Identification of stemness-related LncRNA signature for predicting the prognosis and therapy response in colorectal cancer

Bobin Ning
Chinese PLA General Hospital
Ruibao Zhu
Tsinghua University
Yonggan Xue
Chinese PLA General Hospital
Yajun Cao
Tsinghua University
Huihui Jia
Tsinghua University
Boqing Jia (baoqingjia@126.com)
Chinese PLA General Hospital

Research Article

Keywords: cancer stem cell, lncRNA, prognosis, tumor microenvironment, chemotherapy response, HOXB4

Posted Date: January 16th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2455922/v1

Abstract

Background: Cancer stem cells (CSC) carry out a vital responsibility throughout the entire progress of colorectal cancer (CRC), and fulfil an essential biological function. However, lncRNAs participate in regulating CRC stem cells (CCSCs) and correlate strongly with the patients' prognosis. Therefore, it is crucial to identify the CCRC-related lncRNAs in CRC.

Methods: We identified CCRCs-related lncRNAs through the Cell marker and TCGA databases. And the CCSC-related lncRNAs model was constructed by the differential, cox survival , and lasso regression analysis. Combining the GEO dataset, we determined the prognostic value by Kaplan-Meier analysis, univariate and multivariate cox survival analysis. Moreover, principal component analysis (PCA), clinical characterization, nomogram, gene mutation, gene set enrichment analysis (GSEA), immune microenvironment (TME), chemotherapy, intergroup differential gene, and protein-protein interaction (PPI) analysis were conducted to analyze the risk model. Furthermore, the core genes in the sub-module were comprehensively characterized.

Results: In this research, abnormally expressed, prognostic and CSC-related lncRNAs were firstly identified. Through the lasso regression model, we obtained a robust risk signature consisting of 4 CCSC-related lncRNAs (ZEB1-AS1 LINC00174 FENDRR and ALMS1-IT1). Then, the risk model was confirmed applicable in both TCGA and GEO cohorts. Further verification, the signature can be verified as a independent prognostic factor for CRC. Based on the CCSC-related lncRNA model, the high- and low-risk groups exhibited different stemness statuses, including gene expression, mutation status, signaling pathways, TME and chemotherapy response. The HOX family and HOX4 were centrally located in the PPI interaction and had an influential contribution in CRC.

Conclusions: We established a 4 CCSC-related lncRNA signature with a promising prognosis. And the signature can appropriately estimate the gene mutation, TME, and chemotherapy outcomes for CRC patients. Furthermore, the CCSC-related lncRNAs and HOX4 can serve as noble biomarkers and promote the management of therapy clinically.

Introduction

Colorectal cancer (CRC) is the third most common cancer in the world, and the fatality rate has risen to the second place[1]. It is widely accepted that CRC originate from the abnormal development of crypt-derived polyps[2]. At the cellular level, the origin cells of CRC are thought to be stem cells or stem cell-like cells, also known as cancer stem cells (CSCs)[3].

CSCs are a small fraction of cells bearing stem-like properties within tumor tissues. This concept was proposed about 160 years ago, but limited by the technology at the time, scientists did not find procure CSCs until Bonnet isolated it in acute myeloid leukemia (AML) in 1997[4]. Due to the widespread presence of CSCs across cancers, increasing attention has been paid to exploring their biological function. In the context of CRCs, O'Brien and Ricci-Vitiani first reported the existence of CSCs in 2007[5]. According to

previous investigations, the defining markers of CSCs in CRC mainly include Lgr5, CD133, CD44, CD29, CD166, EpCAM, etc[6]. It is well illustrated that the existence of CSCs is an essential cause of tumor relapse and metastasis[7]. Studies have shown that the expression of Lgr5, CD44 and EpCAM leads to cells possessing a high tumorigenic ability, and the expression of CD44/CD166 is positively correlated with lymphatic metastasis as well as liver and lung metastasis in CRC patients[8]. In addition, CSCs also have the characteristics of promoting drug resistance[9]. It has been reported that the drug resistance mechanisms of CSCs include ABC transporter-mediated chemotherapeutic drug efflux, enhanced ALDH activity and reactive oxygen species scavenging, activation of pro-survival pathways, and more efficient DNA repair[10]. All these biological processes contribute to poorer prognosis for patients. Therefore, the deepening insight into the characteristics of CSCs derives benefits for the clinical treatment of CRC patients.

CSCs can modulate their surrounding microenvironment, achieving self-renewal and maintenance[11]. Within the complex interaction of CSCs and tumor microenvironment, lncRNA as a crucial regulatory factor, is implicated in the regulation of many characteristics of tumors. Studies have shown that lncRNA EPIC1 can interact with MYC through its 129–283 nt region, jointly regulating the transcription of MYC target genes and, thus, promoting the cell cycle progression of tumor[12]. The lncRNA-MVIH has been reported to facilitate tumor angiogenesis by inhibiting the secretion of PGK1[13]. In liver cancer, TGF-β can induce the expression of lncRNA-ATB, up-regulate the expression of epithelial-mesenchymal transition (EMT)-related protein ZEB1/2 and promote the invasive ability of liver cancer[14]. It has been reported that lncRNAs can also inhibit the activity of killer T cells to indirectly mediate tumor immune escape mechanisms. In addition to the tumor-promoting effects, some types of lncRNAs can also impede tumor progression. Experiments have proved that lncRNA-MEG3 plays a tumor suppressor role by activating the expression of p53[15]. The lncRNA-GAS5 can bind to the DNA-binding domain of glucocorticoid receptors, thereby blocking the action of glucocorticoids and promoting tumor cell apoptosis[16]. Given the essential role of lncRNAs in tumor regulation and different types of lncRNAs mediating different biological functions, it is particularly important to investigate the activity of lncRNAs in the biology of CSCs.

In this study, we identified the CCSC-related lncRNAs by TCGA and Cell marker database, and established a CSC-lncRNA risk model composed of 4 lncRNAs (ZEB1-AS1, LINC00174, FENDRR and ALMS1-IT1) by the Lasso regression model. The clinical and mutation information were also retrieved from TCGA database and Gene Expression Omnibus (GEO) database to validate the CSC-related lncRNA signature. Our works showed that the risk model operate well in predicting the survival outcomes of CRC patients, and is highly correlated with gene mutation, drug resistance, and tumor microenvironment landscape. In addition, we focused on HOX family genes involving the cancer stemness through Metascape enrichment analysis and PPI network interaction analysis. Our observations indicated that HOXB4 is highly correlated with CSC markers and is involved in regulating the tumor microenvironment and the infiltration of immune cells in CRC. These results suggested that the four stem-related lncRNAs and HOX family are potential biomarkers and therapeutic targets for CRC.

Methods

Datasets

The data of the experimental group was collected from Colon Adenocarcinoma (COAD) cohorts in The Cancer Genome Atlas (TCGA), including the RNA-seq, mutation matrix and clinical information. The data of validation group was retrieved from the GSE39582 dataset in Gene Expression Omnibus (GEO) database. R package was used to annotate transcriptome data and distinguish mRNA and lncRNA, then extract the expression matrix of lncRNAs in COAD. Twenty-two cancer stem cell (CSC) genes were obtained from the CellMarker database (http://bio-bigdata.hrbmu.edu.cn/CellMarker/), and a matrix of CSC-related lncRNAs were acquaired by Pearson correlation analysis (Supplementary 1). Co-expression analysis and visualization were performed, and the CSC-related lncRNAs were defined by coefficient $|r| > 0.4$ and $p < 0.001$ as the screening criteria. Ethical approval is not required according to the guidelines of these two databases.

Differential and survival analysis of the CCSC-related lncRNAs in COAD

In the TCGA database, the R packages "Limma" and "phearmap" were used to perform differential analysis of CSC-related lncRNAs between the colon cancer and adjacent tissues, $|logFC|= 0.585$, $FDR=0.05$ was considered statistically significant. Next, the “survival” and “Glmnet” packages were used to perform univariate Cox regression analysis on the differential CSC-related lncRNAs for the prognosis of CRC patients. Then, the the core CSC-related lncRNAs of CRC patients were obtained.

Construction and evaluation of a CRC prognostic model based on CCSC-related lncRNAs

From the TCGA and GEO database, we extracted the transcriptome datas as the Train group and the Test group respectively. The Train group (TCGA) was used to construct the prognostic model and then the Test group was applied to verifying the accuracy. We first constructed the LASSO regression model through the "glmnet" R package, and took the point with the smallest cross-validation error. The lncRNAs corresponding to this point is the prognostic signature. The CSC-related lncRNAs model can be acquired by adding up each lncRNA multiplied by its corresponding coefficient (Table 1). According to the expression of the signature lncRNAs, we can calculate the CCSC-related lncRNA score of the individual patient, and then we divided the patients into high- and low- risk groups based on the median value. Similarly, the GEO cohort was also divided into two groups according the CCSC-related lncRNA score. Then, PCA analysis was used to validate the accuracy of the classification. To evaluate the prognostic predication of the CSC-related lncRNA signature, we made the Kaplan-Meier curves of overall survival (OS) in the Train group. Furthermore, the OS analysis in the Test group and the disease-free survival (DFS) in the Train group were also performed. Not only that, the univariate and multivariate cox survival analysis, as well as the time-dependent and clinical trait-dependent receiver operating characteristic (ROC) curves were conducted to evaluate the correlation between the CSC-related lncRNA models and clinical prognosis. The results were visualized by “limma”, “ggplot2”, “survival”and

“survminer” R packages. The differences of the clinical properties, including age, gender, stage and TNM stage, between the high- and low- risk groups were thoroughly revealed.

Concerning the age, TNM stage, and clinical grade and CSC-related lncRNA signature, a nomogram was comprehensively established for the CRC patients. On this basis, the calibration curves was determined by the "survival", "regplot" and "rms" R packages. In the nomogram, a nomogram score was gained according to the personal rating scale. Then, the scores for all clinical characteristics were summed to obtain the patient's composite risk score. We evaluated the 1, 3, and 5 year survival rates against the scale of the composite risk score above. And the calibration curves for 1, 3, and 5 years were also published for the demonstration. The closer the calibration curve is to the gray line, the more accurate the prediction from the nomogram. "Survival", "survminer", and "timeROC" R packages were used to draw the ROC curve of the nomogram. The area under the ROC curve represented the accuracy of survival predication. Nest, "survival" R package was run to check the ability of the nomogram as an independent prognostic factor, P values less than 0.05 were determined to be independent prognostic factors.

Mutation signature and functional enrichment analysis

We carried out GSEA enrichment analysis, immune microenvironment analysis, gene mutation signature analysis, immune function difference analysis, KEGG pathway analysis, and microsatellite instability difference analysis respectively between high and low CCSC-related lncRNA score groups. The R package “limma”, ”reshape2”, ”ggpubr”, “GSEABase”, ”GSVA”, ”pheatmap”, ”pRRophetic”, and ”ggplot2” were applied for the above researches.

Chemotherapy and immunotherapy analysis

CSCs are considered as a drug-resistant cancer cell population, which is the primary cause of the therapeutic failure. To determine the sensitivity with different CCSC realted-lncRNA score, the drug sensitivity analysis was performed using the R packages "limma", "ggpubr", "PRrophic" and "ggplot2". In addition, differences of CSC-lncRNA scores between the mutant versus wild-type genes were also analyzed, which can be assessed to the guidance of the immunotherapy. **Protein-protein interaction (PPI) network analysis and identification of hub genes in CSC subsets**

To begin with, we conducted the deferential expression analysis between the high- and low- risk groups. Further, the up-regulated and down-regulated genes were respectively analyzed and visualized through the Matascape website (https://metascape.org/gp/index.html#/main/step1). In addition, the DEGs were input to construct a PPI network on the String website (https://string-db.org/). For systematically analyze the interaction between proteins, we used the MCOD plug-in of the cytoscape software to modular the functionally similar genes. And the cytohubba plug-in was utilized to focus the most essential genes. After we focused the critical gene, a series of single-gene analyses were performed, which contained differential expression analysis, clinical correlation analysis, nomogram analysis, calibration analysis, survival analysis, stratified survival analysis, CSC genes correlation analysis, and immune microenvironment analysis.

Results

Identification of CCSC-related lncRNAs in CRC

According to the screening conditions, we isolated a total of 1768 CCSC-related lncRNAs from the TCGA database and displayed the correlation of CSC markers through a Sankey diagram (Figure 1A). EdgeR and the "DESeq2" R packages were used to screen 669 differential lncRNAs as differentially expressed CCSC-related lncRNAs, of which 104 lncRNAs were down-regulated and 565 lncRNAs were up-regulated (Figure 1 B, C). Univariate Cox regression analysis showed that a total of 5 lncRNAs were identified to be closely associated with patients' survival in CRC, including LINC00174, ZEB1-AS1, MCM3AP-AS1, ALMS1-IT1 and FENDRR. Among the above core CCSC-related lncRNAs, only FENDRR (HR=0.651, p=0.043) was a protective factor for CRC patients, and the remaining four CCSC-related lncRNAs were risk factors for CRC patients (Figure 1D).

Establishment of a prediction model for CCSC-related lncRNAs in CRC

Based on Lasso regression model and iterative analysis of cross-validation method, we constructed a CCSC-related lncRNA risk model composed of 4 lncRNAs (ZEB1-AS1, LINC00174, FENDRR and ALMS1-IT1) (Figure 2 A, B). According to the risk score formula, we took the median of the score (2.25) as the cutoff value to classify the CRC patients. Following that, the TCGA and GEO cohorts were separated into high- and low- risk groups, respectively. In the PCA distribution diagram, the red scatter points (high-risk) and the blue scatter points (low-risk) can be distinguished effectively (Figure 2C). In the training group, survival analysis indicated that the OS of high-risk CRC patients was both inferior than the low-risk group ($p<0.001$). Consistently, the OS analysis in the test group and the PFS analysis in the train group all presented the same results ($p<0.001$) (Figure 2 D~F). In the cox univariate analysis, age, T stage, N stage, M stage, and CSC-lncRNA risk score were all associated with the prognosis of CRC patients ($p < 0.001$). By multivariate Cox regression analysis, the above factors can still served as survival independent factors against the other clinical characters (Figure 3 A, B). The CCSC-related lncRNA model was further evaluated by ROC curve analysis, and the results suggested that this model had the best accuracy in predicting the 5-year survival of CRC patients, with AUC=0.751 (Figure 3C). Notably, our CCSC-related lncRNA model had better predictive ability than any other clinical features (Figure 3D). In addition, clinical traits such as distant metastasis, lymph node metastasis, and T stage had significant statistical differences between high- and low- risk groups (Figure 3E).

According to the scoring scale in Figure 4A, a clinically characteristic score was applied. The patient's outcome can be estimated based on the composite scoring scale. In Figure 4B, The distance of the calibration curve from the grey line indicated that the accuracy of predicting the surival rates. As we can seen, the 5-year survival rate of patients is comparatively accurate. ROC curves and univariate Cox regression analysis further confirmed the reliability of this nomogram (Figure 4C, 4D). The AUC comparison results showed that the nomogram (0.796) is prominently higher than other clinical traits (Figure 4D). Therefore, our model has strong reliability and applicability for the survival prediction of CRC patients.

Mutation signature and functional enrichment analysis

We conducted the GSEA enrichment analysis between the high- and low- risk groups. The top 5 enrichment signaling pathways of the up-regulated and down-regulated genes were respectively listed in Figure 5A and 5B. As the results showed, cell adhesion molecules, cytokine receptor interaction and ECM receptor interaction were significantly activited in the high-risk group. Among the differences of 22 types of immune cells, B cells naive was statically increased in the high-risk group (Figure 5C). Intriguingly, most immune-related functions are suppressed at high-risk group, such as stimulation of antigen-presenting cells (APCs), chemokine receptors (CCRs), checkpoints, inflammation, stimulation of T cells, and type II interferons (IFNs) (Figure 5F). The above investigations re-confirmed that the CSCs can promote tumor progression by regulating the immune microenvironment. Mutation signature analysis listed the top 20 genes with the highest mutation frequency (Figure 5D, E). The results demonstrated that the mutation frequency of *TP53 (a classic tumor suppressor gene)* in the high-risk group was higher than the low-risk group. Whereas, the mutation frequencies of the other 19 oncogenes were lower than the low-risk group. In addition, GSVA analysis indicated that the cancer related pathways, such as P53 signaling pathway, cell cycle, DNA replication, PPAR signaling pathway, apoptosis death and biological metabolism, possessed a higher activity in the high-risk group (Figure 5G). Moreover, the microsatellite instability of the low-risk group is more unstable (Figure 5H), suggesting that the low-risk group may be more sensitive to immunotherapy.

Analysis of chemotherapy and immunotherapy in CCSC-related lncRNA risk model

Consistent with conventional cognition, the IC50s of the chemotherapy drugs (Fluorouracil, Bleomycin, Gemcitabine and Sunitinb) were all positively correlated with the CSC risk scores of patients (Figure 6A~D). The IC50 of the high-risk group is significantly higher than that of the low-risk group (Figures 6E~6H). These results indicated that the high-risk group is less sensitive to the effects of chemotherapy drugs and targeted therapy, which may be related to the consequence of CSCs. We also analyzed the correlation between risk scores and gene mutation status (Figure 6I~6L). To further validate our observations, we selected four genes among the genes with high mutation frequency including *BIRC6*, *PIK3CA*, *SOX9*, and *TP53*. Patients with wild-type *BIRC6*, *PIK3CA*, and *SOX9* consistently had higher risk scores than mutants (Figure 6I~K). Conversely, patients with wild-type *TP53* had lower risk scores than those with mutant *TP53* (Figure 6L), which is consistent with previous findings. Therefore, intervention strategies targeting CSC are critical in oncology.

Metascape enrichment analysis, PPI network and *Hub* genes

To further explore the mechanism of CSC in each subgroup, we performed Metascape enrichment analysis of DEGs between the high- and low- risk groups. The results indicated that the over-expressed genes in the low-risk group were mainly involved in immune regulatory responses. Defense response to bacterium, regulation of leukocyte-mediated immunity and adaptive immune response were mainly enriched (Figure 7A). The over-expressed genes in the high-risk group are not only involved in cancer

pathways (DNA damage, telomere stress-induced senescence, histone H_3K_{27} trimethylation, and negative regulation of epithelial to mesenchymal), but also in the CSC-related pathways growth (ovarian follicle development, and ncRNAs involved in STAT3 signaling in hepatocellular carcinoma) (Figure 7B). The interactions of the DEGs between the CSC-subsets were obtained from the String database and visualized by Cytoscape software (Figure 7C). The top 10 genes at the core position in the network were acquired by the cytohubba plugin: *FABP1*, *GUCA2A*, *PNISR*, *AGT*, *HOXB4*, *CLCA4*, *GUCA2B*, *ZG16*, *AQP8* and *HOXA5* (Figure 7D). The HOX family members are remarkably clustered into a relatively complete module and their expression is up-regulated in the high-risk group. Interestingly, *HOXB4* is at the center of this module.

On this basis, we found that the expression of *HOXB4* gene was not only higher in tumor tissue, but also in the high age group, and its expression level also exhibited a increasing trend with the progression of T stage (Figure 8A~8D). Nomogram and calibration analysis of *HOXB4* revealed that an accurate prediction was met at 5-year survival (Figure 8E, F). The COAD cohort showed that, in terms of PFS, DSS, OS, over-expressed *HOXB4* indicate a worse prognosis ($P<0.05$). And the predictive ability consisted a statistical difference in both early and advanced stages (Figure 8G~8L). These above results suggested that *HOXB4* has a predictive potential for CRC patients. Moreover, the correlation analysis between *HOXB4* and CSC genes revealed that *HOXB4* was positively correlated with *CD44*, *ALCAM*, *PROM1*, *ITGB1* and *SOX2* (Figure 8M). Analysis of the tumor immune microenvironment showed that patients with abundant *HOXB4* have increased infiltration of immune cells, including T cells, $CD8^+$ T cells, macrophages, NK cells and TH1 cells (Figure 8N). Interestingly, *HOXB4* is positively correlated with the infiltration of the vast majority of immune cells, except Th17cells (Figure 8O). Finally, the single-cell RNA sequencing (scRNA-seq) analysis of CRC showed that *HOXB4* mainly exist in NK cells (Figure 8P~Q). Therefore, *HOXB4* is expected to be one of the candidate genes for potential biomarkers of CRC.

Discussion

Drug resistance and metastasis are two critical characteristics of tumors that lead the tumor difficult to cure[17]. Tumors are composed of heterogeneous cell subsets that display various responses to different treatments[18]. Among them, CSC subsets are pluripotent, capable of self-renewal, highly resistant to cytotoxic therapy, and promote tumorigenesis[19]. Therefore, a comprehensive understanding of the molecular biological mechanism characteristic of CSCs will help to develop new specific targeted therapies to eradicate CSCs. As a class of non-coding RNAs, the functions mainly rely on their RNA expression, which provides a solid foundation for studying their potential functions by RNA-seq and bioinformatics. Studies have shown that lncRNAs play an important role in the regulation of CSCs and their surrounding microenvironment.

The CSC markers in CRC were achieved in the Cell Marker datasets, which is a dedicated collection of cell sorting markers that have been used to define. For identifying CSC-related lncRNAs, we firstly isolated a total of 1768 lncRNAs correlated with CSC markers from the RNA-seq expression profiles of COAD, and retained the lncRNAs sequenced simultaneously in the GEO cohort. Differential expression analysis and

univariate Cox regression analysis screened out five core lncRNAs, including LINC00174, ZEB1-AS1, MCM3AP-AS1, ALMS1-IT1 and FENDRR (Fig. 1A ~ D). According to existing research, LINC00174, ZEB1-AS1, ALMS1-IT1 have been reported in CRC, breast cancer, liver cancer and lung cancer and other tumors. Among them, ALMS1-IT1 can promote the malignant progression of lung adenocarcinoma by activating the cyclin-dependent kinase pathway[20], while LINC00174[21] and ZEB1-AS1[22] have been reported to promote tumor proliferation, metastasis and drug resistance by regulating microRNA. In contrast, the expression of FENDRR is absent in most tumors and often acts as a tumor suppressor. Not only that, FENDRR has been reported to directly inhibit the stemness of CRC through the SOX family (SOX2/SOX4) [23]. For the first time, we link the CCSC-related lncRNAs to the recognized markers of CRC stem cells.

We constructed a CCSC-related lncRNA model by Lasso regression model and iterative analysis with cross-validation method. Through PCA analysis, we could clearly observe that the model can significantly divide the patients into two different CSC-subtypes (Fig. 2C). According to the median value of the CCSC-lncRNA risk score, we divided the two different cohorts into high- and low- risk groups. In the Train group (TCGA), both OS and PFS in the high-risk group suffered a deteriorated prognosis ($p < 0.001$). The predictive power remained robust in the validation cohort (GEO) (Fig. 2D ~ F). Multivariate cox regression analysis and ROC curve analysis further confirmed the risk model. And the ability of the CSC-signature was superior than any other clinical characteristics including age, gender and TNM (Fig. 3D). Furthermore, this signature can serve as a prognostic factor in CRC, independent of other clinical features. In the nomogram constructed after overall consideration of other clinical information, the AUC of this model reached 0.796, far exceeding the models constructed by Xu (0.701) and Li (0.745).

To further explore the reasons for the significant clinical prognostic disparity in patients between CSC subtypes of CRC, we performed mutational signature and functional enrichment analysis on high and low risk groups of CRC patients identified by the CCSC-related lncRNA model. (Fig. 5D, 5E). It is well known that mutated *TP53* in CRC not only loses its tumor suppressor function, but the mutated protein may drive oncogenic mechanisms in tumorigenesis[24]. In this project, we found that patients of the high-risk group correspond to a higher frequency of *TP53* mutation, which reflects that patients with high tumor stemness can lose tumor suppressor function along with *TP53* mutation. Studies have revealed that *TP53* mutants can drive the expression of CSC genes, which may also be one of the reasons for the higher frequency of *TP53* mutations in patients of the high-risk group in our model. *PIK3CA* is also a common mutation in CRC[25], and in our model, the low-risk group had a lower mutation frequency of *PIK3CA* than the high-risk group, which is consistent with previous reports that patients with PIK3CA mutations have a better prognosis. Apoptotic proteins are also a research hotspot in cancer. Previous studies have shown that in CRC, overexpression of BIRC6 promotes tumor growth and invasion, and is associated with poorer overall survival[26]. In our analysis, the mutation frequency of BIRC6 in the low-risk group was higher than that in the high-risk group (Fig. 6I), which also indicate its promoting role in CRC. In our analysis of the tumor microenvironment, many immune-related functions, such as checkpoint, inflammatory response, type I and type II IFN, etc. were inhibited in the high-risk group of our model (Fig. 5F). This result also confirmed that CSCs can escape the killing function of the immune system by regulating the immune microenvironment and immune cells surrounding them, thereby maintaining and

promoting their growth. In summary, CSC-related lncRNA risk model we established can provide a new platform for the study of immune regulation in the tumor microenvironment and immunotherapy.

5-Fluorouracil is the first-line drug for clinical chemotherapy in patients with CRC today[27]. We analyzed the resistance of patients to 5-fluorouracil in the high- and low-risk groups of our model by pRRophetic package. It is reasonable that patients in the high-risk group which means higher tumor cell stemness have higher IC50 for 5-fluorouracil and are less sensitive to the effect of chemotherapy. We also analyzed two other classic chemotherapy drugs (bleomycin, gemcitabine) and a targeted drug (sunitinib), and patients in the high-risk group have the same trend with 5-fluorouracil for drug resistance (Fig. 6A-D). The reason behind this may be related to a series of drug resistance mechanisms induced by CSCs, which still needs further investigation. All these results indicate that the risk model we established has clinical significance and can be used to formulate more personalized treatment for patients with CRC.

The occurrence of cancerification is accompanied by abnormal regulation of important intrinsic cell signaling. Wnt, Notch and Hedgehog are three highly conserved signaling pathways that affect cell proliferation, differentiation and fate determination. CSCs are usually followed by persistent activation of one or more of these signaling pathways. Mutations in the *APC* gene are found in most CRC patients, which can lead to ectopic activation of the WNT signaling pathway, resulting in excessive proliferation of stem cells and reducing cell adhesion, which is conducive to tumor migration and metastasis[28]. In our Metascape enrichment analysis, the abnormal genes in the high-risk group were activated in cell growth, DNA damage, and EMT (Fig. 7B), all of which were highly correlated with cancer progression. The over-expressed genes in the low-risk group were mainly involved in metabolism and immune regulation-related signaling pathways (Fig. 7A). The differences of the regulation between low- and high-risk groups are relied on the complex microenvironment, which needs to be further studied.

Surprisingly, PPI network analysis Bring us to the attention of the *HOX* gene family. Between the CSC-lncRNAs subgroups, *HOXB4* is at the center of this network module. Previous studies have found that *HOXB4* can promote the self-renewal of hematopoietic stem cells (HSCs), and this function is achieved through the activation of the WNT signaling pathway[29–31]. However, the regulatory mechanism of *HOXB4* in CSCs of CRC is not clear. Our correlation analysis showed that *HOXB4* is highly correlated with typical markers of CSCs of CRC (Fig. 8M), and its role remains to be further explored. The results of single-gene analysis showed that high expression of *HOXB4* was highly associated with worse prognosis of CRC patients, and this phenomenon persisted at different stages of tumor development (Fig. 8G-8L), which is consistent with previous studies reporting that high *HOXB4* expression promotes cell proliferation and migration, drives cell cycle progression, and is associated with poorer survival probability. In addition, studies have shown that *HOXB4* is involved in immune infiltration, especially in tumor-associated macrophages and cancer-associated fibroblasts[32]. Our analysis of the tumor immune microenvironment indicated that patients with high *HOXB4* expression have high infiltration in some immune cells (such as CD8 T cells, macrophages, NK cells, etc.), and were positively correlated with the degree of infiltration of most immune cells (Fig. 8N-O). In conclusion, *HOXB4* can serve as a potential biomarker for CRC to assist in the diagnosis and prognosis of patients.

Conclusion

In summary, we identified four critical lncRNAs (LINC00174, ZEB1-AS1, MCM3AP-AS1, ALMS1-IT1) related to the CSC. And the CCSC-related lncRNA model can reliably predict the prognosis of CRC patients and apply to evaluate the chemotherapeutic responses, TME, and mutation alternations, therefore optimizing the prognosis and treatment. In addition, we also found that HOXB4 can be related with CSC, and a potential marker for the CRC patients.

Abbreviations

TCGA The Cancer Genome Atlas

GEO Gene Expression Omnibus

CRC colorectal cancer

CSC cancer stem cell

CCSC colorectal cancer stem cell

AUC areas under the curve

lncRNA long non-coding RNA

TME the tumor microenvironment

OS overall survival,

DSS disease specific survival

PFI progression free interval

PPI protein-protein interaction

Declarations

Authors' contributions

BN: Methodology & formal analysis; RZ: writing—original draft; YX: Validation; YC: software & editing; HJ: writing—review; HL: Resources, data curation, visualization; BJ: Conceptualization, supervision, project administration. All authors read and approved the final manuscript.

Funding

None.

Availability of data and materials

The materials analyzed during the study are available in COAD cohort from the TCGA dateset (https://portal.gdc.cancer.gov/) and in GSE39582 dateset from the GEO dataset (https://www.ncbi.nlm.nih.gov/geo/).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

None.

References

1. Dekker E, Tanis PJ, Vleugels J, Kasi PM, Wallace MB. Colorectal cancer. Lancet. 2019. 394(10207): 1467-1480.
2. Ponz de Leon M, Di Gregorio C. Pathology of colorectal cancer. Dig Liver Dis. 2001. 33(4): 372-88.
3. Das PK, Islam F, Lam AK. The Roles of Cancer Stem Cells and Therapy Resistance in Colorectal Carcinoma. Cells. 2020. 9(6).
4. Thomas D, Majeti R. Biology and relevance of human acute myeloid leukemia stem cells. Blood. 2017. 129(12): 1577-1585.
5. O&#x27, Brien CA, Pollett A, Gallinger S, Dick JE. A human colon cancer cell capable of initiating tumour growth in immunodeficient mice. Nature. 2007. 445(7123): 106-10.
6. Munro MJ, Wickremesekera SK, Peng L, Tan ST, Itinteang T. Cancer stem cells in colorectal cancer: a review. J Clin Pathol. 2018. 71(2): 110-116.
7. de Sousa e Melo F, Kurtova AV, Harnoss JM, et al. A distinct role for Lgr5(+) stem cells in primary and metastatic colon cancer. Nature. 2017. 543(7647): 676-680.
8. Wang Z, Tang Y, Xie L, et al. The Prognostic and Clinical Value of CD44 in Colorectal Cancer: A Meta-Analysis. Front Oncol. 2019. 9: 309.
9. Katoh M. Canonical and non-canonical WNT signaling in cancer stem cells and their niches: Cellular heterogeneity, omics reprogramming, targeted therapy and tumor plasticity (Review). Int J Oncol. 2017. 51(5): 1357-1369.

10. Van der Jeught K, Xu HC, Li YJ, Lu XB, Ji G. Drug resistance and new therapies in colorectal cancer. World J Gastroenterol. 2018. 24(34): 3834-3848.
11. Jahanafrooz Z, Mosafer J, Akbari M, Hashemzaei M, Mokhtarzadeh A, Baradaran B. Colon cancer therapy by focusing on colon cancer stem cells and their tumor microenvironment. J Cell Physiol. 2020. 235(5): 4153-4166.
12. Wang Z, Yang B, Zhang M, et al. lncRNA Epigenetic Landscape Analysis Identifies EPIC1 as an Oncogenic lncRNA that Interacts with MYC and Promotes Cell-Cycle Progression in Cancer. Cancer Cell. 2018. 33(4): 706-720.e9.
13. Wang Y, Wu Y, Xiao K, et al. RPS24c Isoform Facilitates Tumor Angiogenesis Via Promoting the Stability of MVIH in Colorectal Cancer. Curr Mol Med. 2020. 20(5): 388-395.
14. Yuan JH, Yang F, Wang F, et al. A long noncoding RNA activated by TGF-β promotes the invasion-metastasis cascade in hepatocellular carcinoma. Cancer Cell. 2014. 25(5): 666-81.
15. Wei GH, Wang X. lncRNA MEG3 inhibit proliferation and metastasis of gastric cancer via p53 signaling pathway. Eur Rev Med Pharmacol Sci. 2017. 21(17): 3850-3856.
16. Liu X, She Y, Wu H, Zhong D, Zhang J. Long non-coding RNA Gas5 regulates proliferation and apoptosis in HCS-2/8 cells and growth plate chondrocytes by controlling FGF1 expression via miR-21 regulation. J Biomed Sci. 2018. 25(1): 18.
17. Smith AG, Macleod KF. Autophagy, cancer stem cells and drug resistance. J Pathol. 2019. 247(5): 708-718.
18. Prasetyanti PR, Medema JP. Intra-tumor heterogeneity from a cancer stem cell perspective. Mol Cancer. 2017. 16(1): 41.
19. Yin W, Wang J, Jiang L, James Kang Y. Cancer and stem cells. Exp Biol Med (Maywood). 2021. 246(16): 1791-1801.
20. Luan T, Zhang TY, Lv ZH, et al. The lncRNA ALMS1-IT1 may promote malignant progression of lung adenocarcinoma via AVL9-mediated activation of the cyclin-dependent kinase pathway. FEBS Open Bio. 2021. 11(5): 1504-1515.
21. Ma Y, Li Y, Tang Y, Tang N, Wang D, Li X. LINC00174 Facilitates Proliferation and Migration of Colorectal Cancer Cells via MiR-3127-5p/ E2F7 Axis. J Microbiol Biotechnol. 2021. 31(8): 1098-1108.
22. Jin Z, Chen B. LncRNA ZEB1-AS1 Regulates Colorectal Cancer Cells by MiR-205/YAP1 Axis. Open Med (Wars). 2020. 15: 175-184.
23. Zhao X, Wu J, Li Y, Ye F, Wang C. Long non-coding RNA FENDRR inhibits the stemenss of colorectal cancer cells through directly binding to Sox2 RNA. Bioengineered. 2021. 12(1): 8698-8708.
24. Tsilimigras DI, Ntanasis-Stathopoulos I, Bagante F, et al. Clinical significance and prognostic relevance of KRAS, BRAF, PI3K and TP53 genetic mutation analysis for resectable and unresectable colorectal liver metastases: A systematic review of the current evidence. Surg Oncol. 2018. 27(2): 280-288.

25. Mei ZB, Duan CY, Li CB, Cui L, Ogino S. Prognostic role of tumor PIK3CA mutation in colorectal cancer: a systematic review and meta-analysis. Ann Oncol. 2016. 27(10): 1836-48.
26. Wolff RK, Hoffman MD, Wolff EC, et al. Mutation analysis of adenomas and carcinomas of the colon: Early and late drivers. Genes Chromosomes Cancer. 2018. 57(7): 366-376.
27. Blondy S, David V, Verdier M, Mathonnet M, Perraud A, Christou N. 5-Fluorouracil resistance mechanisms in colorectal cancer: From classical pathways to promising processes. Cancer Sci. 2020. 111(9): 3142-3154.
28. Fodde R. The APC gene in colorectal cancer. Eur J Cancer. 2002. 38(7): 867-71.
29. Elcheva IA, Wood T, Chiarolanzio K, et al. RNA-binding protein IGF2BP1 maintains leukemia stem cell properties by regulating HOXB4, MYB, and ALDH1A1. Leukemia. 2020. 34(5): 1354-1363.
30. Lei D, Yang WT, Zheng PS. HOXB4 inhibits the proliferation and tumorigenesis of cervical cancer cells by downregulating the activity of Wnt/β-catenin signaling pathway. Cell Death Dis. 2021. 12(1): 105.
31. Moore MA, Shieh JH, Lee G. Hematopoietic cells. Methods Enzymol. 2006. 418: 208-42.
32. Wang L, Jin H, Zeng Y, et al. HOXB4 Mis-Regulation Induced by Microcystin-LR and Correlated With Immune Infiltration Is Unfavorable to Colorectal Cancer Prognosis. Front Oncol. 2022. 12: 803493.

Supplementary 1

Supplementary 1 is not available with this version.

Figures

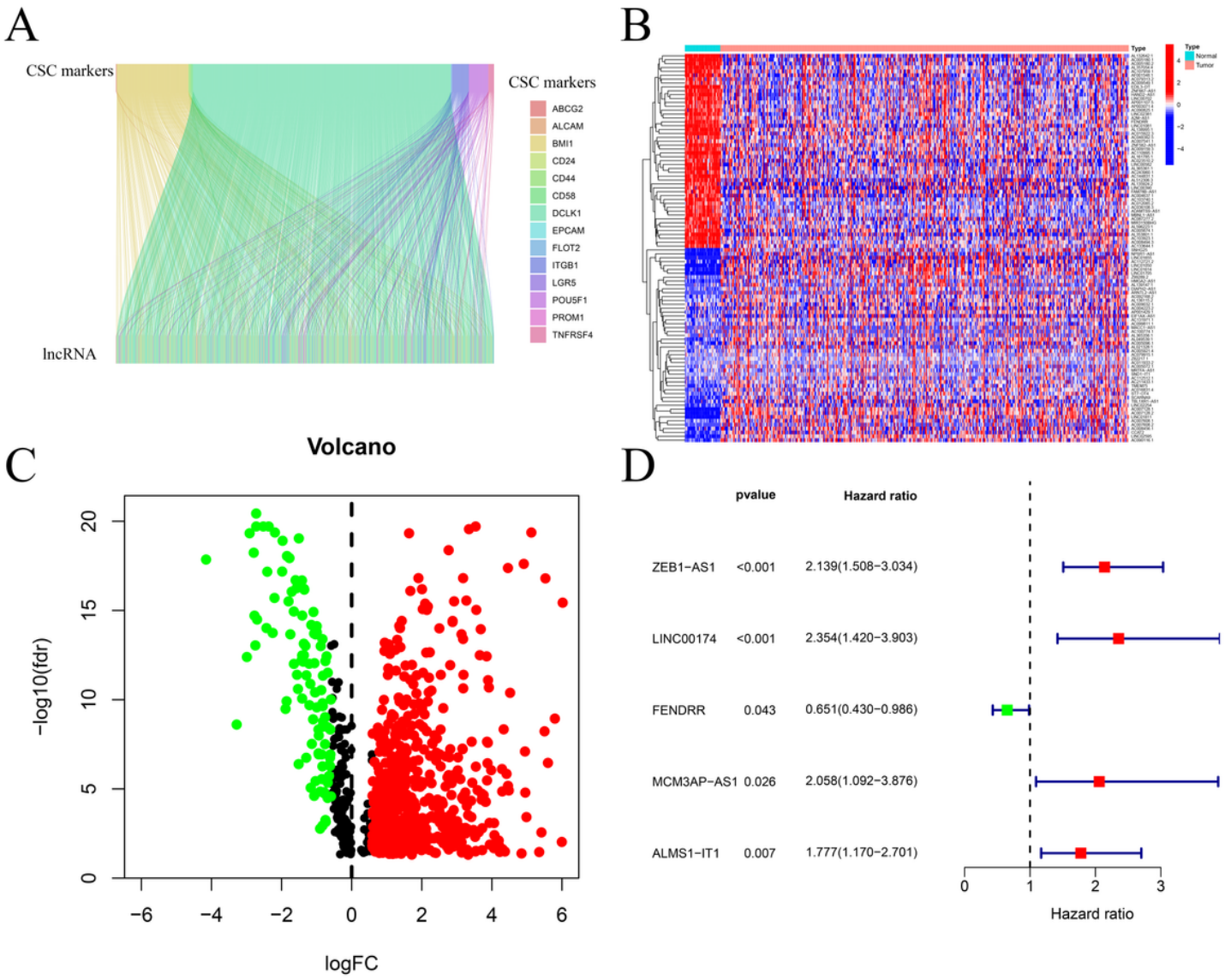

Figure 1

Identification of abnormally expressed, prognostic, cancer stem cell (CSC) related lncRNAs in colorectal cancer (CRC). (A) The ggalluvial diagram of the CSC-related lncRNAs and the CSC markers in CRC. (B) The heap-map of the differential expressed CSC-related lncRNAs. (C) The volcano of the differential expressed CSC-related lncRNAs. (D) The univariate cox analysis of the differential expressed CSC-related lncRNAs.

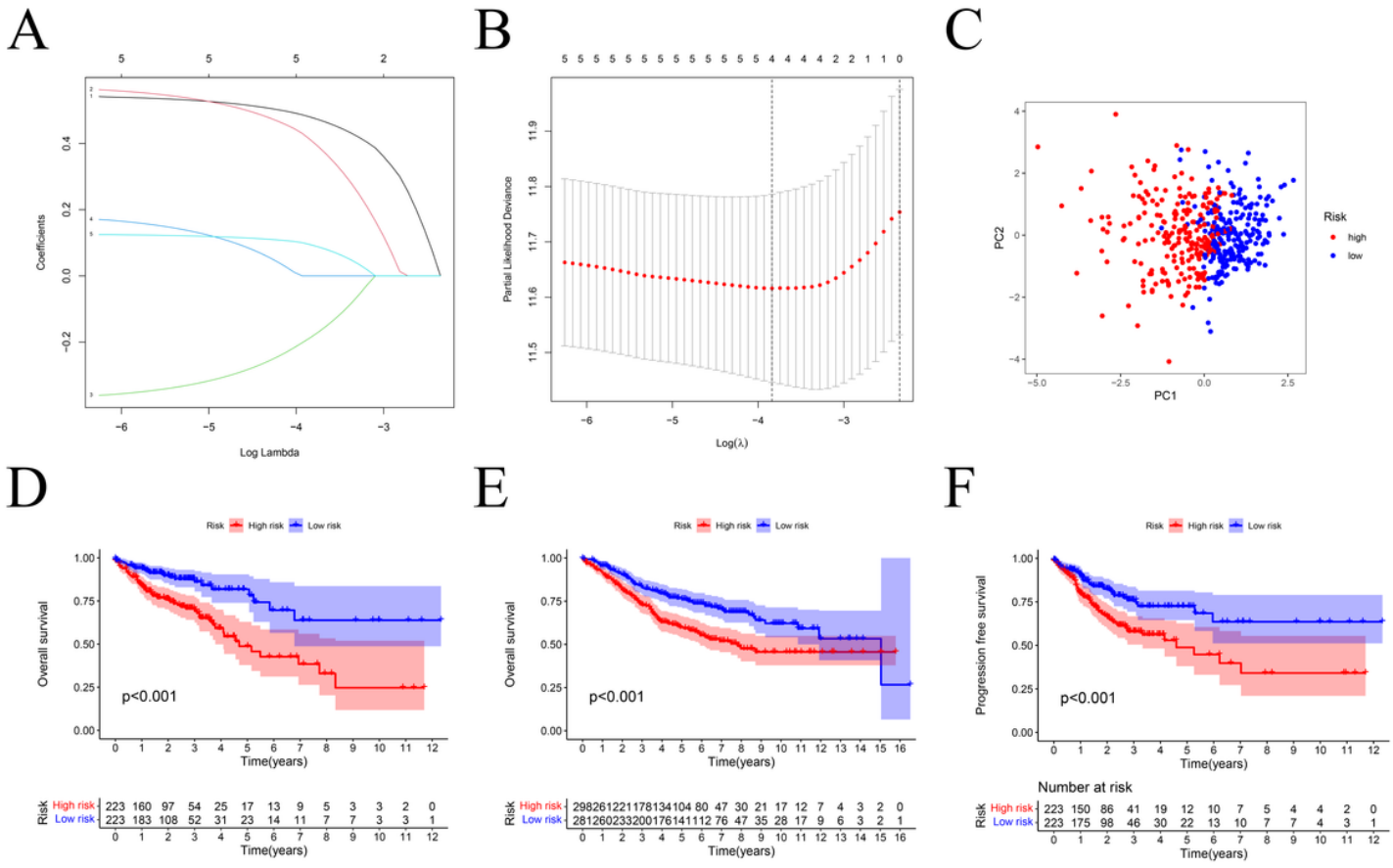

Figure 2

Construction and prognostic validation of the CSC-related lncRNA signature. (A) Lasso coefficient profiles of the CSC-related lncRNAs were determined for constructing a CSC-related lncRNA signature. (B) The selection of the tuning parameter of the Lasso model. (C) The principal component analysis (PCA) of the Train group (TCGA). (D) Kaplan-Meier analysis of the Train group based on overall survival. (E) Kaplan-Meier analysis of the Test group (GEO) based on overall survival. (F) Kaplan-Meier analysis of the Train group based on progression free survival.

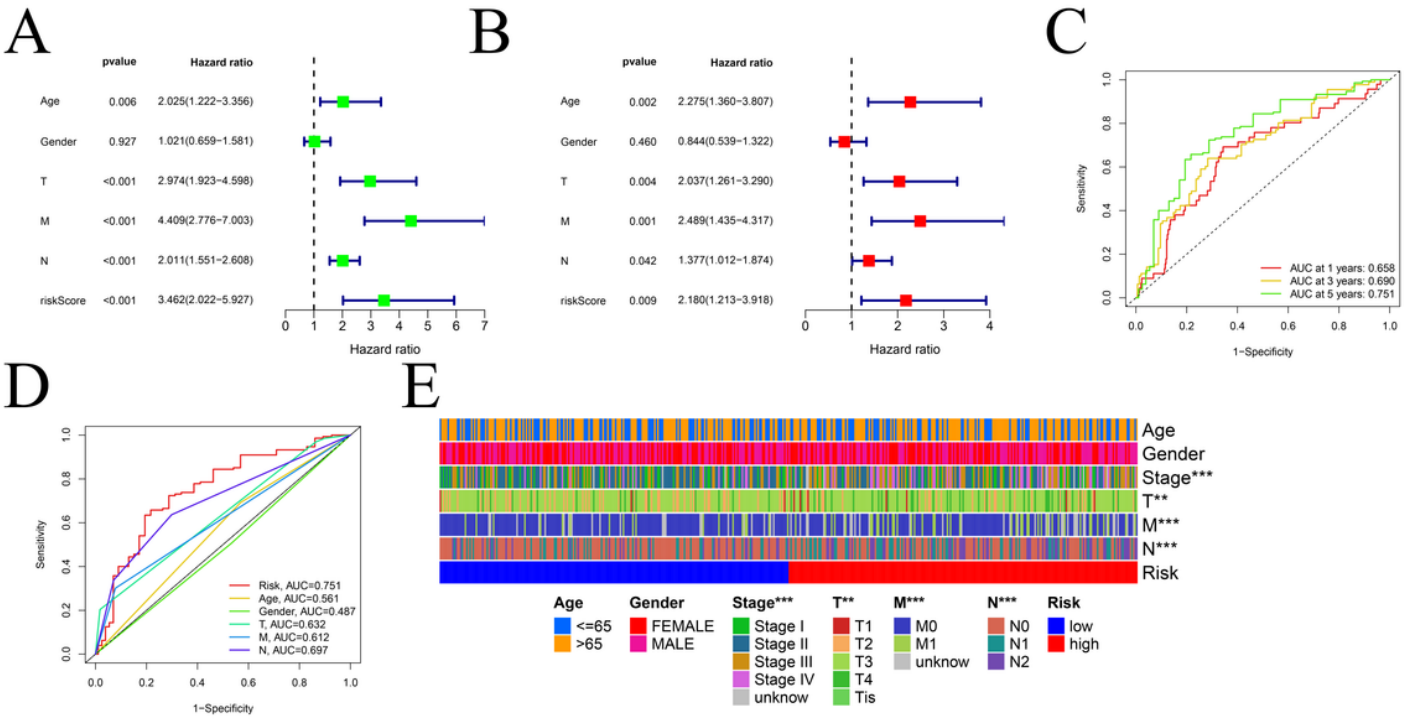

Figure 3

The clinical aspects validation of the CSC-related lncRNA signature. (A) Univariate cox survival analysis of the CSC-related lncRNA signature. (B) Multivariate cox survival analysis of the CSC-related lncRNA signature. (C) The time-dependent ROC curves in terms of 1-, 3- ,5- years. (D) The clinic-dependent ROC curves in terms of age, gender, T, N, and M. (E) Differences of clinical characteristics between high and low risk groups.

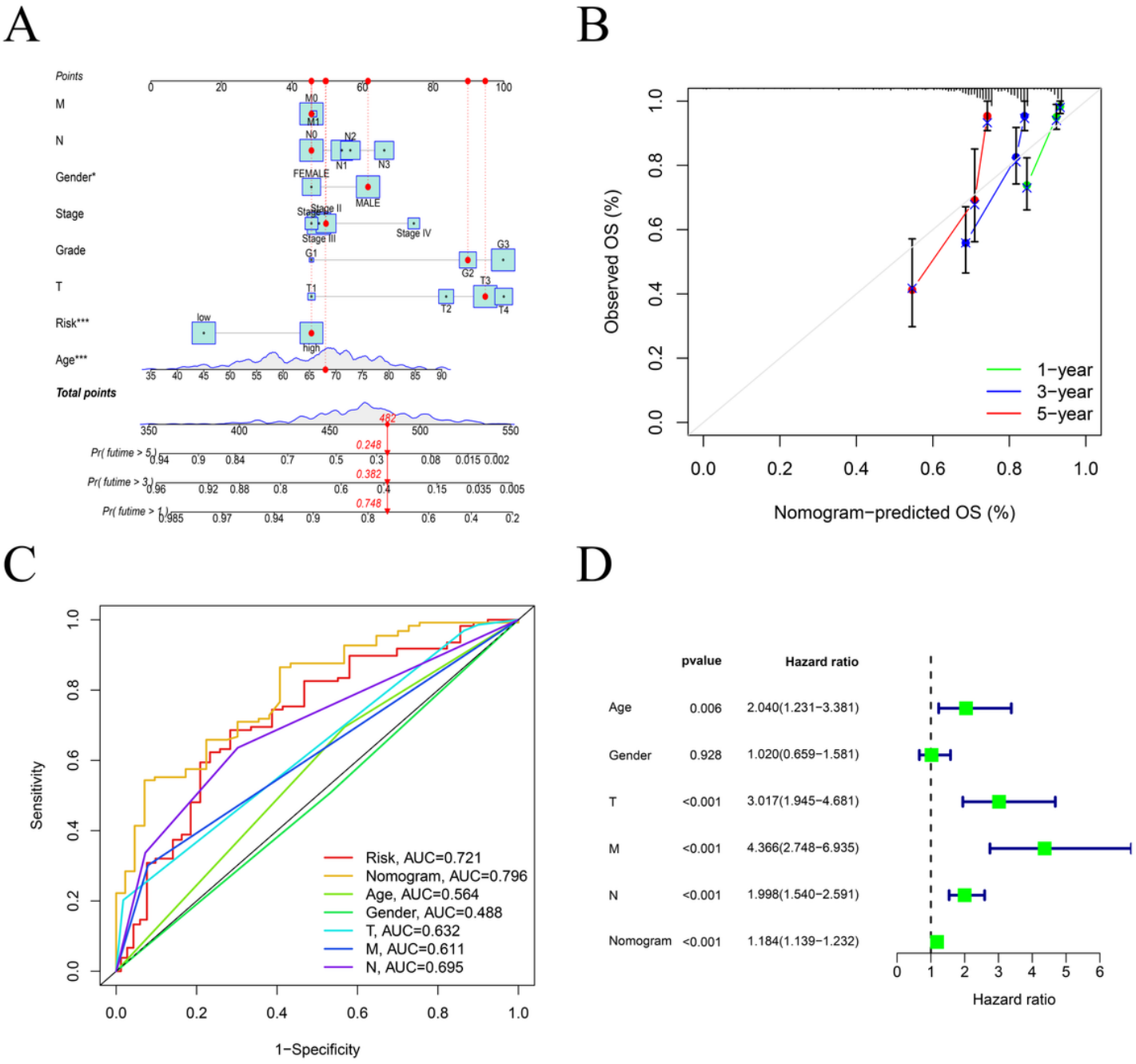

Figure 4

The nomogram analysis of the CSC-related lncRNA signature. (A) The nomogram of the CSC-related lncRNA signature. (B) The calibration of CSC-related lncRNA nomogram. (C) The ROC curve of the CSC-related lncRNA nomogram. (D) Univariate cox survival analysis of the CSC-related lncRNA nomogram.

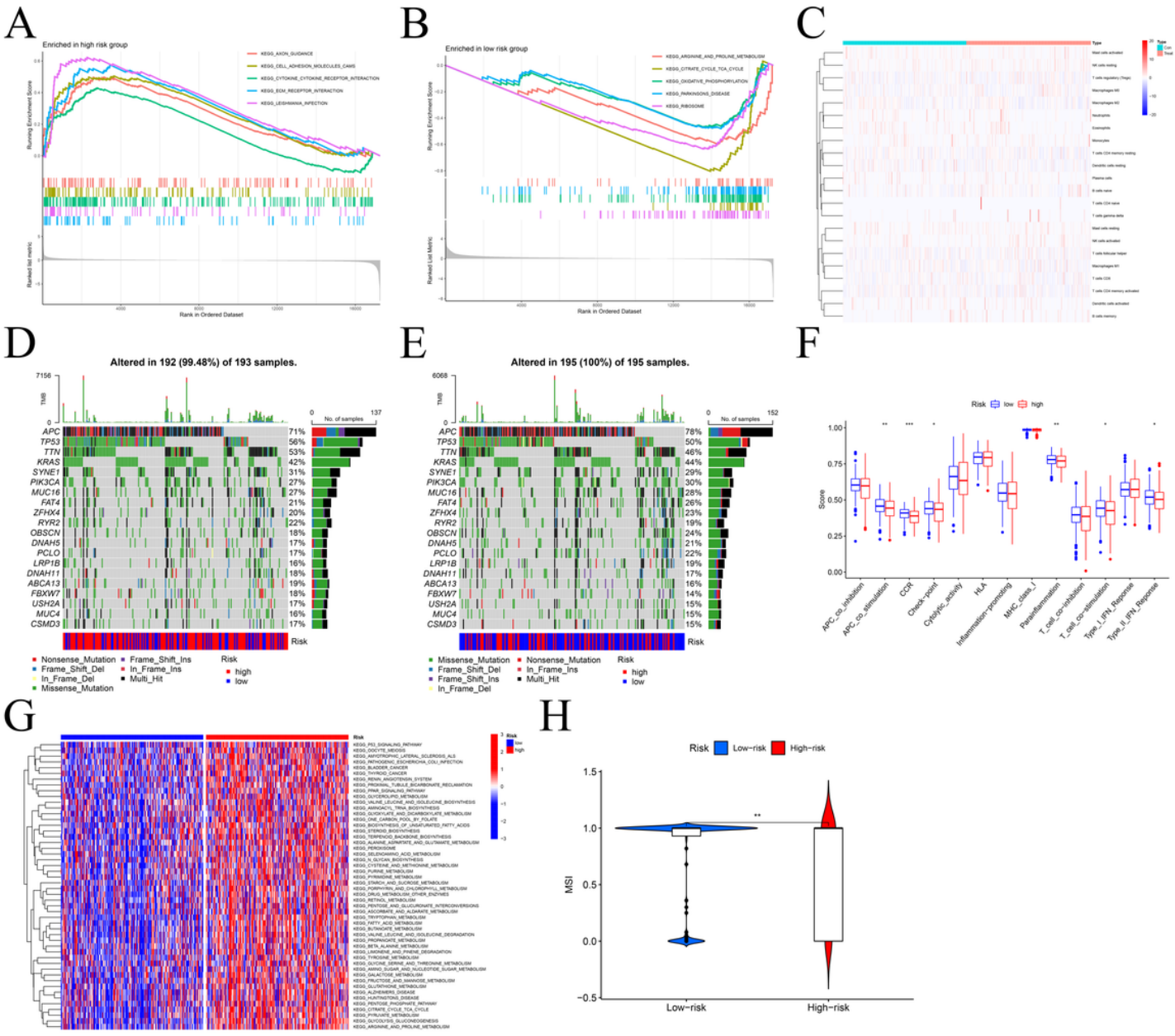

Figure 5

The comprehensive functional analysis of the CSC-related lncRNA signature. (A) The top 5 activated signalling pathways in GSEA. (B) The top 5 inhibited signalling pathways in GSEA. (C) Immunocyte differential analysis between the high and low risk groups. (D) Mutations in the top 20 mutated genes in the low risk group. (E) Mutations in the top 20 mutated genes in the high risk group. (F) The immune-related function analysis between the high and low risk groups. (G) GSVA analysis between the high and low risk groups. (H) The MSI situations between the high and low risk groups.

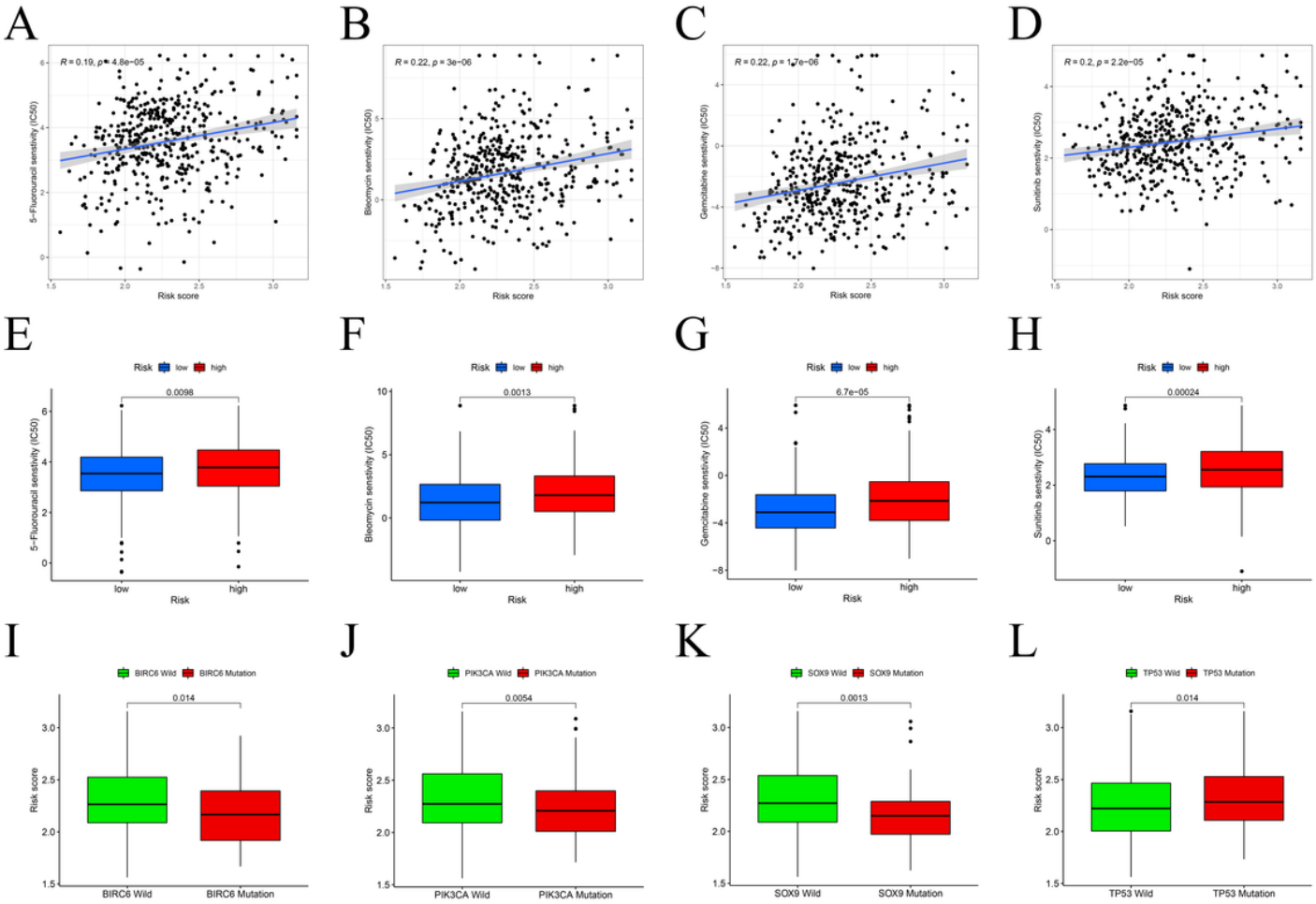

Figure 6

Differences in chemotherapy response and critical gene mutations between high and low colorectal cancer stem cell related lncRNA score (CCSC-related lncRNA score) groups. The correlation between IC50 and CCSC-related lncRNA score. (A) 5-fluorouracil, (B) Bleomycin, (C) Gemcitabine, (D) Sunitinib. Differences in IC50 between high and low CCSC-related lncRNA score groups. (E) 5-fluorouracil, (F) Bleomycin, (G) Gemcitabine, (H) Sunitinib. The differences in critical gene mutations between high and low CCSC-related lncRNA score groups. (I) BIRC6, (J) PIK3CA, (K) SOX9, (L) TP53.

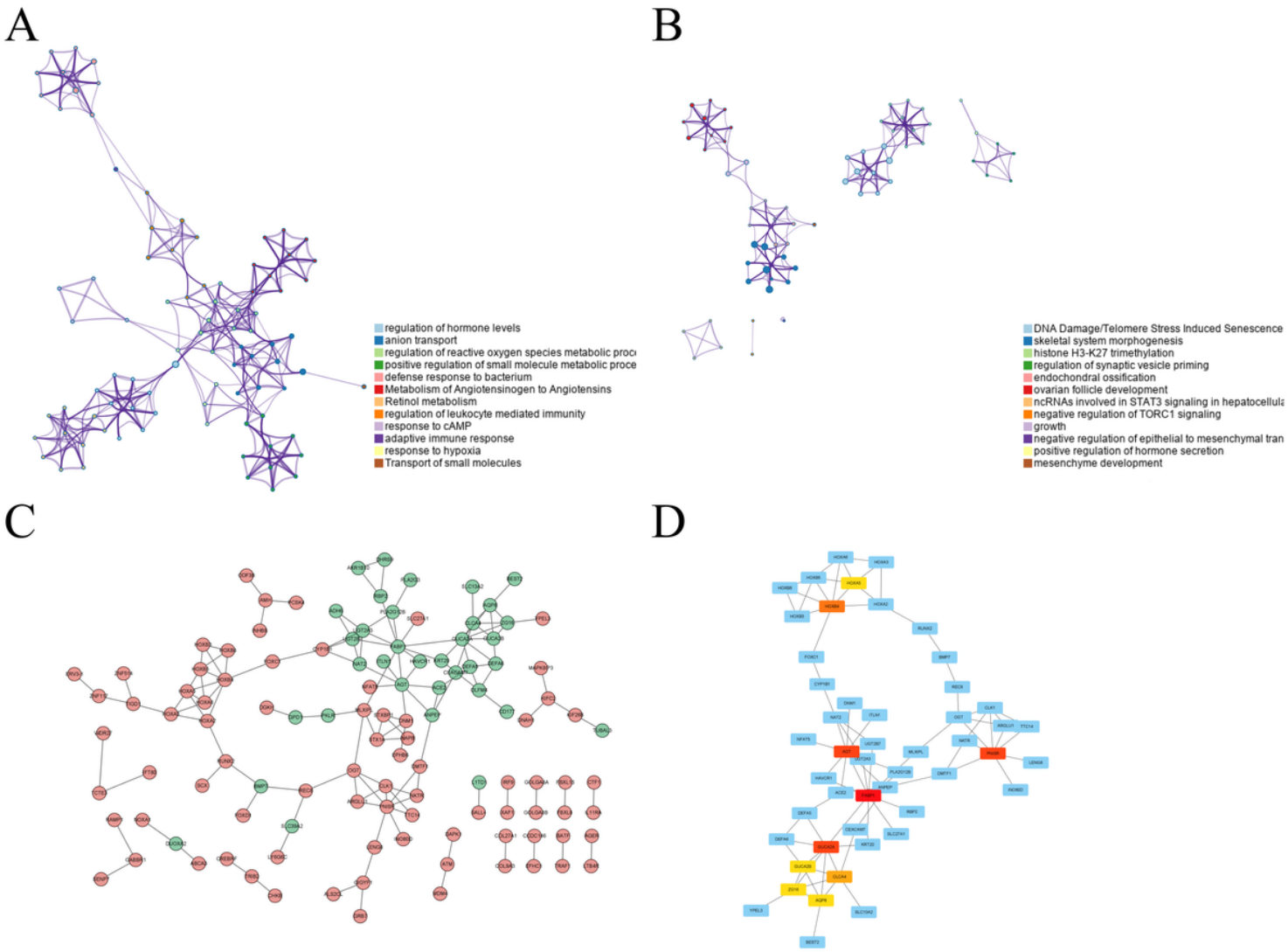

Figure 7

The metascape and protein-protein interaction (PPI) analysis of the differential expressed genes between the high and low CCSC-related lncRNA score groups. (A) The metascape analysis of the over-expressed genes in the high risk group. (B) The metascape analysis of the over-expressed genes in the low risk group. (C) The PPI analysis of the differential expressed genes. (D) The central genes in the PPI network sub-module.

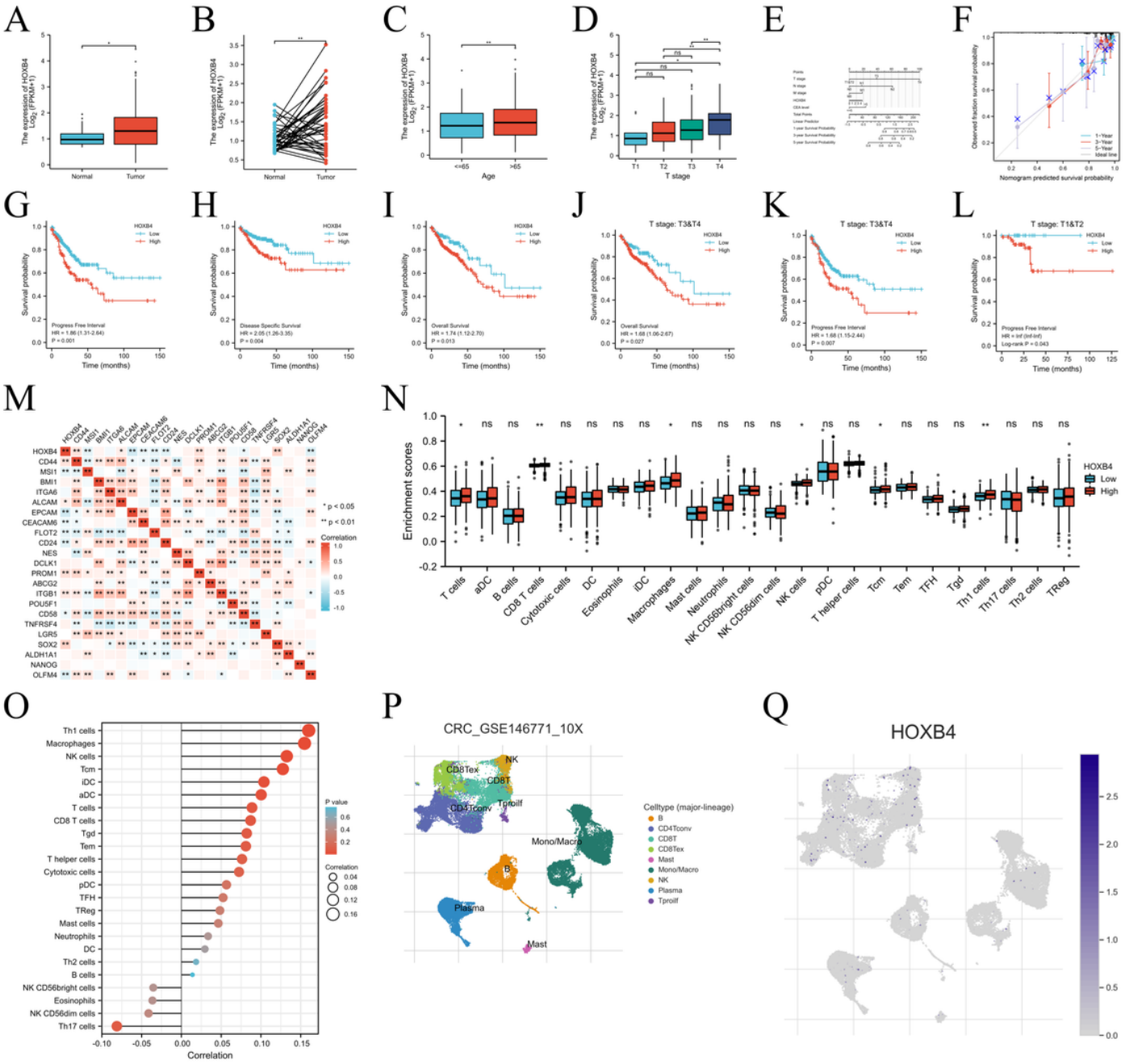

Figure 8

Comprehensive and integrative analysis of HOXB4 in CRC. (A) Expression of HOXB4 in unpaired cancer and paracancer samples. (B) Expression of HOXB4 in paired carcinoma and paracancer samples. (C) Expression of HOXB4 in patients of advanced and junior age. (D) Expression of HOXB4 at different T-stages. (E) The nomogram of HOXB4. (F) The calibration of HOXB4. (G) Kaplan-Meier analysis of HOXB4 in progression free interval (PFI). (H) Kaplan-Meier analysis of HOXB4 in disease specific survival (DSS). (I) Kaplan-Meier analysis of HOXB4 in overall survival (OS). Kaplan-Meier analysis of HOXB4 in advanced CRC patients. (J) OS, (K) PFI. (L) Kaplan-Meier analysis of HOXB4 in early stage CRC patients. (M) Correlation of HOXB4 with CSC markers in CRC. (N) Infiltration of immune cells between high- and low-

HOXB4 expression groups. (O) Correlation of HOXB4 with immune cell infiltration. (P) Single cell distribution map of GSE146771. (Q) Distribution of HOXB4 in single cell sequencing.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- corResult.txt