic stage exerts a decisive role in different therapy strategies as well as clinical treatment outcomes [8-10]. However, the signatures of pathologic progress and stage promotion remain unclear. Surgery is recommended as the preferred treatment in stage I colon cancer, and chemotherapy composes of the therapeutical strategy as an integral part for patients at III or IV stage [5]. Hence, more prognostic markers are required to uncover the decisive genes in pathologic processes, like RAS activation or function loss of TP53 during CRC tumorigenesis [6]. Moreover, with the advances in single-cell RNA sequencing (scRNA-Seq) technique, a novel approach is provided to clarify the strong heterogeneity in predominant cell populations and profound alteration of key gene expression [11, 12]. As is reported, CRC is classified into 4 subtypes based on consensus molecular subtypes (CMS) [7]. The pathways or genes implicated in each subtype are unique: hypermutated, strong immune activation and microsatellite unstable in CMS1; striking WNT and MYC pathway activation in epithelial in CMS2; marked metabolic dysregulation of epithelial in CMS3; overt transforming growth factor-β (TGF-β) activation, stromal angiogenesis and invasion within mesenchymal in CMS4. In brief, dominant gene expression alteration in decisive cell populations attracted increasing attention.

In this study, we combined scRNA-seq and bulk transcriptomic to identify pathology-related cell populations and pathologic progress-related hub genes. A novel specific cell infiltration classifier was established to forecast pathologic stages and prognosis of CRC patients based on the 3 hub genes HOXC6, HOXC8 and BMP5. Our study provided a potential classification and key biomarkers to screen out CRC patients with primary pathology stages that were more likely to obtain a better prognosis.

Material and Methods

Data Preprocessing and Quality Control

The workflow is exhibited in Figure 1. The dataset (GSE81861) consisting of scRNA-Seq and clinical information of 11 primary CRC patients [13] was obtained from Gene Expression Omnibus (GEO) database. A total of 590 single-cell samples derived from CRC tissues or matched normal mucosa were included in this study. The selected cell samples contained different cell populations, including epithelial cells (tumor cells), endothelial cells, T cells, B cells, fibroblasts, macrophages and mast cells. Only cells that possessed at least 10,000 total counts and 500 expressed genes were included in the following analysis. The cell-level and sequencing profile diagnosis were executed sequentially. At the same time, count-per-million standardization was applied to optimize the library size. Quality control of scRNA-Seq was conducted with *scater* package.

Bulk RNA sequencing (RNA-Seq) of CRC cohorts were acquired as independent training dataset including colonic adenocarcinoma (COAD) and rectum adenocarcinoma (READ) data from the cancer genome atlas (TCGA), containing 363 CRC patients, Transcripts Per Million (TPM) standardization method and Z-score normalization were subsequently performed on TCGA RNA-Seq. Validation datasets were bulk genome chip data acquired from GEO database (GSE39582, n=552; GSE37892, n=130; GSE12945, n=62; GSE17537, n=55; GSE17538, n=232) [14-16]. All the above 5 datasets were processed with robust multi-array average (RMA) standardization and Z-score normalization. The outlines of the datasets were displayed in Table 1, including sample capacity, age, gender proportion and survival. The clinical characteristic used in this study was the pathologic stage, which was assessed based on the standard of American Joint Committee on cancer. Stages higher than 2 (including pathologic stage 2) were regarded as high pathologic stages and stage1 was defined as low pathologic stages.

Dimensionality Reduction

Principal components analysis (PCA) was a classical linear dimensionality reduction algorithm [17, 18]. Eigenvalues and eigenvectors of covariance matrix were calculated to estimate correlations between variables. After that, several larger eigenvectors were extracted from the matrix represented as the principal components.

T-distributed stochastic neighbor embedding (t-SNE) was a nonlinear dimensionality reduction algorithm [19] used to dispose of the nonlinear correlations between variables. In the process of dimensionality reduction, instead of Euclidean distance, this algorithm selected conditional probability of choosing another point as the adjacent node to reflect the similarity of 2 nodes [20]. For strict quality estimation, both t-SNE and PCA were applied to investigate the distribution of cell types and tissue types with the log-transformed expression values in R software.

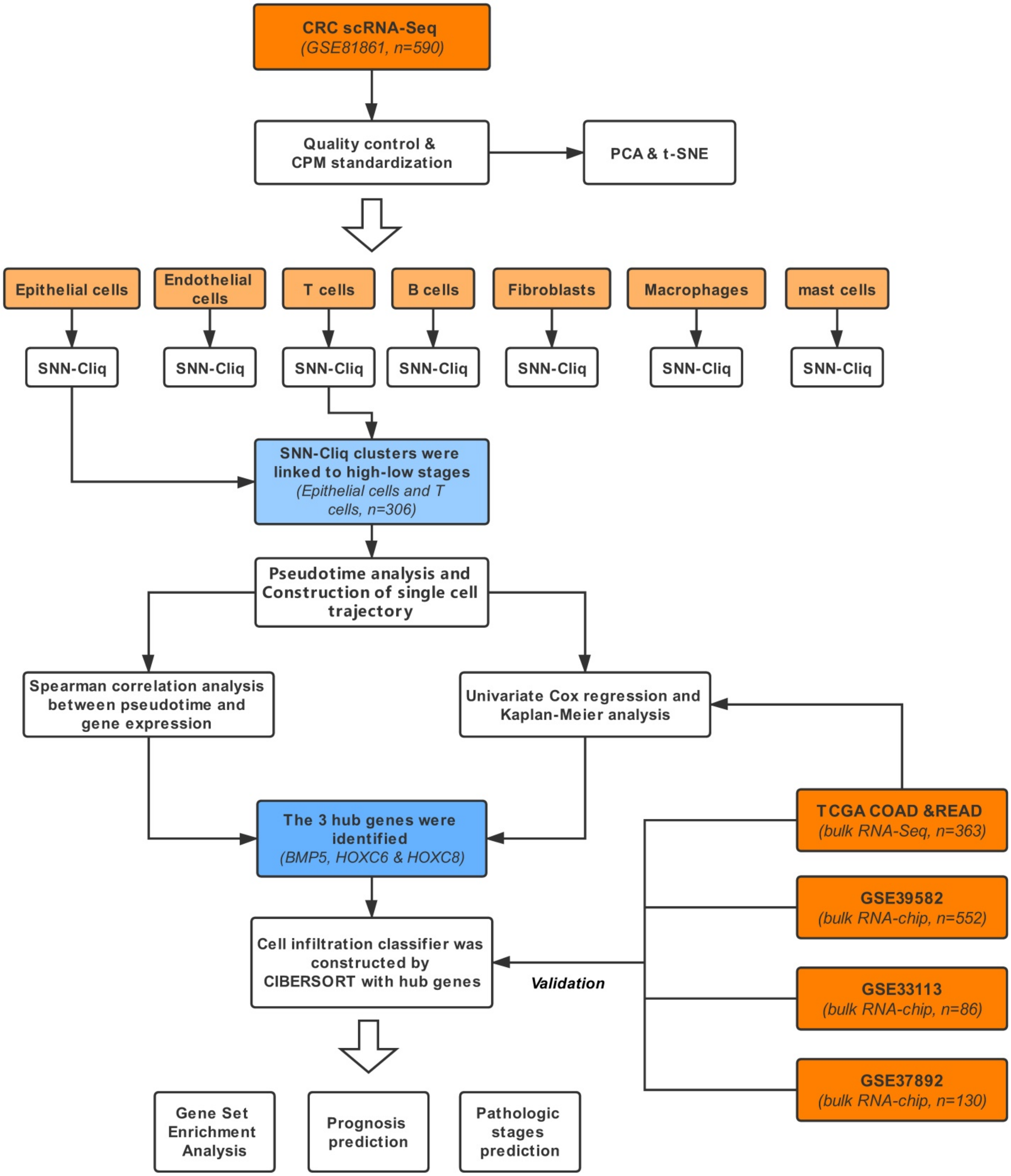
CRC scRNA-Seq
(GSE81861, n=590)
Quality control & CPM standardization
PCA & t-SNE
Epithelial cells
Endothelial cells
T cells
B cells
Fibroblasts
Macrophages
mast cells
SNN-Cliq
SNN-Cliq
SNN-Cliq
SNN-Cliq
SNN-Cliq
SNN-Cliq
SNN-Cliq
SNN-Cliq clusters were linked to high-low stages
(Epithelial cells and T cells, n=306)
Pseudotime analysis and Construction of single cell trajectory
Spearman correlation analysis between pseudotime and gene expression
Univariate Cox regression and Kaplan-Meier analysis
TCGA COAD &READ
(bulk RNA-Seq, n=363)
The 3 hub genes were identified
(BMP5, HOXC6 & HOXC8)
GSE39582
(bulk RNA-chip, n=552)
Cell infiltration classifier was constructed by CIBERSORT with hub genes
Validation
GSE33113
(bulk RNA-chip, n=86)
GSE37892
(bulk RNA-chip, n=130)
Gene Set Enrichment Analysis
Prognosis prediction
Pathologic stages prediction

Figure 1. Workflow of this study.

Table 1. General property information of validation datasets

Dataset	TCGA (COAD&READ)	GSE39582	GSE37892	GSE12945	GSE17537	GSE17538
Sample capacity	363	552	130	62	55	232
Data type	RNA-Seq	Gene chip	Gene chip	Gene chip	Gene chip	Gene chip
Standardization method	TPM	RMA	RMA	RMA	RMA	RMA
Clinical characteristic	overall survival	overall survival	pathologic stage	overall survival	overall survival& pathologic stage	overall survival& pathologic stage
Median age (IQR)	56.57-76.29	59.00- 76.00	59.25-76.00	59.00-73.75	54.00- 72.00	56.00- 74.00
Gender ratio (M/F)	1.18	1.22	1.13	1.21	0.90	1.11
Survival (Median year)	6.94	12.08	NA	not arrive median survival	not arrive median survival	11.24

Graph-Based Clustering Analysis

A novel algorithm named shared nearest neighbor (SNN)-Cliq was developed combining SNN method with quasi-clique-based clustering method [21]. In SNN-Cliq, input nodes were the vectors of gene expression within an individual cell. Similarity (Euclidean distance) between points was utilized as a weighted edge to construct SNN graph. In addition, graph-theoretic techniques were included to cluster the sparse SNN graph [22]. This algorithm not only included Euclidean distance as a similarity measure but also combined with a quasi-clique-based clustering algorithm to accurately identify the highly similar nodes in the same cluster. Graph-based clustering analysis was implemented to subclassify the pathologic stage-related cell populations with *scran* package in R.

Pseudotime Analysis

Pseudotime analysis was an algorithm to construct the development trajectory of a single-cell lineage according to the gene expression profile [23]. It infers the cell development trajectory from the expression level changes of the pre-defined phenotype-related gene list and then distributes every single cell with its proper pseudotime in this trajectory. Cell development processes were exhibited on the trajectory plot. The above process was realized by *monocle* package using R to restore the development trajectory of pathologic evolution-related cell populations.

Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) was a function annotation approach aiming to identify probable significant biologic expression-changed signatures based on known gene sets [24]. The enrichment score was calculated based on whether a certain gene belonged to the known gene sets or not. P values and normalized enrichment scores of certain pathways were obtained via permutation test. GSEA was achieved by *clusterProfiler* package to elucidate the key genes among differential pathologic evolution-related cell population clusters [25].

Cell Infiltration Abundance Estimation

Cell-type identification by estimating relative subsets of RNA transcripts (CIBERSORT) was a computational deconvolution approach [26] to characterize cell composition of a mixture based on bulk gene expression profiles. The expression matrix of pre-defined cell markers worked as the reference to estimate the relative proportion of specific cell types in bulk RNA-seq profiles. A linear support vector regression, a machine learning method to denoise, was applied to deconvolve the bulk gene expression matrix. To establish the specific cell infiltration classifier, a gene matrix defining 2 different populations (C1 and C2 populations) was created using expression quartiles of the pathologic hub genes in CRC bulk transcriptome and then was input as the reference matrix, representing for 2 pathologic evolution-related cell classifications. The concrete process was performed using *CIBERSORT* package in R.

Statistical Analysis

To identify the hub genes and validate the novel cell infiltration classifier, survival analysis was performed using both Kaplan-Meier survival estimation for classified variables, and Cox proportional hazards regression for quantitative index by R *survival* package. Chi-square test was implemented to verify the relevance between pathologic evolution-related cell classifications and realistic pathology stages. Spearman and Pearson correlation analysis were calculated respectively via the stats package. The above statistical analyses were completed with R 3.6.1. In all hypothesis tests, *P*-values less than 0.05 were regarded as statistical significance. All the *P*-values were two-sided.

Results

Quality estimation of scRNA-Seq

The quality of GSE81861 containing 590 single-cell samples from 11 primary CRC patients with tumor and microenvironmental cell populations were first evaluated. The ERCC spike-in and mitochondria genes serving as known control signatures were detected in each cell to calculate the percentage of counts that come from the feature control set (Figure 2A, Figure 2B). Well-behaved cells, which contained a large proportion of expressed features as well as a small ratio expression of spike-in and mitochondria features, while others were discarded. At the same time, the mean expressions of features and coefficient of variation in cells were calculated to acquire highly expressed features (Figure 2C). Five bins with similar expression levels were encapsulated and shown in Figure 2D. Genes in bin 3, 4, and 5 that fell below the 2 * background were excluded. The filtered 1261 genes and 590 cells with high quality were preserved to the next step. Dimensional reduction results implied that cells from CRC tissues and normal mucosa were distinctly separated (PCA, Figure 2E; t-SNE, Figure S1A). The same results were observed in cell populations (PCA, Figure 2F; t-SNE, Figure S1B). The quality estimation outcome demonstrated that GSE81861 scRNA-Seq had a good performance in the strict process of quality control.

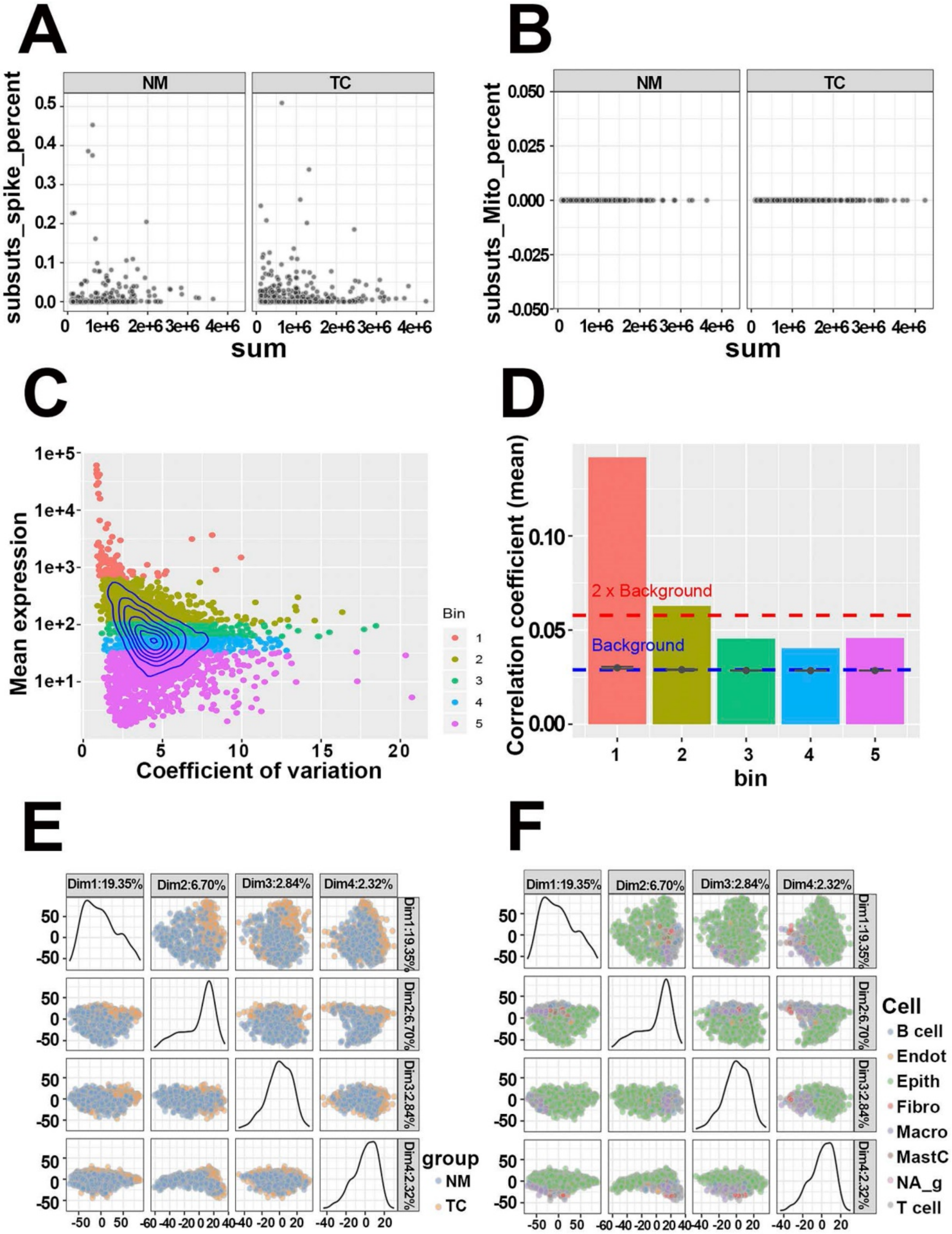

Figure 2. Quality control of single-cell RNA-Seq from GSE81861 dataset. (**A**) Respective detection of ERCC spike-in genes as control features in tumors tissues and matched normal mucosa. (**B**) Respective detection of mitochondria genes as control features in tumors tissues and matched normal mucosa. (**C**) Topographic map of features expression. The relationships between gene mean expression and coefficient of variation were displayed in this map with 5 different colors bins representing for corresponding expression levels. (**D**) Histogram of correlation coefficients of each bin calculated by Pearson correlation. Every feature in each bin was correlated to every other feature in the same bin. Mean value of the correlations were taken as the vertical axis variable. (**E**) Principal components extracted from covariance matrix between colorectal tumor tissue and normal mucosa. (**F**) Principal components extracted from covariance matrix between different cell populations.

Identification of pathologic stage-related cell populations

We selected 375 CRC cells with pathologic stage information of patients for downstream analyses. First, each cell type was clustered using graph-based methods. Tumor epithelial cells were classified into 6 clusters (Table S1), which distinguished high and low stages (P<0.0001, Fisher exact test). At the same time, T cells were sorted into 2 clusters respectively (Table S2). As expected, the 2 clusters completely corresponded to different stages (P<0.0001, Fisher exact test). However, other cell populations, including endothelial cells, B cells, fibroblasts, macrophages and mast cells exhibited no statistically distinct association with pathologic stages (all, P>0.05, Fisher exact test).

To gain better insight into signature expression in heterogeneous cell clusters, we detected maker genes of each pathologic stage-related cell cluster in T cells and tumor epithelial cells. Several signatures were significantly elevated in the clusters of epithelial cells with a low pathologic stage (Table S3): tumor necrosis factor receptor superfamily member 11b and kallikrein-related peptidase 7 in cluster 1, cadherin related family member 2 in cluster 5, growth factor receptor-bound protein 14, lymphoid enhancer-binding factor 1 and dynamin 1 in cluster 6. Within the cluster of T cells related to low pathologic stages, C-C chemokine receptor type 8, CD27 molecule, cell division cycle 34 and cyclin-dependent kinase 4 in cluster 2 were overexpressed compared to high pathologic stage-related cluster 1 (Table S4).

Definition of pathologic stage evolution-related hub genes

To identify pathologic stage evolution-related hub genes, we next applied pseudotime analysis to the 306 cells that belonged to either epithelial or T cells in scRNA-Seq. A total of 2356 differentially expressed genes in response to pathologic stage evolution process were filtered based on the average expression level and dispersion empirical across cells (Figure 3A). The inferred developmental trajectory was demonstrated as a tree-like structure, exhibiting different cell states and gene expressions (Figure 3B). Pathologic stages presented a divergent distinction between high and low stages, with stage 3 located distant from stages 1 and 2 in trajectory plot (Figure 3C). The box plot between pseudotime and stages supported our speculation that pathologic stage classification was negatively correlated with pseudotime (Figure 3D, *P*<0.0001, variance analysis). On the other hand, Spearman correlation analysis was applied to 2356 differentially expressed genes to determine the potential pseudotime-related signatures. According to above analyses, a total of 64 pathologic stage positive-related genes and 20 negative-related genes were extracted from the single-cell expression profile (Table S5, *P*<0.05, |R|>0.2). Heatmap visualized all the pseudotime-dependent genes into 4 clusters based on their pseudotemporal expression pattern (Figure 3E). Taken together, our results indirectly linked pathologic stage to gene expression bridged by pseudotime, screening out the 84 pathologic stage evolution-related genes at a single-cell level. Subsequently, both univariate cox proportional hazards regression and Kaplan-Meier survival estimation were executed to assess the prognostic effect of the 84 genes on primary CRC patients from the TCGA bulk transcriptomics database. HOXC6, HOXC8 and BMP5 exhibited statistical significance on prognosis prediction in both methods (all, *P*<0.05, Table 2). Based on the aforementioned process, the 3 hub genes with correlation to both pathologic evolution and prognosis in CRC were finally determined.

More specifically, the scatter diagrams separately depicted the distribution of the 3 hub genes in high and low pathologic stages, varying with pseudotime in scRNA-Seq (Figure 4A-C). The results confirmed that highly expressed HOXC6 and HOXC8 as well as downregulated BMP5 were associated with higher pathologic stages. In addition, survival analysis suggested that both HOXC6 and HOXC8 were associated with poor prognosis (both HR=1.90, P<0.005, Table 2), whereas BMP5 played a positive role in better clinical outcome (HR=0.57, P=0.012, Table 2). Mean expression value calculation indicated that HOXC6 and HOXC8 were expressed only in epithelial cells, while BMP5 was activated in both cell populations (Figure S2A). The detailed expression levels of the 3 hub genes in different cell populations were displayed in Figure S2B-D. Moreover, correlation analysis validated the statistically significant relationship between pseudotime and expression levels again, separately in both cell populations (Table 3). Finally, T cell immune infiltration portraits of hub genes in COAD and READ were explored in Tumor Immune Estimation Resource database [27] (Figure 4D-F).

Table 2. Results of prognosis analyses with the 3 hub genes in TCGA colorectal cancer

Gene symbol	*P*-value (KM)	HR (KM)	*P*-value (COX)	HR (COX)
HOXC6	3.00E-03	1.90	3.12E-04	1.16
HOXC8	2.90E-03	1.90	2.92E-04	1.23
BMP5	1.20E-02	0.57	3.70E-04	0.83

Table 3. Results of correlation analysis between pseudotime and expression levels of hub genes

Gene symbol	Cell population	*P* value	R
HOXC6	epithelial cells	1.95E-02	-0.14
HOXC8	epithelial cells	7.50E-03	-0.16
BMP5	epithelial cells	1.33E-05	0.26
BMP5	T cells	4.10E-03	0.48

Cell infiltration classifier construction, evaluation and validation

According to the differential expression pattern of hub genes, we defined 2 cell classifications as the reference matrix. C1 population represented a cell category, in which HOXC6 and HOXC8 were highly expressed, while BMP5 was downregulated while C2 cells behaved oppositely (Table S6). Based on the support vector regression algorithm, CIBERSORT accurately calculated the relative proportions of distinct cell classifications in TCGA CRC bulk transcriptomics (Table S7). To confirm the efficiency of the 2 cell classifications on prognosis prediction, survival analysis was applied to these patients. Patients with C1/C2 cell infiltration coefficients larger than 1 were considered as high C1 infiltration group, and results suggested that patients in higher C1/C2 ratio group suffered from a significantly unfavorable survival (Figure 5A, *P*=0.0029). Meanwhile, the outcome of univariate Cox regression with the same TCGA dataset supported the above conclusion (*P*=0.002). Correlation analysis implied that the lower C1/C2 ratio exhibited a statistical correlation with the lower pathologic stage in CRC patients (*P*=0.028, Spearman correlation test). Of note, all the 3 hub genes were involved in the TGF-β signaling pathway. GSEA results suggested that cellular response and regulation to TGF-β stimulus contributed to the distinct clinical results of the 2 infiltration groups (Table S8). Thus, a practical cell infiltration classifier based on the expression level of HOXC6, HOXC8 and BMP5 was eventually established to predict prognosis, as well as pathologic stage differences in primary CRC patients.

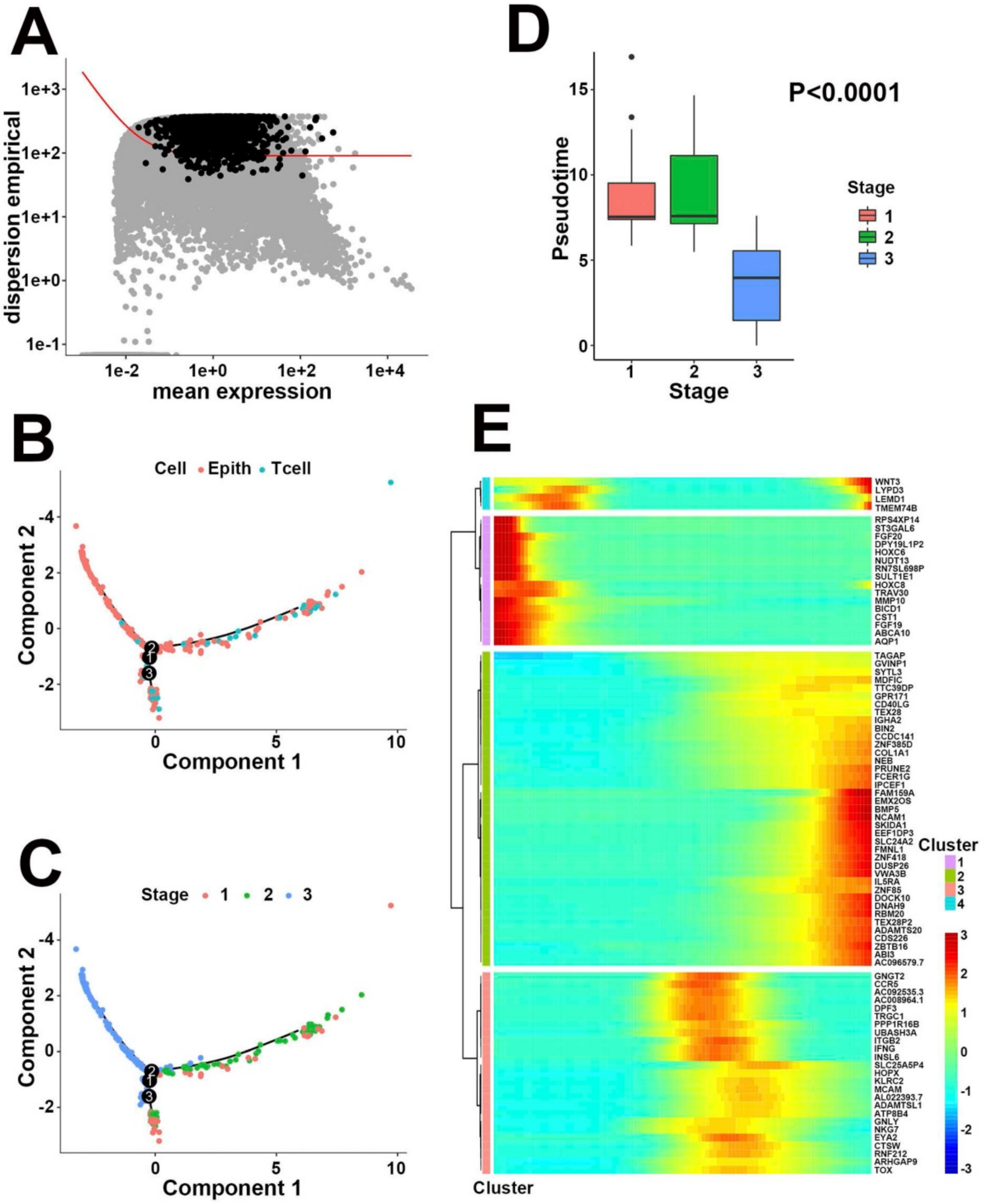

Figure 3. Negative correlation between pseudotime and pathologic stage. (**A**) The identification of 2356 differentially expressed features based on the average expression level and unusually variable expression across cells. (**B**) Trajectory analysis colored by cell types. (**C**) Trajectory analysis colored by pathologic stages. Stage 3 was located far distant from stage 1 and 2. (**D**) Box plot between pseudotime and stage. Stages 1 and 2 showed similar pseudotime levels while stage 3 exhibited lower pseudotime compared to other stages. (**E**) Heatmap of gene expression level varying with pseudotime in 4 clusters. Enhanced genes (in red) with lower pseudotime in cluster 1 represented for the higher pathologic stages, while upregulated signatures (in red) in cluster 2 performed completely the opposite.

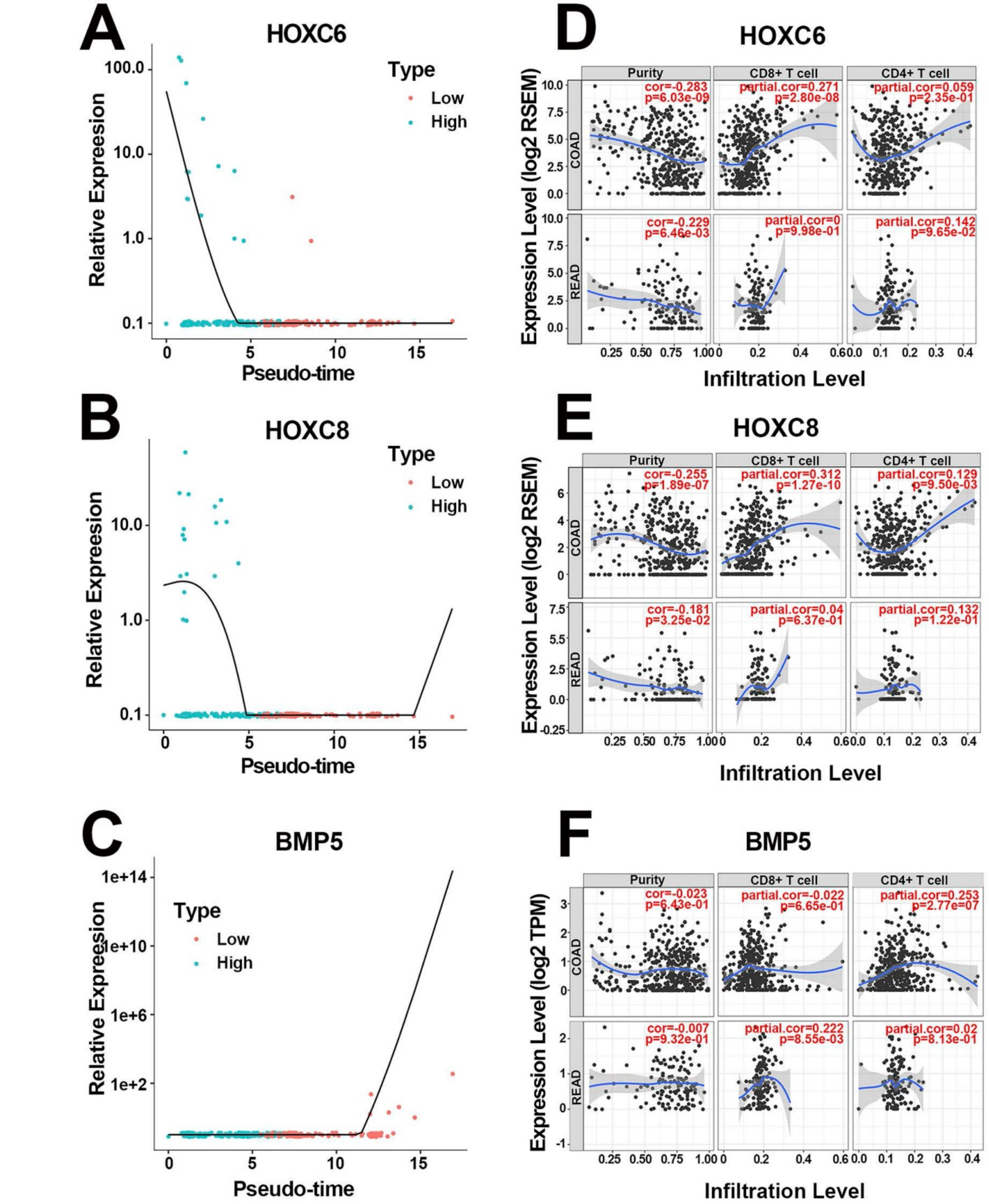

Figure 4. Portraits of the 3 hub genes. (**A**) The pseudotime-expression scatter diagrams of HOXC6. Cells with higher pathologic stage possessed smaller pseudotime and higher HOXC6 expression level. (**B**) The pseudotime-expression scatter diagrams of HOXC8. Cells with higher pathologic stage possessed smaller pseudotime and higher HOXC8 expression level. (**C**) The pseudotime-expression scatter diagrams of BMP5. Cells with higher pathologic stage possessed smaller pseudotime and lower BMP5 expression level. (**D**) Immunocyte infiltration map of HOXC6 in COAD and READ. (**E**) Immunocyte infiltration map of HOXC8 in COAD and READ. (**F**) Immunocyte infiltration map of BMP5 in COAD and READ.

To validate the efficiency of our cell infiltration classifier, a set of independent primary CRC bulk genome chip datasets (GSE39582, GSE33113, GSE37892, GSE12945, GSE17537, and GSE17538) was collected from the GEO database. The 3 hub gene expression profiles were relatively extracted out as the input data of our classifier and the cell infiltration estimation classifications of each patient were defined based on C1/C2 proportion. Subsequent survival analysis exhibited the lower C1/C2 proportion were statistically linked to better prognosis of CRC patients (HR= 1.48, *P*=0.0077, GSE39582; HR=3.36, *P*=0.096, GSE12945; HR= 1.52, *P*=0.072, GSE17538; HR= 3.16, *P*=0.1, GSE17537; Figure 5B-D, Figure 6A). Variance analysis was applied to the sample to validate the association between the infiltration C1/C2 ratio and pathologic stages. Box plot implied lower pathologic stages were associated with lower C1/C2 ratio (*P*=0.0024; *P*=0.0027; *P*=0.016; Figure 6B-D). These results validated the prediction efficiency of our novel constructed cell infiltration classifier based on the expression of the 3 hub genes, confirming that CRC patients with upregulated BMP5 and downregulated HOXC6/8 were related with lower pathologic stages and better prognosis.

Discussion

With the iteration of cell sequencing technology, more details about differentiation map and transcriptional heterogeneity have been clarified [28, 29]. The application of scRNA-Seq facilitates the excavation of tumorigenesis and molecular classifications compared to bulk sequencing [30-32]. By combining scRNA-Seq with clinical characteristics, we revealed the hub genes in the course of CRC pathologic stage evolution and clarified the impact of differential cell infiltration classifications on CRC prognosis based on hub gene expression. A novel classifier to estimate specific cell infiltration was established, which might indicate the pathologic stages and clinical outcomes.

Based on the application of scRNA-Seq, 2 distinct subtypes of cancer-associated fibroblasts associated with prognosis and an mRNA-miRNA regulatory network of CRC were identified in previous studies [13, 33]. However, in this study, tumor epithelial and T cells were discovered to exert an essential role in the evolution of CRC pathologic stage for the first time by graph-based clustering. Besides, the 3 hub genes were later identified

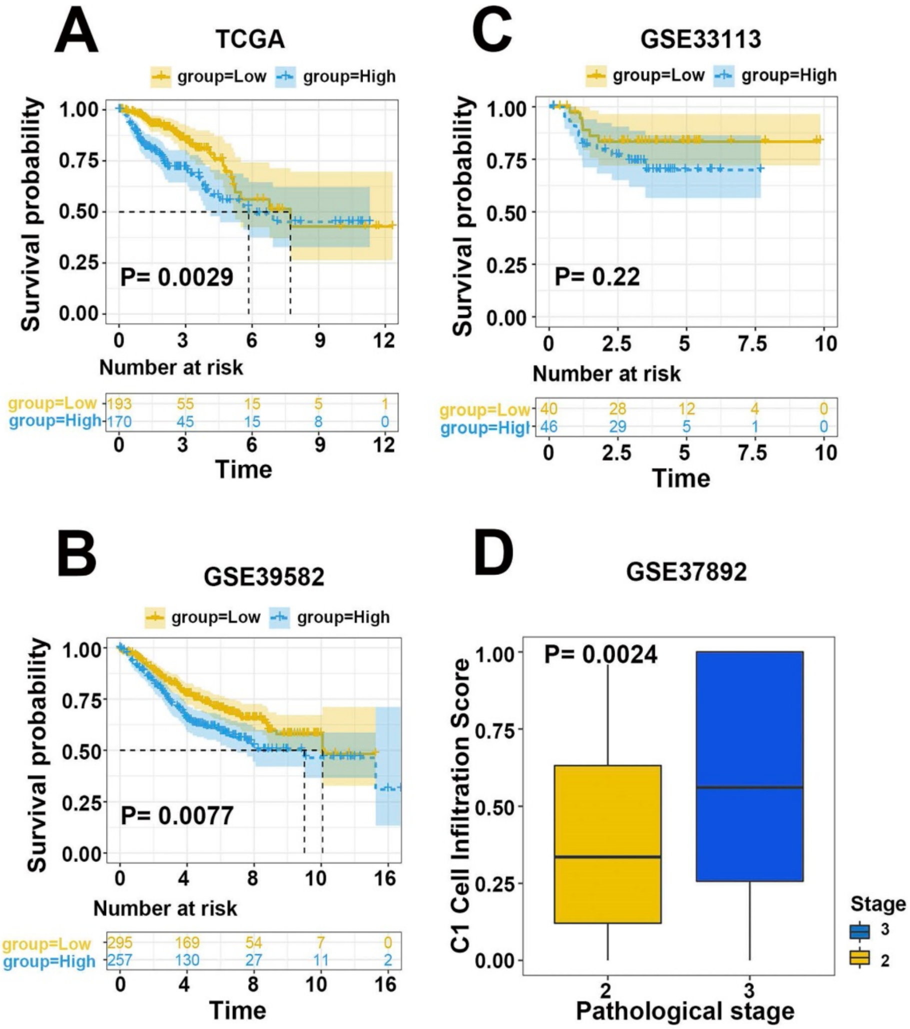

Figure 5. Evaluation and validation of efficiency of the cell infiltration classifier on prognosis.(**A**) Prognosis analysis with TCGA CRC training dataset. Patients in low C1 cells infiltration group had significantly better survival than those in the high group. (**B-D**) Prognosis analysis with independent validation datasets GSE39582, GSE12945 and GSE17538.

(HOXC6, HOXC8 and BMP5) via pseudotime, correlation and survival analysis. Importantly, instead of Multi-Omics Matrix Factorization [34], a more stable and effective algorithm CIBERSORT was performed to construct the cell infiltration classifier, which divided cells in bulk tumor tissue into 2 classifications (C1: HOXC6 ^{high}, HOXC8 ^{high}, BMP5 ^{low}; C2: HOXC6 ^{low}, HOXC8 ^{low}, BMP5 ^{high}). Independent GEO datasets strongly validated that patients in C2 group with upregulated BMP5 and downregulated HOXC6/8 were related to lower pathologic stage and better clinical survival. Our study demonstrated that HOXC6, HOXC8 as well as BMP5 were implicated in the pathologic evolution and a classifier on the basis of the expression of the 3 genes could serve as a prognostic factor, which would facilitate the clinical survival forecast for CRC patients.

Notably, the 3 hub genes were all involved in TGF-β signaling pathway [35-38]. Recent researches emphasized that TGF-β signaling inhibited proliferation and proceeded apoptosis in CRC epithelial cells [39-41]. Escaping the growth-inhibiting effect of TGF-β signaling in tumor epithelial cells promoted CRC development [7]. This signaling was also involved in epithelial-mesenchymal transition (EMT) to induce CRC metastasis [42] as well as T cell exclusion and immune failure in tumor immune microenvironment [43]. A Pan-Cancer Analysis highlighted that gene alteration in TGF-β pathway was carried by 39% cancers, especially gastrointestinal (GI) cancer, and BMP5 was one of the 6 recurrent hotspot mutations in GI cancers [44]. In fact, BMP5 was identified as a tumor suppressor in sporadic CRC and the loss of BMP5 happening at early stages of CRC was linked to the poor survival of patients [45]. In addition, Romagnoli M revealed that BMP5 was repressed by Blimp-1 during EMT process via TGF-β1 in breast cancer and that the poor prognosis of breast cancer was associated with BMP5 low expression [46]. HOXC6 was overexpressed in multiple solid tumors, like hepatocellular carcinoma, cervical carcinoma (CC), head and neck cancer and GI carcinoids [47-50]. In CC, HOXC6 silencing repressed the activation of TGF-β signaling pathway via blocking smad-4 phosphorylation, thus inhibiting cell proliferation and EMT in CC cells [51]. A human tissue microarray

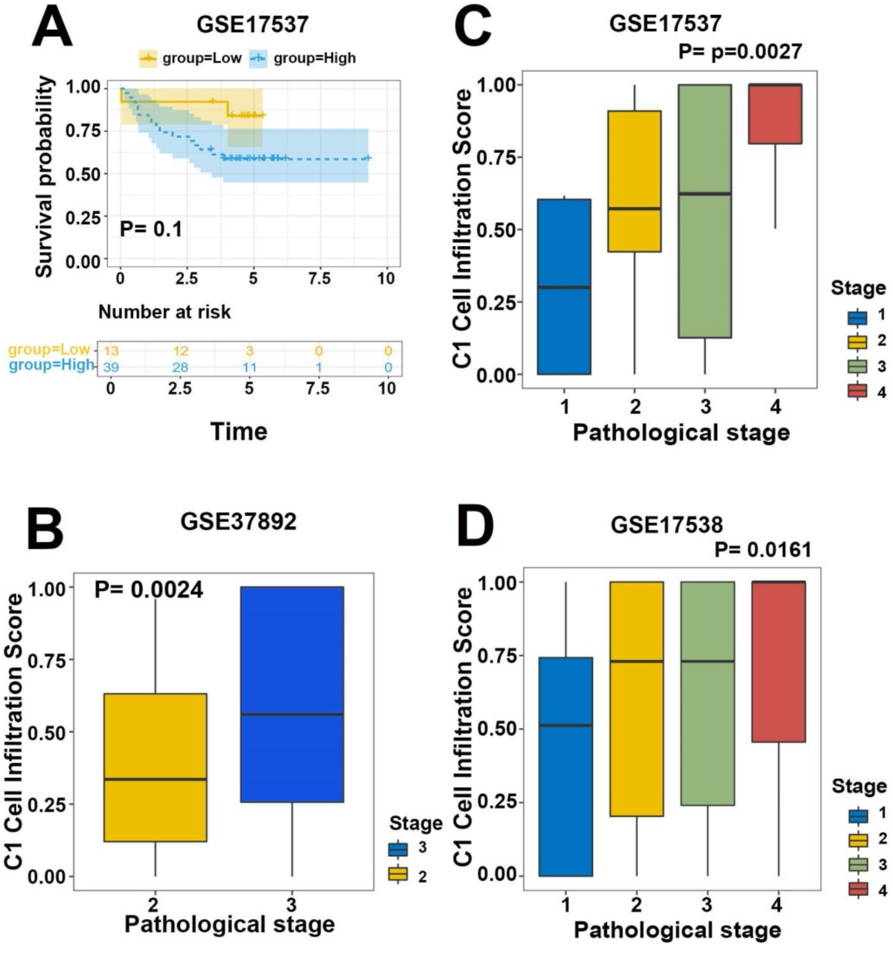

Figure 6. Validation of efficiency of the cell infiltration classifier on prognosis and pathologic stages. (**A**) Patients in low C1 cells infiltration group in GSE17537 had better survival tendency. (**B-D**) Box plot between pathologic stage and C1 cell infiltration with independent validation datasets GSE37892, GSE17537 and GSE17538. Lower pathologic stages were associated with lower C1 cell infiltration scores.

containing 462 samples from CRC patients detected significantly higher HOXC6 expression in tumor tissues compared to matched normal mucosa (P<0.001), which also acted as an independent prognostic marker to poor overall survival [52]. Moreover, HOXC6 deregulation decreased CRC cell growth mediated by the suppression of autophagy directly or indirectly [52]. HOXC8, in the same homeobox family with HOXC6, was reported to serve as a transcription activator to boost the expression of TGFβ1, leading to an increase of the proliferation, anchorage-independent growth and migration in NSCLC [53]. In CRC, BMP signaling functioned as a crucially inhibitory element in tumorigenesis [54], where HOXC8 was discovered to be a negative regulator together with smad6 [55]. Moreover, GSEA results of high and low C1/C2 ratio groups implied the contribution of TGF-β signaling pathway in clinical prognosis of CRC patients. Hence, it made sense to elucidate that the 3 genes were recruited in our cell infiltration classifier and performed well in pathologic stage and prognosis forecast. In terms of the differential expression in epithelial and T cells, it was revealed that highly transcriptional BMP family signatures including BMP5 promoted T cell infiltration in estrogen receptor-positive breast cancer [56]. However, the effect and mechanism of BMP5 expressed in T cells remained unclear.

There were still some limitations in our current study that should be considered when elucidating the results of our findings. First, an essential step might be verification with clinical single-cell data, taking into account of the distinction between bulk sequencing and single-cell technique. In addition, although the function of HOXC6 has been verified in clinical CRC patients [52], further experiments *in vivo* and *in vitro* are still required. More explorations aimed at the crosstalk of the 3 hub genes might shed light on the underlying mechanism.

Collectively, technology advance toward single-cell sequencing enhances our recognition of tumor heterogeneity. The excavation of pathologic stage-related genes facilitates the discovery of novel therapeutic targets. Our study provided a CRC classifier of pathologic stage and survival with potential clinical significance.

Conclusion

In this study, we identified pathologic evolution related genes in scRNA-Seq and proposed a novel specific cell infiltration classifier to prognosis prediction of CRC patients. Patients with upregulated BMP5 and downregulated HOXC6 and HOXC8 were related to lower pathologic stages and better prognosis. These hub genes also suggested the potentially crucial role of TGF-β signaling pathway in CRC tumorigenesis and progression.

Abbreviations

CRC: colorectal cancer; CMS: consensus molecular subtypes; TGF-β: transforming growth factor-β; scRNA-Seq: single-cell RNA sequencing; GEO: gene expression omnibus; TCGA: the cancer genome atlas; COAD: colonic adenocarcinoma; READ: rectum adenocarcinoma; RNA-Seq: RNA sequencing; TPM: Transcripts Per Million; RMA: robust multi-array average; PCA: principal components analysis; t-SNE: t-distributed stochastic neighbor embedding; SNN: shared nearest neighbor; GSEA: gene set enrichment analysis; CIBERSORT: cell-type identification by estimating relative subsets of RNA Transcripts; EMT: epithelial-mesenchymal transition; GI: gastrointestinal; CC: cervical carcinoma.

Supplementary Material

Supplementary figures and tables.
http://www.jcancer.org/v11p6861s1.pdf

Acknowledgements

We thank Yingming Sun for assistance with writing guidance.

Data availability

The datasets generated and/or analyzed during this study are available in the GEO and TCGA repository.

Funding

This study was supported by the National Natural Science Foundation of China (81773236, 81800429 and 81972852), Key Research & Development Project of Hubei Province (2020BCA069), Natural Science Foundation of Hubei Province (2020CFB612), Health Commission of Hubei Province Medical Leading Talent Project, Health Commission of Hubei Province Scientific Research Project (WJ2019H002 and WJ2019Q047), Young & Middle-Aged Medical Backbone Talents of Wuhan (WHQG201902), Application Foundation Frontier Project of Wuhan (2020020601012221), and Zhongnan Hospital of Wuhan University Science, Technology, and Innovation Seed Fund (znpy2018028, znpy2019070, znpy2019001, znpy2019048 and ZNJC201922).

Authors' contributions

J-L (substantial contributions to conception, design of single-cell analyses and data analysis), Z-Z (bioinformatics algorithms and analysis instruction), J-C (datasets acquisition and data preprocessing), X-L (data preprocessing and GSEA analysis), X-J (data preprocessing), W-S (data preprocessing), YL (validation of data analysis), J-R (validation of data analysis), YG (supervision of the study and critical revision of manuscript), CH-X (conception of the study, and design the workflow). All authors contributed to manuscript revision, read and approved the submitted version.

Competing Interests

The authors have declared that no competing interest exists.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018; 68: 394-424.
2. Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global patterns and trends in colorectal cancer incidence and mortality. Gut. 2017; 66: 683-91.
3. Stoffel EM, Murphy CC. Epidemiology and Mechanisms of the Increasing Incidence of Colon and Rectal Cancers in Young Adults. Gastroenterology. 2020; 158: 341-53.
4. Brody H. Colorectal cancer. Nature. 2015; 521: S1.
5. Cunningham D, Atkin W, Lenz HJ, Lynch HT, Minsky B, Nordlinger B, et al. Colorectal cancer. Lancet (London, England). 2010; 375: 1030-47.
6. Markowitz SD, Bertagnolli MM. Molecular origins of cancer: Molecular basis of colorectal cancer. The New England journal of medicine. 2009; 361: 2449-60.
7. Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. Nature medicine. 2015; 21: 1350-6.
8. Rohatgi PR, Mansfield PF, Crane CH, Wu TT, Sunder PK, Ross WA, et al. Surgical pathology stage by American Joint Commission on Cancer criteria predicts patient survival after preoperative chemoradiation for localized gastric carcinoma. Cancer. 2006; 107: 1475-82.
9. Cibula D, Potter R, Planchamp F, Avall-Lundqvist E, Fischerova D, Haie Meder C, et al. The European Society of Gynaecological Oncology/European Society for Radiotherapy and Oncology/European Society of Pathology guidelines for the management of patients with cervical cancer. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. 2018; 127: 404-16.
10. Wang G, McKenney JK. Urinary Bladder Pathology: World Health Organization Classification and American Joint Committee on Cancer Staging Update. Archives of pathology & laboratory medicine. 2019; 143: 571-7.
11. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Molecular cell. 2017; 65: 631-43.e4.
12. Yan L, Yang M, Guo H, Yang L, Wu J, Li R, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nature structural & molecular biology. 2013; 20: 1131-9.
13. Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJL, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nature genetics. 2017; 49: 708-18.
14. Marisa L, de Reynies A, Duval A, Selves J, Gaub MP, Vescovo L, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS medicine. 2013; 10: e1001453.
15. Kemper K, Versloot M, Cameron K, Colak S, de Sousa e Melo F, de Jong JH, et al. Mutations in the Ras-Raf Axis underlie the prognostic value of CD133 in colorectal cancer. Clinical cancer research : an official journal of the American Association for Cancer Research. 2012; 18: 3132-41.
16. Laibe S, Lagarde A, Ferrari A, Monges G, Birnbaum D, Olschwang S. A seven-gene signature aggregates a subgroup of stage II colon cancers with stage III. Omics : a journal of integrative biology. 2012; 16: 560-5.
17. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006; 38: 904-9.
18. Lenz M, Muller FJ, Zenke M, Schuppert A. Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data. Scientific reports. 2016; 6: 25696.
19. Linderman GC, Rachh M, Hoskins JG, Steinerberger S, Kluger Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. Nature methods. 2019; 16: 243-5.
20. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nature communications. 2019; 10: 5416.
21. Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics (Oxford, England). 2015; 31: 1974-80.
22. Wu W, Sun X, Gao Y, Jiang J, Cui Z, Ge B, et al. Genome-Wide De Novo Prediction of Cis-Regulatory Binding Sites in Mycobacterium tuberculosis H37Rv. PloS one. 2016; 11: e0148965.
23. Haghverdi L, Buttner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. Nature methods. 2016; 13: 845-8.
24. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102: 15545-50.
25. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics : a journal of integrative biology. 2012; 16: 284-7.
26. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nature methods. 2015; 12: 453-7.
27. Li T, Fan J, Wang B, Traugh N, Chen Q, Liu JS, et al. TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. Cancer research. 2017; 77: e108-e10.
28. Svensson V, Natarajan KN, Ly LH, Miragaia RJ, Labalette C, Macaulay IC, et al. Power analysis of single-cell RNA-sequencing experiments. Nature methods. 2017; 14: 381-7.
29. Kyrochristos ID, Roukos DH. Comprehensive intra-individual genomic and transcriptional heterogeneity: Evidence-based Colorectal Cancer Precision Medicine. Cancer treatment reviews. 2019; 80: 101894.
30. Kim C, Gao R, Sei E, Brandt R, Hartman J, Hatschek T, et al. Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. Cell. 2018; 173: 879-93.e13.
31. Kyrochristos ID, Ziogas DE, Goussia A, Glantzounis GK, Roukos DH. Bulk and Single-Cell Next-Generation Sequencing: Individualizing Treatment for Colorectal Cancer. Cancers. 2019; 11.
32. Tieng FYF, Baharudin R, Abu N, Mohd Yunos RI, Lee LH, Ab Mutalib NS. Single Cell Transcriptome in Colorectal Cancer-Current Updates on Its Application in Metastasis, Chemoresistance and the Roles of Circulating Tumor Cells. Front Pharmacol. 2020; 11: 135.
33. Tang H, Zeng T, Chen L. High-Order Correlation Integration for Single-Cell or Bulk RNA-seq Data Analysis. Frontiers in genetics. 2019; 10: 371.
34. Sun X, Sun S, Yang S. An Efficient and Flexible Method for Deconvoluting Bulk RNA-Seq Data with Single-Cell RNA-Seq Data. Cells. 2019; 8.
35. Liu H, Zhang M, Xu S, Zhang J, Zou J, Yang C, et al. HOXC8 promotes proliferation and migration through transcriptional up-regulation of TGFbeta1 in non-small cell lung cancer. Oncogenesis. 2018; 7: 1.
36. Lei H, Wang H, Juan AH, Ruddle FH. The identification of Hoxc8 target genes. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102: 2420-4.
37. Chen L, Sun DZ, Fu YG, Yang PZ, Lv HQ, Gao Y, et al. Upregulation of microRNA-141 suppresses epithelial-mesenchymal transition and lymph node metastasis in laryngeal cancer through HOXC6-dependent TGF-beta signaling pathway. Cellular signalling. 2020; 66: 109444.
38. Lowery JW, Rosen V. The BMP Pathway and Its Inhibitors in the Skeleton. Physiological reviews. 2018; 98: 2431-52.
39. Tauriello DVF, Palomo-Ponce S, Stork D, Berenguer-Llergo A, Badia-Ramentol J, Iglesias M, et al. TGFbeta drives immune evasion in genetically reconstituted colon cancer metastasis. Nature. 2018; 554: 538-43.
40. Jung B, Staudacher JJ, Beauchamp D. Transforming Growth Factor beta Superfamily Signaling in Development of Colorectal Cancer. Gastroenterology. 2017; 152: 36-52.
41. Calon A, Lonardo E, Berenguer-Llergo A, Espinet E, Hernando-Momblona X, Iglesias M, et al. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. Nature genetics. 2015; 47: 320-9.
42. Siraj AK, Pratheeshkumar P, Divya SP, Parvathareddy SK, Bu R, Masoodi T, et al. TGFbeta-induced SMAD4-dependent Apoptosis Proceeded by EMT in CRC. Molecular cancer therapeutics. 2019; 18: 1312-22.
43. Chen J, Ye X, Pitmon E, Lu M, Wan J, Jellison ER, et al. IL-17 inhibits CXCL9/10-mediated recruitment of CD8(+) cytotoxic T cells and regulatory T cells to colorectal tumors. Journal for immunotherapy of cancer. 2019; 7: 324.
44. Korkut A, Zaidi S, Kanchi RS, Rao S, Gough NR, Schultz A, et al. A Pan-Cancer Analysis Reveals High-Frequency Genetic Alterations in Mediators of Signaling by the TGF-beta Superfamily. Cell systems. 2018; 7: 422-37.e7.
45. Chen E, Yang F, He H, Li Q, Zhang W, Xing J, et al. Alteration of tumor suppressor BMP5 in sporadic colorectal cancer: a genomic and transcriptomic profiling based study. Molecular cancer. 2018; 17: 176.
46. Romagnoli M, Belguise K, Yu Z, Wang X, Landesman-Bollag E, Seldin DC, et al. Epithelial-to-mesenchymal transition induced by TGF-beta1 is mediated by Blimp-1-dependent repression of BMP-5. Cancer research. 2012; 72: 6268-78.
47. Li PD, Chen P, Peng X, Ma C, Zhang WJ, Dai XF. HOXC6 predicts invasion and poor survival in hepatocellular carcinoma by driving epithelial-mesenchymal transition. Aging. 2018; 10: 115-30.
48. Wang Y, Wang C, Liu N, Hou J, Xiao W, Wang H. HOXC6 promotes cervical cancer progression via regulation of Bcl-2. FASEB journal : official publication of the Federation of American Societies for Experimental Biology. 2019; 33: 3901-11.

49. Moon SM, Kim SA, Yoon JH, Ahn SG. HOXC6 is deregulated in human head and neck squamous cell carcinoma and modulates Bcl-2 expression. The Journal of biological chemistry. 2012; 287: 35678-88.
50. Fujiki K, Duerr EM, Kikuchi H, Ng A, Xavier RJ, Mizukami Y, et al. Hoxc6 is overexpressed in gastrointestinal carcinoids and interacts with JunD to regulate tumor growth. Gastroenterology. 2008; 135: 907-16, 16.e1-2.
51. Zhang F, Ren CC, Liu L, Chen YN, Yang L, Zhang XA. HOXC6 gene silencing inhibits epithelial-mesenchymal transition and cell viability through the TGF-beta/smad signaling pathway in cervical carcinoma cells. Cancer cell international. 2018; 18: 204.
52. Ji M, Feng Q, He G, Yang L, Tang W, Lao X, et al. Silencing homeobox C6 inhibits colorectal cancer cell proliferation. Oncotarget. 2016; 7: 29216-27.
53. Liu H, Zhang M, Xu S, Zhang J, Zou J, Yang C, et al. HOXC8 promotes proliferation and migration through transcriptional up-regulation of TGFβ1 in non-small cell lung cancer. Oncogenesis. 2018; 7: 1.
54. Hardwick JC, Kodach LL, Offerhaus GJ, van den Brink GR. Bone morphogenetic protein signalling in colorectal cancer. Nature reviews Cancer. 2008; 8: 806-12.
55. Kang M, Bok J, Deocaris CC, Park HW, Kim MH. Hoxc8 represses BMP-induced expression of Smad6. Molecules and cells. 2010; 29: 29-33.
56. Katsuta E, Maawy AA, Yan L, Takabe K. High expression of bone morphogenetic protein (BMP) 6 and BMP7 are associated with higher immune cell infiltration and better survival in estrogen receptorpositive breast cancer. Oncology reports. 2019.