Integrative analysis from multi-center studies identities a consensus machine learning-derived lncRNA signature for stage II/III colorectal cancer

Zaoqu Liu,[a,b,c] ChunGuang Guo,[d] Qin Dang,[e] Libo Wang,[f] Long Liu,[f] Siyuan Weng,[a] Hui Xu,[a] Taoyuan Lu,[g] Zhenqiang Sun,[e,*] and Xinwei Han,[a,b,c,**]

[a]Department of Interventional Radiology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan 450052, China
[b]Interventional Institute of Zhengzhou University, Zhengzhou, Henan 450052, China
[c]Interventional Treatment and Clinical Research Center of Henan Province, Zhengzhou, Henan 450052, China
[d]Department of Endovascular Surgery, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan 450052, China
[e]Department of Colorectal Surgery, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan 450052, China
[f]Department of Hepatobiliary and Pancreatic Surgery, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan 450052, China
[g]Department of Cerebrovascular Disease, Zhengzhou University People's Hospital, Zhengzhou, Henan 450003, China

Summary

Background Long non-coding RNAs (lncRNAs) have recently emerged as essential biomarkers of cancer progression. However, studies are limited regarding lncRNAs correlated with recurrence and fluorouracil-based adjuvant chemotherapy (ACT) in stage II/III colorectal cancer (CRC).

Methods 1640 stage II/III CRC patients were enrolled from 15 independent datasets and a clinical in-house cohort. 10 prevalent machine learning algorithms were collected and then combined into 76 combinations. 109 published transcriptome signatures were also retrieved. qRT-PCR assay was performed to verify our model.

Findings We comprehensively identified 27 stably recurrence-related lncRNAs from multi-center cohorts. According to these lncRNAs, a consensus machine learning-derived lncRNA signature (CMDLncS) that exhibited best power for predicting recurrence risk was determined from 76 kinds of algorithm combinations. A high CMDLncS indicated unfavorable recurrence and mortality rates. CMDLncS not only could work independently of common clinical traits (e.g., AJCC stage) and molecular features (e.g., microsatellite state, *KRAS* mutation), but also presented dramatically better performance than these variables. qRT-PCR results from 173 patients further verified our in-silico findings and assessed its feasible in different centers. Comparisons of CMDLncS with 109 published transcriptome signatures further demonstrated its predictive superiority. Additionally, patients with high CMDLncS benefited more from fluorouracil-based ACT and were characterized by activation of stromal and epithelial-mesenchymal transition, while patients with low CMDLncS suggested the sensitivity to bevacizumab and displayed enhanced immune activation.

Interpretation CMDLncS provides an attractive platform for identifying patient at high risk of recurrence and could optimize precision treatment to improve the clinical outcomes in stage II/III CRC.

Funding This study was supported by the National Natural Science Foundation of China (81,972,663); Henan Province Young and Middle−Aged Health Science and Technology Innovation Talent Project (YXKC2020037); and Henan Provincial Health Commission Joint Youth Project (SB201902014).

Keywords: Stage II/III colorectal cancer; LncRNA; Machine learning; Recurrence; Chemotherapy

EBioMedicine 2022;75: 103750
Published online xxx
https://doi.org/10.1016/j.ebiom.2021.103750

*Corresponding author.
**Corresponding author at: Department of Interventional Radiology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan 450052, China.
E-mail addresses: liuzaoqu@163.com (Z. Liu), fccsunzq@zzu.edu.cn (Z. Sun), fcchanxw@zzu.edu.cn (X. Han).

Research in context

Evidence before this study

Long non-coding RNAs (lncRNAs) have recently emerged as essential biomarkers of cancer progression. However, studies are limited regarding lncRNAs correlated with recurrence and fluorouracil-based adjuvant chemotherapy (ACT) in stage II/III colorectal cancer (CRC).

Added value of this study

Our study enrolled 1640 stage II/III CRC patients from 15 datasets and an in-house cohort. In total, 27 stably recurrence-related lncRNAs were identified from multi-center cohorts. Subsequently, a consensus machine learning-derived lncRNA signature (CMDLncS) that exhibited best power for predicting recurrence risk was determined from 76 algorithm combinations. CMDLncS not only could work independently of common clinical and molecular factors, but also presented better performance. Comparisons of CMDLncS with 109 published transcriptome signatures further demonstrated its predictive superiority. A high CMDLncS indicated unfavorable recurrence and mortality rates, sensitivity to fluorouracil-based ACT, resistance to bevacizumab, stronger stromal activation and epithelial-mesenchymal transition, and inferior immune activation. Overall, CMDLncS provides an attractive platform for optimizing decision-making of stage II/III CRC.

Implications of all the available evidence

CMDLncS provides an attractive platform for identifying patient at high risk of recurrence and could optimize precision treatment to improve the clinical outcomes in stage II/III CRC.

Introduction

Colorectal cancer (CRC) represents one of the most prevalent tumors and the second leading cause of global cancer death according to the GLOBOCAN 2020 statistics.[1] Approximately 70% of CRC patients are diagnosed with stage II/III tumors. Currently, the therapeutic management of CRC mainly relies on the tumor-node-metastasis (TNM) staging system. After curative surgery, fluorouracil-based adjuvant chemotherapy (ACT) remains the standard of care treatment for stage III CRC and some stage II CRC with high-risk clinical features (e.g., T4, high grade), leading to prolonged overall survival (OS) and decreased recurrence risk for these patients.[2] Nevertheless, the selection of patients is currently suboptimal, which further gives rise to either over- or undertreatment.[3] A previous study has reported that only 20% of stage III CRC patients benefit from ACT, while 80% of patients are exposed to unnecessary toxicity.[4] For patients with stage II CRC, the application of ACT remains controversial because only a subset of patients will yield considerable benefit. Although the QUASAR clinical trial demonstrated that ACT could improve the OS of stage II CRC, the absolute improvement was quite limited (approximately 3.6%).[5] Additionally, up to 30% of stage II patients will develop recurrence after surgery and succumb to their disease.[6] Thus, the current staging system is insufficient for clinical management in stage II/III CRC, and it is imperative to identify reliable biomarkers for detecting high-risk patients who might benefit from ACT.

Recently, accumulated evidence has revealed that genetic and epigenetic alterations are closely implicated in the prognosis and treatment of CRC.[2] Mutational biomarkers such as *TP53*, *KRAS*, *BRAF*, microsatellite instability (MSI), and tumor mutational burden (TMB) are commonly applied in clinical settings.[7] Our team has reported a *TTN*/*OBSCN* "double-hit" tumor that is significantly correlated with better prognosis and superior immune infiltration.[8] However, the high cost, small proportion, and moderate performance hinder the clinical utilization of mutational biomarkers. The transcriptomic-based consensus molecular subtype (CMS) and CRC intrinsic subtype (CRIS) have been developed to reveal the heterogeneity of molecular features and clinical outcomes of CRC.[9,10] These classification systems are currently limited in clinical practice due to a lack of standardization and the requirement of bioinformatics resources.[11] An immunohistochemistry-based scoring approach used to assess the recurrence risk, termed Immunoscore®, has been established and validated, which measures the tumor core and invasive margin of CD3+ and CD8+ T cells.[12] Although Immunoscore® exhibits stable power in assessing the recurrence risk of early-stage CRC, its performance remains moderately accurate according to the C-index evaluation in an international trial.[12] Circulating tumor DNA (ctDNA) released by tumors into the bloodstream also has potential for liquid biopsy in assessing prognosis and guiding treatment but still needs further exploration and validation.[2]

As is well-known, CRC is a complex disease with both inter- and intra-tumor heterogeneity. An ideal biomarker should maintain homogenous expression within and between tumor tissues. Hence, a multigene panel might be a promising approach to address this issue.[2] Long noncoding RNAs (lncRNAs) represent a newly discovered class of noncoding RNAs with > 200 nucleotides that include the majority of human RNAs (approximately 76%).[13,14] Previous studies have incorporated lncRNAs into preclinical signatures to identify multigene panels associated with tumor recurrence, prognosis, and chemoresistance.[15,16] However, due to underutilized data, inapposite machine learning algorithms, a lack of rigorous validation, and no clinical testing, these multigene expression signatures are usually hard for clinical interpretation.[17–19]

To tackle the abovementioned considerations, our study tried to comprehensively explore the clinical significance of lncRNAs in stage II/III CRC and systematically identify a consensus machine learning-derived lncRNA signature (CMDLncS) from 76 kinds of algorithm combinations. CMDLncS was tested in a total of 1640 stage II/III CRC patients from 16 independent datasets to evaluate the recurrence risk, OS, and benefits of fluorouracil-based ACT and bevacizumab. We compared CMDLncS with common clinical traits, molecular features, and 109 published signatures to further verify its robustness and translation. We also revealed the latent biological mechanisms underlying CMDLncS. Overall, our study offers an attractive platform for detecting patients at a high risk of recurrence and could optimize precision treatment to improve the clinical outcomes of stage II/III CRC.

Methods

Data collection

The overall design of this study is displayed in Figure 1. Our study retrospectively enrolled 15 independent CRC cohorts from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO). A total of 1467 patients were retained according to the following criteria: (1) Primary CRC; (2) Possessed gene expression profiles and clinical annotations; (3) AJCC stage II/III; (4) No preoperative chemotherapy or radiotherapy received. The detailed baseline was summarized in Table S1. Among these cohorts, eight cohorts (TCGA-CRC, GSE143985, GSE161158, GSE17536, GSE29621, GSE31595, GSE39582, and GSE92921) with complete recurrence-free survival (RFS) information were utilized to develop and validate the CMDLncS model. Four cohorts, TCGA-CRC, GSE17536, GSE29621, and GSE39582, also containing overall survival (OS), were further applied to explore the predictive value of CMDLncS for OS. Seven cohorts were treated with fluorouracil-based ACT (FOLFOX or FOLFIRI) alone, including TCGA-CRC, GSE19860, GSE28702, GSE45404, GSE62080, GSE69657, and GSE72970, while three cohorts were treated with fluorouracil-based ACT in combination with bevacizumab, including GSE19860, GSE19862, and GSE72970. These drug-related cohorts were employed to evaluate the performance of CMDLncS in predicting ACT and bevacizumab benefits in stage II/III CRC.

Genome-wide lncRNA and mRNA expression

For TCGA-CRC, RNA-seq raw read count from the TCGA portal was converted to transcripts per kilobase million (TPM) and further log-2 transformed. Data from the GEO database were all retrieved from the Affymetrix® GPL570 platform (Human Genome U133 Plus 2.0 Array). The raw data from Affymetrix® were processed via the robust multiarray averaging (RMA) algorithm implemented in the *Affy* package. The *ComBat* algorithm implemented in the *sva* package was utilized to remove batch effects from nonbiological technical biases. The principal component analysis (PCA) before and after batch correction is shown in Fig. S1a, b. The GENCODE database (https://www.gencodegenes.org/) was applied to lncRNA and mRNA annotations. Furthermore, the intersection of two platforms (Illumina and GPL570) was taken, and ultimately, 3390 lncRNAs and 17,046 mRNAs were retained for the subsequent analysis.

Signature generated from machine learning integrative approaches

Prior to constructing a consensus machine learning-derived signature (CMDLncS), we transformed lncRNA expression into z-score in all cohorts, which enhanced the comparability between different datasets. Eight cohorts (TCGA-CRC, GSE143985, GSE161158, GSE17536, GSE29621, GSE31595, GSE39582, and GSE92921) with complete recurrence information were utilized to develop the CMDLncS model according to the following pipeline:

(1) Univariate Cox analysis was performed on all lncRNAs in these eight cohorts. Given the strictness of multiple testing correction and the small sample problem of some cohorts that might filter out latent lncRNAs associated with recurrence, lncRNAs possessing both an unadjusted $P < 0.1$ for more than six cohorts and the same hazard ratio (HR) direction for more than five cohorts were considered stable recurrence-related lncRNAs (SRRLs).

(2) The initial signature discovery was performed in GSE39582. Ten single machine learning algorithms, including random survival forest (RSF), elastic network (Enet), Lasso, Ridge, stepwise Cox, CoxBoost, partial least squares regression for Cox (plsRcox), supervised principal components (SuperPC), generalized boosted regression modeling (GBM), and survival support vector machine (survival-SVM), were applied. A few algorithms possessed the ability of feature selection, such as Lasso, stepwise Cox, CoxBoost, and RSF. Thus, we combined these algorithms to generate a consensus model. In total, 76 algorithm combinations were conducted on SRRLs to fit prediction models based on 10-fold cross-validation. The parameter tuning details are described in the Supplementary Material.

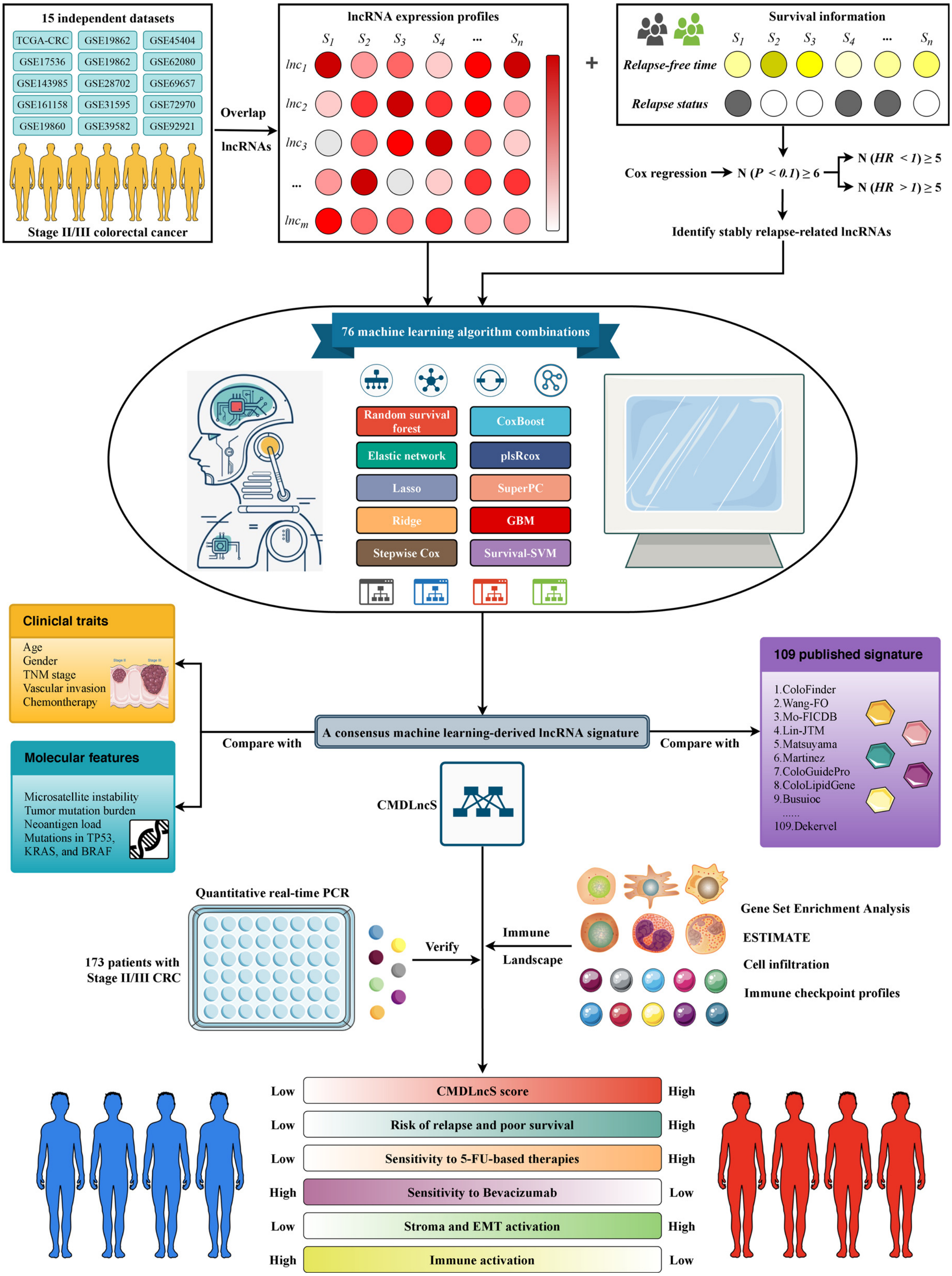


Figure 1. The overall flow of this study.

(3) These 76 models were further tested in the other seven cohorts. For each model, its C-indices across all validation datasets were calculated, and the model with the highest average C-index was considered as the optimal one.

Collection of published signatures

To compare the performance of CMDLncS with other signatures, we comprehensively retrieved published signatures. The miRNA signatures were excluded due to the severe lack of miRNA information in the validation datasets. Eventually, a total of 109 signatures (including mRNA and lncRNA signatures) were collected (Table S2). These signatures were fitted by diverse algorithms, such as stepwise Cox, Lasso, RSF, and single sample gene set enrichment analysis (ssGSEA). In addition, these signatures were derived from various biological processes, such as the tumor microenvironment, autophagy, ferroptosis, stemness, epithelial-mesenchymal transition (EMT), Toll-like receptor signaling, hypoxia, metastasis, glycolysis, lipogenesis, vitamin D, epigenetics, N6-methyladenosine, aging, and drug sensitivity. For each signature, we performed univariate Cox regression and calculated the C-index in all cohorts.

Ethics

The human cancer tissues used in this study were approved by Ethnics Committee of The First Affiliated Hospital of Zhengzhou University on December 19, 2019, and the TRN is 2019-KW-423.

Human tissue specimens

A total of 173 frozen surgically resected CRC tissues were enrolled from The First Affiliated Hospital of Zhengzhou University. All patients gave written informed consent, and none of the patients received any preoperative chemotherapy or radiotherapy. After radical surgery, patients received available standard systemic therapies, such as FOLFOX (n = 35), FOLFIRI (n = 37), and bevacizumab (n = 48). Drug responses were evaluated based on the Response Evaluation Criteria in Solid Tumors (RECIST, version 1.1). Responders and nonresponders were defined as having a complete response (CR)/partial response (PR) and stable disease (SD)/progressive disease (PD), respectively. The detailed baseline characteristics of the patients are displayed in Table S1.

Quantitative real-time PCR (qRT-PCR)

Total RNA was isolated with the RNAisoPlus reagent (Takara, Dalian, China) as described previously.[20] RNA quality was evaluated using a NanoDrop One C (Waltham, MA, USA), and RNA integrity was assessed by agarose gel electrophoresis. An aliquot of 1 μg of total RNA was reverse-transcribed into complementary DNA (cDNA) using a High-capacity cDNA Reverse Transcription kit (TaKaRa Bio, Japan), according to the manufacturer's protocol. The primer sequences are shown in Table S3. See the Supplementary Material for more details.

Cell line and cell transfection

Human CRC cell lines HCT-116 (RRID: CVCL_0291) and SW480 (RRID: CVCL_0546) were obtained from the Chinese Academy of Science (Shanghai, China). Cell lines were authenticated by short tandem repeat polymerase chain reaction (STR-PCR). Mycoplasma infection status was tested by 4′, 6-diamidino-2-phenylindole (DAPI) staining in the laboratory. All cells were cultured in DMEM (high glucose) (HyClone, Logan, Australia) supplemented with 10% fetal bovine serum (Gibco; Thermo Fisher Scientific, Inc.), 100 U/ml penicillin, and 100 mg/ml streptomycin in a humidified incubator at 37ˉC with 5% CO2. Silencer select small interfering RNAs (siRNAs) specific for *ENSG00000232995* (*lnc-RGS5−1*), *ENSG00000249035* (*CLMAT3*), and inhibitor control were generated from RiboBio (Guangzhou, China). To silence lncRNAs in cancer cells, *lnc-RGS5−1* specific siRNA (si-lnc-RGS5−1, target sequence: GACATGGCCCAGAAAAGAA), *CLMAT3* specific siRNA (si-CLMAT3, target sequence: GGATGTTAGTGAGATCTA) and control siRNA were transfected into HCT-116 and SW480 cells. Lipofectamine 3000 (Invitrogen; Thermo Fisher Scientific, L3000−015) was utilized as a transfection carrier. The transfected cells were harvested after 60 h. The transfection efficiency was confirmed via qRT-PCR analysis.

Cell proliferation assay

In Cell Counting Kit-8 (CCK-8) assay, all cells were plated at 1.5 × 103 cells per well in 96-well plates and incubated overnight in DMEM (high glucose) (HyClone, Logan, Australia) supplemented with 10% FBS. The cell proliferation index was measured using a CCK-8 (Dojin Laboratories, Tokyo, Japan) at 0, 24, 48, 72, and 96 h post-transfection according to the manufacturer’s instruction. Then, 10 μl of CCK-8 solution was added to the culture medium and incubated for 2 h at 37ˉC. Absorbance was measured at a wavelength of 450 nm with a reference wavelength of 570 nm. Each experiment was repeated ≥ 3 times.

Transwell assay

The migration and invasive abilities of CRC cells were determined via transwell assays after transfection with siRNAs. Transwell chambers (Corning, NY, USA) were prepared with or without Matrigel. Then, blood serum medium (10% FBS) was added to the lower chamber.

After transfection with si-lnc-RGS5−1, si-CLMAT3 and inhibitor control, HCT-116 and SW480 cells were digested to prepare a cell suspension, and this suspension was added to the upper chamber and incubated for 24 h. At the end of this incubation, the residual cells in the upper chamber were gently wiped with cotton swabs. The cells were fixed with 4% paraformaldehyde and stained with 1% crystal violet for 30 min. After washing three times with PBS, the cells were imaged and counted with an OLYMPUS FV1000 confocal microscope. Every experiment was performed three times for statistical analysis.

Gene set enrichment analysis (GSEA)

A total of 21,338 gene sets were generated from the MSigDB resource (version 7.4, h.all.v7.4.symbols.gmt, c2.all.v7.4.entrez.gmt, and c5.all.v7.4.entrez.gmt). The correlation coefficients between the CMDLncS score and all mRNAs were calculated. Subsequently, all mRNAs were sorted in descending order via their correlations with the CMDLncS score. The ranked gene list was further subjected to the *clusterProfiler* package to perform GSEA. Gene sets with a false discovery rate (FDR) < 0.001 were considered to be significantly enriched.

EMT pathway activity

Gene set variation analysis (GSVA) is an unsupervised and nonparametric algorithm that quantifies activity variation, which is broadly utilized in bioinformatics analysis.[21] Based on the GSVA approach, the epithelial mesenchymal transition (version 7.4, Hallmark) gene set retrieved from the MSigDB resource was utilized to assess the EMT activity of each sample.

Tumor microenvironment

According to the gene expression profiles, the *ESTIMATE* package was utilized to infer the fraction of stromal and immune fractions in the tumor microenvironment.[22]

T-cell inflammatory signature (TIS)

TIS is composed of 18 inflammatory genes associated with antigen presentation, chemokine expression, cytotoxic activity, and adaptive immune resistance, which was used to quantify the T cell-inflamed microenvironment via ssGSEA and predict the putative response to pembrolizumab.[23]

Cell infiltration

The ssGSEA algorithm implemented in the *GSVA* package was employed to measure the relative infiltration of 28 immune cells in CRC.[24]

Statistical analysis

All data processing, statistical analysis, and plotting were conducted in the R 4.0.5 software. Correlations between two continuous variables were evaluated via Pearson's correlation coefficients. The Wilcoxon rank-sum test or T test was applied to compare the difference between two groups. The *survminer* package was used to determine the optimal cutoff value. Cox regression and Kaplan-Meier analysis were performed via the *survival* package. The time-dependent area under the ROC curve (AUC) for survival variables was determined by the *timeROC* package. The ROC curve used to predict binary categorical variables was implemented via the *pROC* package. The consensus molecular subtypes were inferred via the *CMSclassifier* package.[9] All statistical tests were two-sided. $P < 0.05$ was regarded as statistically significant. Error bars span 95% confidence intervals.

Role of funders

The funding agencies had no further role in study design, in the collection, analysis and interpretation of data, in the writing of the report and in the decision to submit the paper for publication.

Results

Integrative construction of a consensus signature

Univariate Cox regression analysis was performed in eight cohorts with complete recurrence information, revealing a total of 27 SRRLs that were stably associated with recurrence in stage II/III CRC (Fig. S2). Subsequently, the expression profiles of these 27 SRRLs were subjected to the machine learning integrative procedure to develop a consensus signature. In the GSE39582 cohort, our study fitted 76 kinds of prediction models via the 10-fold cross-validation framework. The C-index was utilized to evaluate the predictive performance of the 76 models. A few algorithms displayed extreme accuracy in the training cohort, such as RSF, with its C-index even reaching 0.95−0.99 in the training cohort, which was likely due to overfitting. In addition, evaluating a model mainly depends on whether it still maintains robust performance in different validation cohorts, which is known as the generalization ability of the model. Therefore, we only calculated the C-indices of each model in the other seven validation cohorts (Figure 2a), and the model with the highest average C-index was considered the optimal one. Notably, the optimal model was a combination of Lasso and stepwise Cox with the highest average C-index (0.777), which was also the only model with a C-index above 0.7 in all validation cohorts (Figure 2a).

In the Lasso regression, based on the 10-fold cross-validation, the regression partial likelihood deviance reached the minimum value when lambda = 0.012

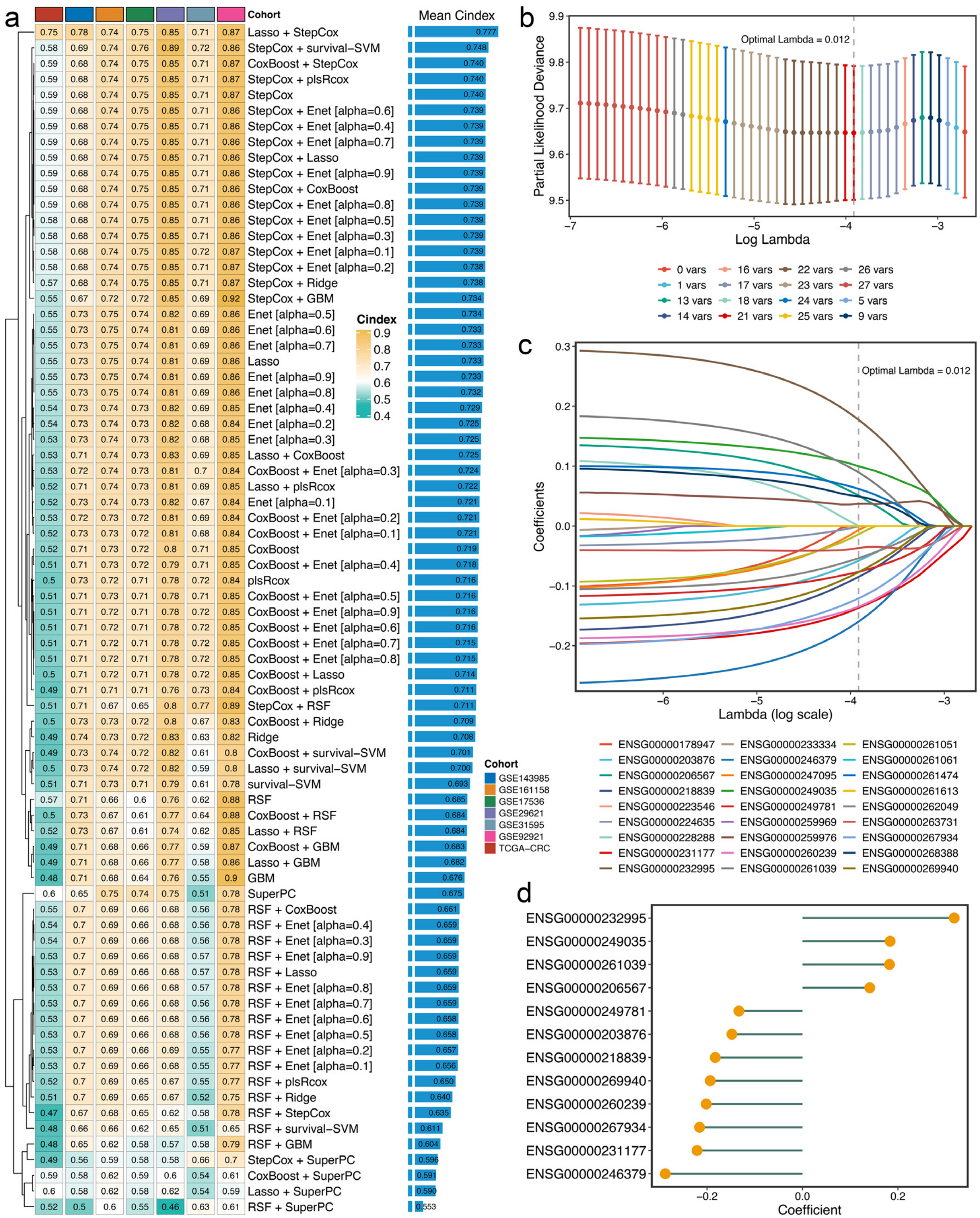


Figure 2. Integrative construction of a consensus signature. a. C-indices of 76 kinds of prediction models in seven validation cohorts. b. Determination of the optimal lambda was obtained when the partial likelihood deviance reached the minimum value, and further generated the key lncRNAs with nonzero coefficients. c. LASSO coefficient profiles of the candidate lncRNAs for CMDLncS construction. d. Coefficients of 12 lncRNAs finally obtained in stepwise Cox regression.

(Figure 2b). Twenty-one SRRLs with nonzero Lasso coefficients were defined as the key factors of recurrence in stage II/III CRC (Figure 2c). Subsequently, these SRRLs were further subjected to stepwise Cox proportional hazards regression based on the Akaike information criterion (AIC), which identified a final set of 12 SRRLs, including *ENSG00000232995*, *ENSG00000249035*, *ENSG00000261039*, *ENSG00000206567*, *ENSG00000249781*, *ENSG00000203876*, *ENSG00000218839*, *ENSG00000269940*, *ENSG00000260239*, *ENSG00000267934*, *ENSG00000231177*, and *ENSG00000246379* (Figure 2d). Therefore, a risk score for each patient was calculated using the expression of 12 SRRLs weighted by their regression coefficients (Figure 2d). Furthermore, we selected two lncRNAs, *ENSG00000232995* (*lnc-RGS5–1*) and *ENSG00000249035* (*CLMAT3*), to further explore their roles in CRC cells. HCT-116 and SW480 cells were transfected with si-lnc-RGS5–1, si-CLMAT3 and inhibitor control. The knockdown efficiency was confirmed via qRT-PCT analysis (Fig. S3). CCK-8 and transwell assays suggested that lnc-RGS5–1 and CLMAT3 could promote the proliferation, migration, and invasion of CRC cells (Fig. S4).

Independent prognostic value of CMDLncS

Our study dichotomized all patients into high- and low-risk groups based on the optimal cutoff value determined by the *survminer* package. Kaplan-Meier survival analysis showed that the rate of recurrence in the high-risk group was dramatically higher than the low-risk group in the training cohort GSE39582 (n = 465, log-rank test: P < 0.0001), and similar results were also observed in the validation cohorts TCGA-CRC (n = 169, log-rank test: P < 0.0001), GSE143985 (n = 91, log-rank test: P < 0.0001), GSE161158 (n = 141, log-rank test: P < 0.0001), GSE17536 (n = 111, log-rank test: P < 0.0001), GSE29621 (n = 40, log-rank test: P < 0.0001), GSE31595 (n = 37, log-rank test: P < 0.0001), GSE92921 (n = 59, log-rank test: P < 0.0001), and meta-cohort (n = 1113, log-rank test: P < 0.0001) (Figure 3a–i). Likewise, comparisons of OS demonstrated that the mortality rate in the high-risk group was significantly higher than that in the low-risk group in GSE39582 (n = 465, log-rank test: P < 0.0001), TCGA-CRC (n = 169, log-rank test: P = 0.0029), GSE17536 (n = 111, log-rank test: P = 0.0039), GSE29621 (n = 40, log-rank test: P = 0.033), and meta-cohort (n = 785, log-rank test: P < 0.0001) (Figure 3j–n).

To assess whether the prognostic significance of the CMDLncS model was independent of common clinical traits and molecular features, multivariate Cox regression analysis was conducted on age, gender, T, N, AJCC stage, vascular invasion (VI), ACT, TMB, neoantigen load (NAL), microsatellite state, *TP53*, *KRAS*, or *BRAF* mutations, and our CMDLncS model. The results showed that CMDLncS remained statistically significant for RFS and OS in all cohorts when adjusted for these clinical and molecular variables, indicating that it could serve as an independent risk factor in stage II/III CRC (Tables S4, S5).

Robust performance of CMDLncS

As illustrated in Figure 4a, our CMDLncS model demonstrated powerful performance in the training cohort GSE39582, and the time dependent AUCs were 0.755/0.770/0.739 at 1/3/5 years. Similar results were also observed in the validation cohorts TCGA-CRC (0.749/0.753/0.867), GSE143985 (0.767/0.788/0.784), GSE161158 (0.749/0.798/0.782), GSE17536 (0.811/0.799/0.806), GSE29621 (0.802/0.885/0.848), GSE31595 (0.736/0.763/0.715), GSE92921 (0.805/0.875/0.871), and meta-cohort (0.756/0.778/0.760) (Figure 4a). The C-indices [95% confidence interval] were 0.729 [0.698–0.760], 0.752 [0.656–0.849], 0.776 [0.679–0.873], 0.739 [0.689–0.790], 0.746 [0.692–0.800], 0.846 [0.728–0.965], 0.712 [0.533–0.892], 0.866 [0.754–0.979], and 0.740 [0.718–0.762] in nine cohorts, respectively (Figure 4b). Hence, the above results suggested that our CMDLncS model possessed the stable and robust performance in multiple independent cohorts.

In clinical settings, clinicians usually apply the clinical traits (e.g., AJCC stage, VI) and molecular features (e.g., microsatellite state, *KRAS* mutation) for prognostic evaluation and management.[7] Thus, we compared the predictive superiority of CMDLncS with common clinical traits and molecular features for predicting the recurrence risk of stage II/III CRC after radical surgery. In eight cohorts, CMDLncS presented significantly superior accuracy than these variables, such as age, gender, T, N, AJCC stage, VI, TMB, NAL, microsatellite state, ACT, and *TP53*, *KRAS*, or *BRAF* mutations (Figure 4c–j). This indicated that our CMDLncS model could be a promising surrogate for predicting the recurrence risk of stage II/III CRC in clinical practice.

Comparisons of gene expression signatures

Recently, with advancements in high-throughput sequencing techniques and computational biology, numerous predictive gene expression signatures have been proposed according to various machine learning approaches.[25] To compare the performance of CMDLncS with other signatures, we systematically enrolled a total of 109 signatures, mainly encompassing lncRNA and mRNA signatures (Table S2). These signatures were developed via multiple algorithms, such as stepwise Cox, Lasso, RSF, and ssGSEA. Univariate Cox regression analysis was performed on each signature, and notably, only our CMDLncS model maintained complete significance across all datasets (Figure 5a), which suggested the stability of CMDLncS for assessing

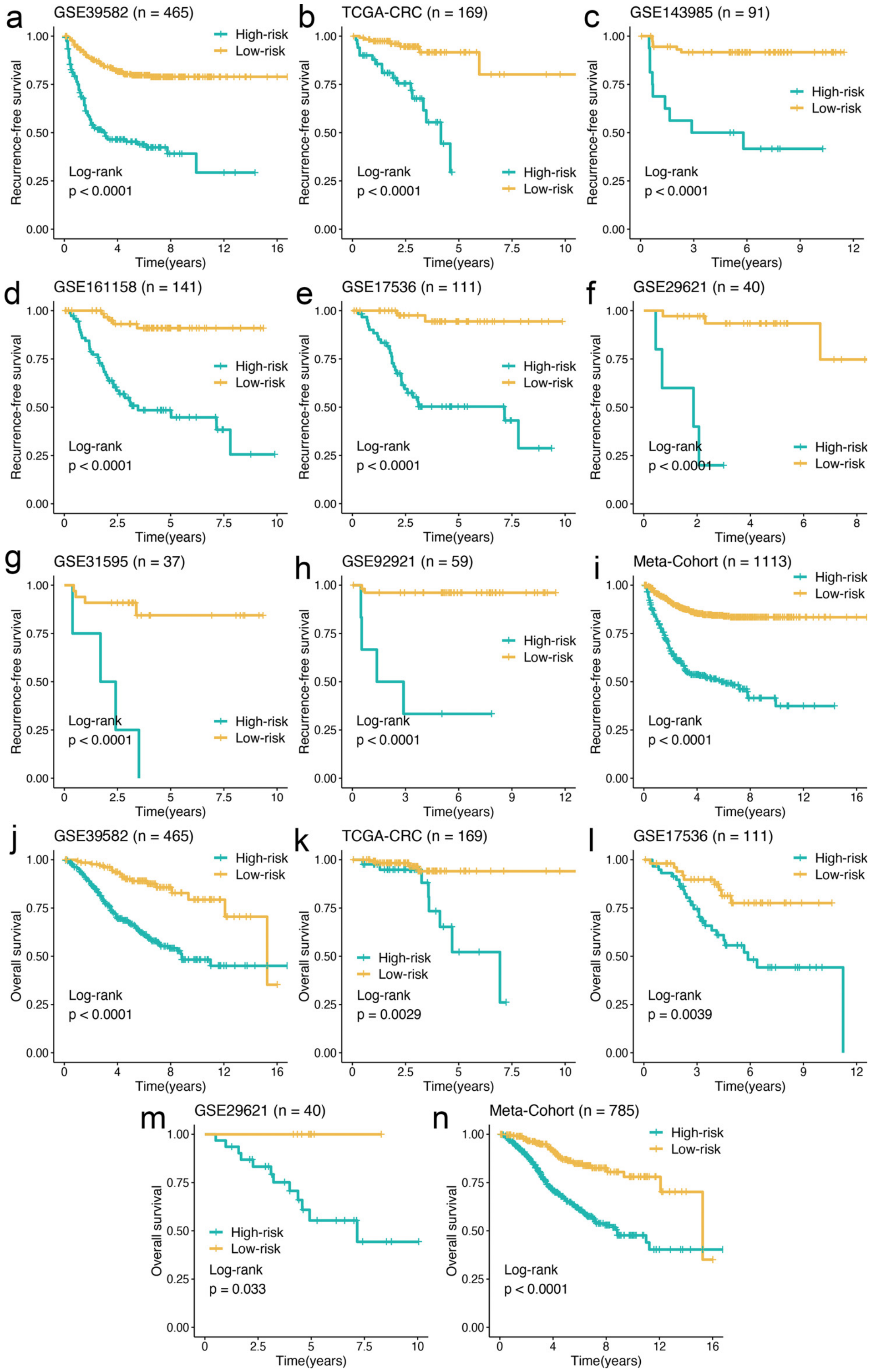


Figure 3. Kaplan-Meier survival analysis of CMDLncS. a–i. Kaplan-Meier curves of RFS according to the CMDLncS in GSE39582 (a), TCGA-CRC (b), GSE143985 (c), GSE161158 (d), GSE17536 (e), GSE29621 (f), GSE31595 (g), GSE92921(h), and meta-cohort (i). j–n. Kaplan-Meier curves of OS according to the CMDLncS in GSE39582 (j), TCGA-CRC (k), GSE17536 (l), GSE29621 (m), and meta-cohort (n).

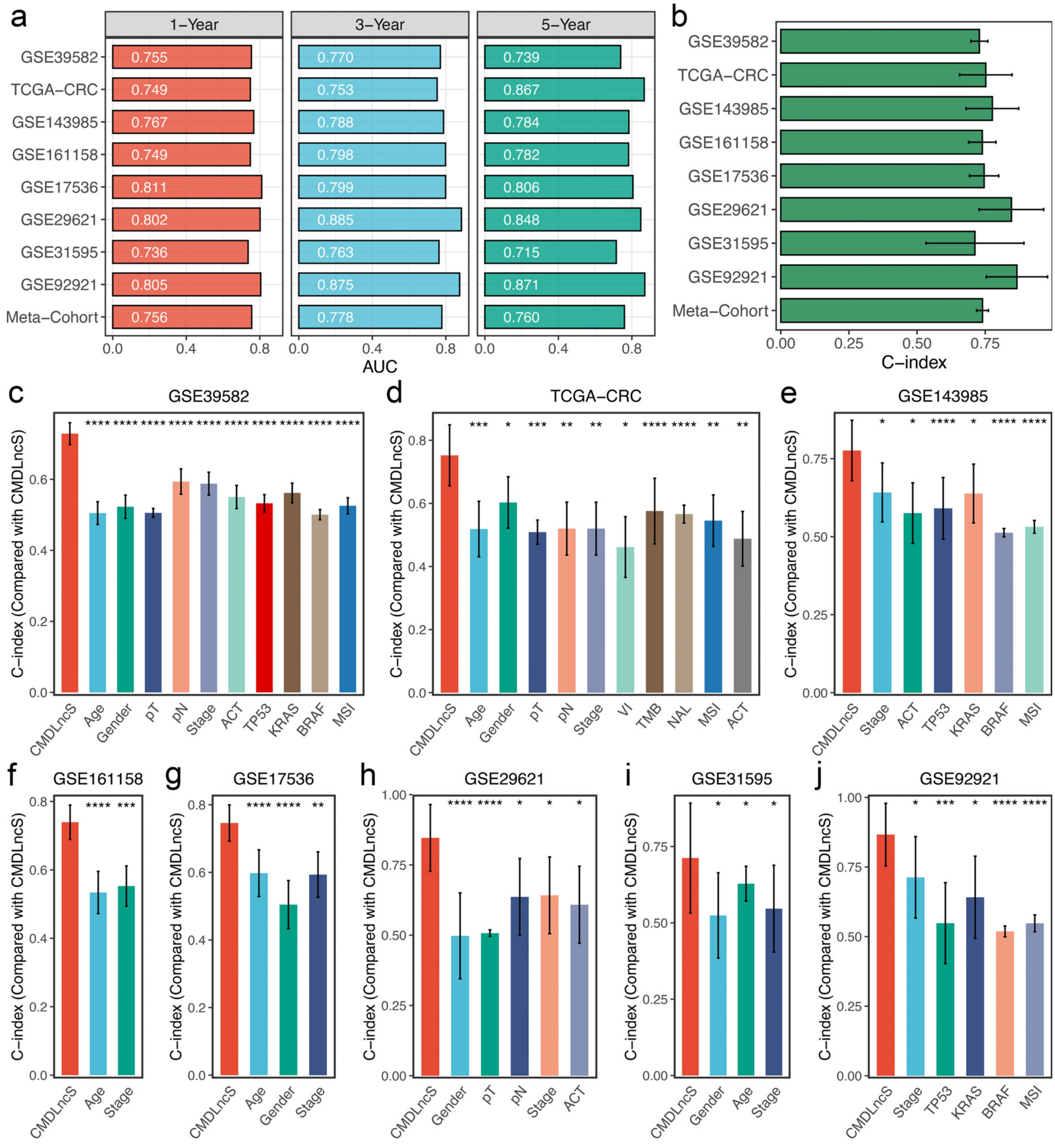


Figure 4. Robust performance of CMDLncS. a. Time-dependent ROC analysis for predicting RFS at 1, 3, and 5 years. b. C-indices of CMDLncS across all datasets. c−j. The performance of CMDLncS was compared with common clinical and molecular variables in predicting prognosis across all training and validation cohorts. Z-score test: **P* < 0.05, ***P* < 0.01, ****P* < 0.001, *****P* < 0.0001.

the recurrence risk of stage II/III CRC. For each signature, we further calculated the C-indices of all cohorts. As illustrated in Figure 5b, our CMDLncS model ranked first in predictive power among GSE39582, GSE143985, GSE161158, GSE17536, GSE 29,621, GSE92921, and meta-cohort. Meanwhile, CMDLncS ranked second in TCGA-CRC, weaker than ColoFinder, and fourth in GSE31593, following Chu, Ge, Liu-FIMB-1 (Figure 5b). Our CMDLncS model demonstrated the better performance in each cohort than almost all models (z-score test: $P < 0.05$, Figure 5b). Of note, the performance of most signatures was powerful in their own training cohort but weak in some external cohorts. For example, the C-index of ColoFinder was 0.979 [0.970−0.987] in its own training cohort TCGA-CRC, which suggested extremely accurate performance, but the C-indices in other cohorts were all less than 0.6 (Figure 5b). Similarly, Chu performed best in

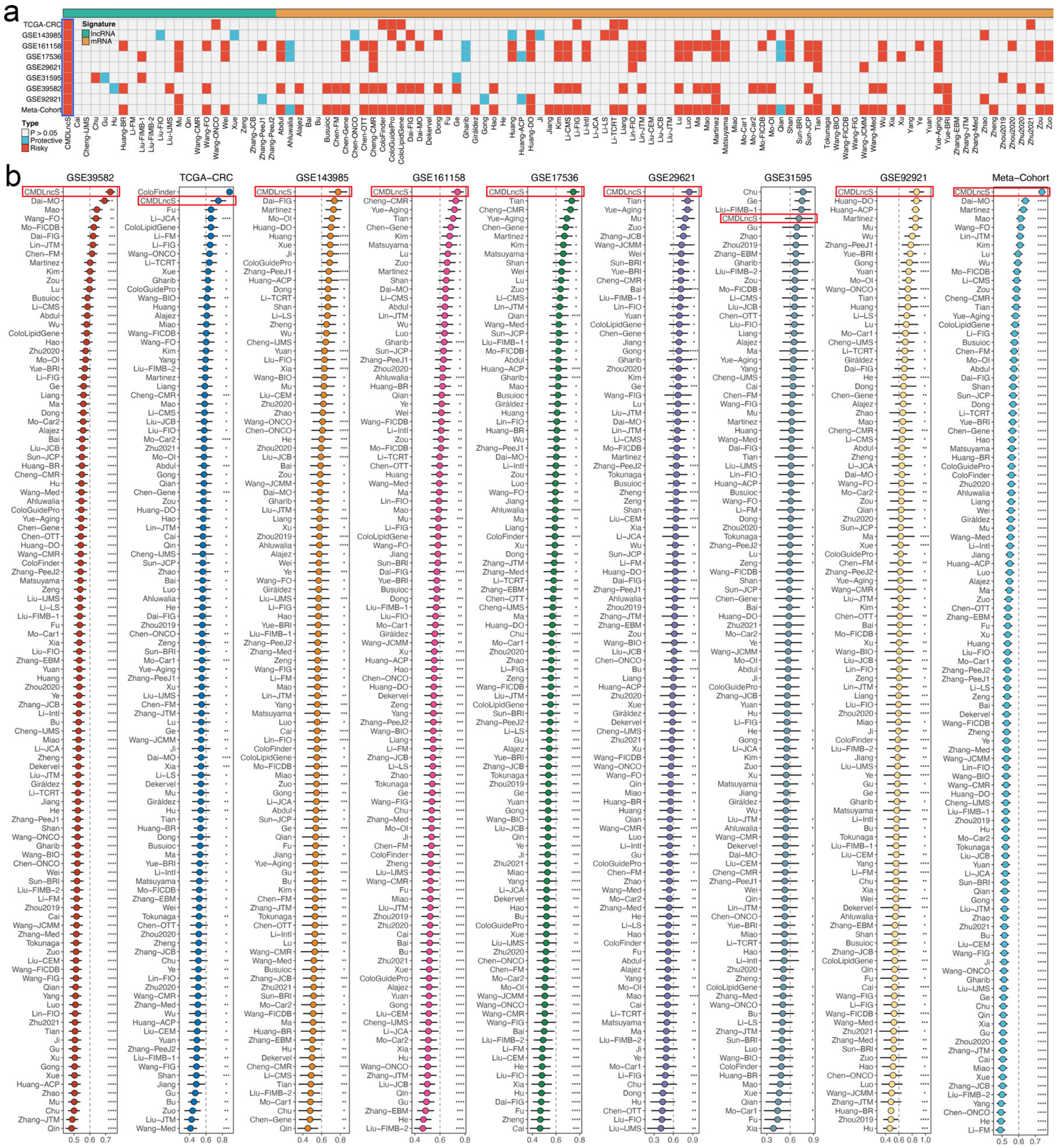


Figure 5. Comparisons of gene expression signatures. a. Univariate Cox regression analysis of CMDLncS and 109 published signatures. b. C-indices of CMDLncS and 109 published signatures in GSE39582, TCGA-CRC, GSE143985, GSE161158, GSE17536, GSE29621, GSE31595, GSE92921, and meta-cohort. Z-score test: *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$.

GSE31595, superior to CMDLncS, but its performance in other cohorts was quite poor, and the C-indices of some cohorts were even less than 0.5 (Figure 5b). The weak generalization ability may arise from model overfitting, though our signature was dimensionally reduced by two machine learning algorithms and thus had a better extrapolation possibility.

Verification of CMDLncS via qRT-PCR

To further validate the robustness and reproducibility of our CMDLncS model as a clinically translatable tool, we next quantified the expression of these 12 SRRLs in a clinical in-house cohort of 173 patients with stage II/III CRC via qRT-PCR assay. In line with our prior results, Kaplan-Meier survival analysis revealed that patients

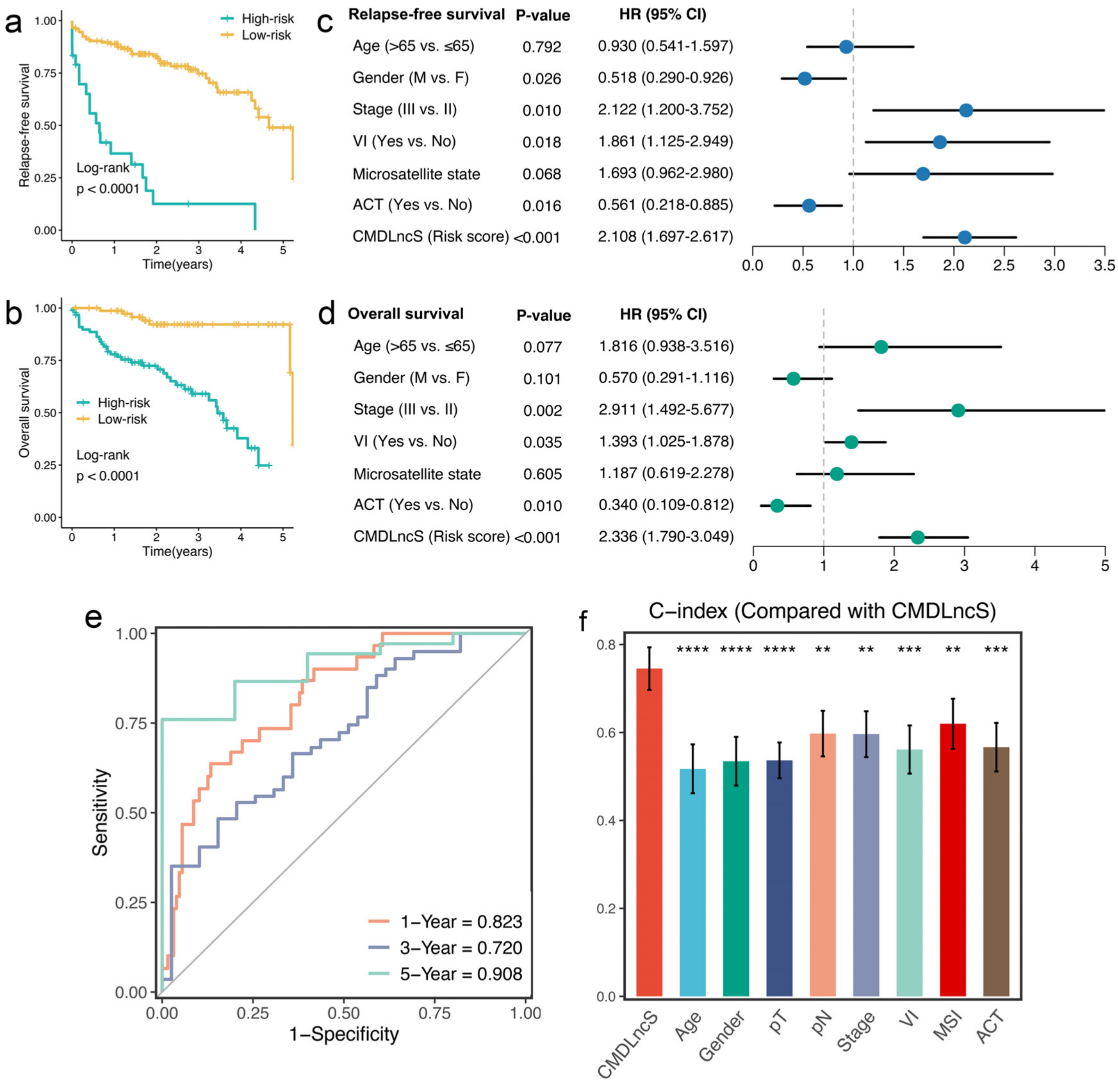


Figure 6. Verification of CMDLncS via qRT-PCR. a, b. Kaplan-Meier curves of RFS (a) and OS (b) according to CMDLncS. c, d. Multivariable Cox regression analysis of RFS (c) and OS (d). e. Time-dependent ROC analysis for predicting RFS at 1, 3, and 5 years. f. The performance of CMDLncS was compared with common clinical and molecular variables in predicting prognosis. Z-score test: **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$.

with high CMDLncS possessed significantly dismal RFS and OS (log-rank test: $P < 0.0001$) (Figure 6a, b). After adjusting for confounding variables, our CMDLncS model remained statistically significant for RFS and OS (Wald test: $P < 0.001$) (Figure 6c, d). The time-dependent ROC analysis showed that the AUCs were 0.823, 0.720, and 0.908 at 1, 3 and 5 years of recurrence, respectively (Figure 6e). The C-index of CMDLncS was 0.745 [0.697–0.793], which was pronouncedly superior to common variables of clinical settings, including age, gender, T, N, AJCC stage, VI, microsatellite state, and ACT (Figure 6f). Overall, the results from the clinical in-house cohort supported our in-silico findings, which further verified that our CMDLncS model was quite feasible and reproducible for stage II/III CRC patients after radical surgery.

Predictive value of fluorouracil-based ACT and bevacizumab benefits

Elegant studies have revealed that lncRNAs are intimately associated with the sensitivity and resistance of fluorouracil-based ACT and bevacizumab.[15,26–28] Therefore, we further explored the predictive value of CMDLncS for assessing the benefits of fluorouracil-based ACT and bevacizumab in stage II/III CRC. Seven cohorts treated with fluorouracil-based ACT were retrieved, which contained 187 nonresponders and 191 responders. Among these seven cohorts and the combined meta-cohort, responders displayed higher CMDLncS than nonresponders (Figure 7a), suggesting that patients with a high CMDLncS might benefit more from fluorouracil-based ACT. Notably, the CMDLncS difference between the two groups in GSE62080 was not statistically significant (T-test: $P = 0.087$, Figure 7a), which might be due to the small sample size ($n = 21$). ROC analysis displayed that CMDLncS had a robust classification performance for assessing the benefit of fluorouracil-based ACT, with relatively high AUC, sensitivities, and specificities in seven therapy cohorts and the combined meta-cohort (Figure 7b). Moreover, compared with the four published signatures, our CMDLncS model showed more stable and better performance (Fig. S5a).

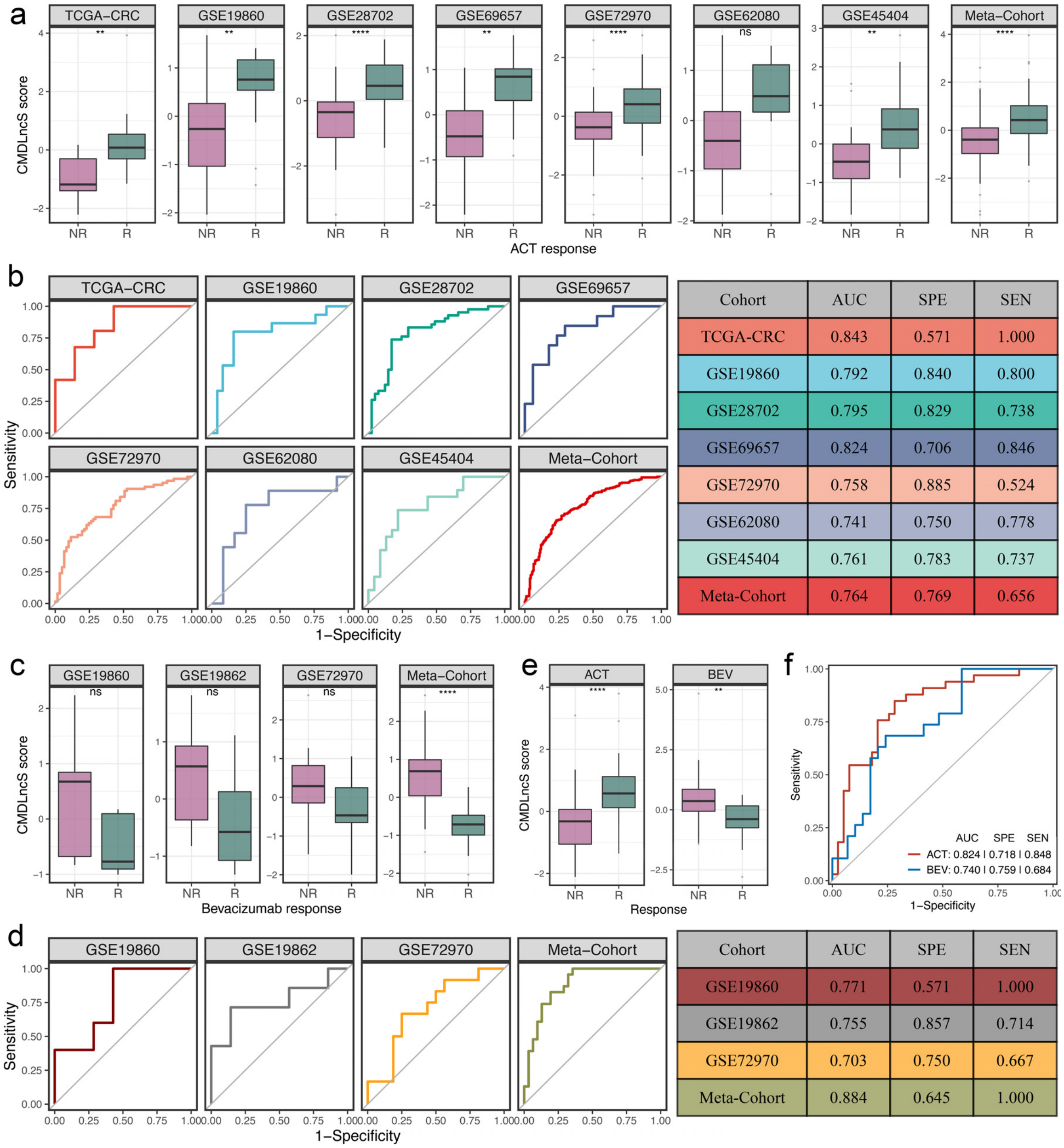
In addition, three cohorts (GSE19860, GSE19862, and GSE72970) treated with bevacizumab, including 30 nonresponders and 24 responders, were also enrolled. Compared to fluorouracil-based ACT alone, the CMDLncS level of patients sensitive to bevacizumab exhibited a trend of lower CMDLncS levels compared with those patients resistant to bevacizumab in GSE19860 (T-test: $P = 0.150$), GSE19862 (T-test: $P = 0.130$), GSE72970 (T-test: $P = 0.074$), and the combined meta-cohort (T-test: $P < 0.0001$) (Figure 7c). This indicated that patients with a low CMDLncS might be sensitive to bevacizumab. In parallel, our CMDLncS model also demonstrated high AUCs, sensitivities, and specificities in predicting the benefit of bevacizumab across these three cohorts and the combined meta-cohort (Figure 7d). Likewise, compared with the three published signatures, we observed the similar results (Fig. S5b). Of note, the AUC of CMDLncS was slightly lower than that of Abajo-BJC (0.703 vs. 0.750), but the AUCs of CMDLncS were the highest in the GSE19860, GSE19862, and meta-cohort (Fig. S5b). Overall, our CMDLncS model still had an obvious advantage in predicting fluorouracil-based ACT and bevacizumab benefits.

To further verify the clinical significance of CMDLncS for assessing the benefits of fluorouracil-based ACT and bevacizumab in our clinical in-house cohort. We collected 39 nonresponders and 33 responders to fluorouracil-based ACT as well as 29 nonresponders and 19 responders to bevacizumab (Table S1). In line with our in-silico findings, patients with a high CMDLncS tended to benefit from fluorouracil-based ACT and were resistant to bevacizumab (Figure 7e). ROC analysis further demonstrated the accurate performance of our CMDLncS model (Figure 7f).

Biological mechanisms underlying CMDLncS

To decipher the biological mechanisms underlying CMDLncS, we performed GSEA on 21,338 gene sets from the MSigDB resource. Based on the normalized enrichment score (NES), we identified the top 20 pathways that were positively and negatively correlated with CMDLncS. As displayed in Figure 8a, patients with a high CMDLncS enriched numerous pathways related to stroma and EMT activation, such as epithelial mesenchymal transition, extracellular matrix structural constituent, and collagen containing extracellular matrix, while patients with a low CMDLncS were mainly associated with immune activation, such as inflammatory response, myeloid leukocyte migration, and cytokine binding. We next measured EMT, stromal, immune, and TIS scores via multiple bioinformatic algorithms (including GSVA, ESTIMATE, and ssGSEA). Consistently, CMDLncS was significantly positively associated with EMT and stromal scores, and negatively asscoaited with immune and TIS scores (Figure 8b–e), which further verified the prior results. In addition, we investigated the associations between CMDLncS and consensus molecular subtypes (CMS1–4). As illustrated in Fig. S6a, CMS4 subtype displayed higher CMDLncS score relative to the other subtypes. As is well known, CMS4 was associates with a mesenchymal phenotype and poor prognosis, in line with the indications of high CMDLncS. Furthermore, we plotted ROC curves to further evaluate the accuracy of CMDLncS in the identification of CMS4 CRC patients, and the AUC for CMDLncS was remarkably high: 0.962 (Fig. S6b).

To gain more detailed insights into the tumor microenvironment, our study further quantified the infiltration abundance of 28 immune cell populations. With the increase in CMDLncS, the overall immune infiltration level was weakened (Figure 9a). At both ends of CMDLncS, we distinctly observed small subsets of immune-hot and immune-cold phenotypes (Figure 9a). Indeed, most immune cells demonstrated significantly inverse correlations with CMDLncS and possessed superior infiltration in the low-risk group (Figure 9b, c). Overall, patients with a low CMDLncS displayed a stronger immune infiltration, which might explain their better prognosis. Furthermore, we extended our analysis to encompass 27 immune checkpoint members, including the B7-CD28 family, TNF superfamily, and several other molecules, and the results are displayed in Fig. S7a. Patients with a low CMDLncS presented significantly higher *CD27, CD40, CD70, ENTPD1, HAVCR2, PD-L1, PD-L2*, and *TNFRSF4* expression levels, while patients with a low CMDLncS displayed a higher *HHLA2* expression level (Fig. S7b, c). These findings might provide references for the development and management of immunotherapy.



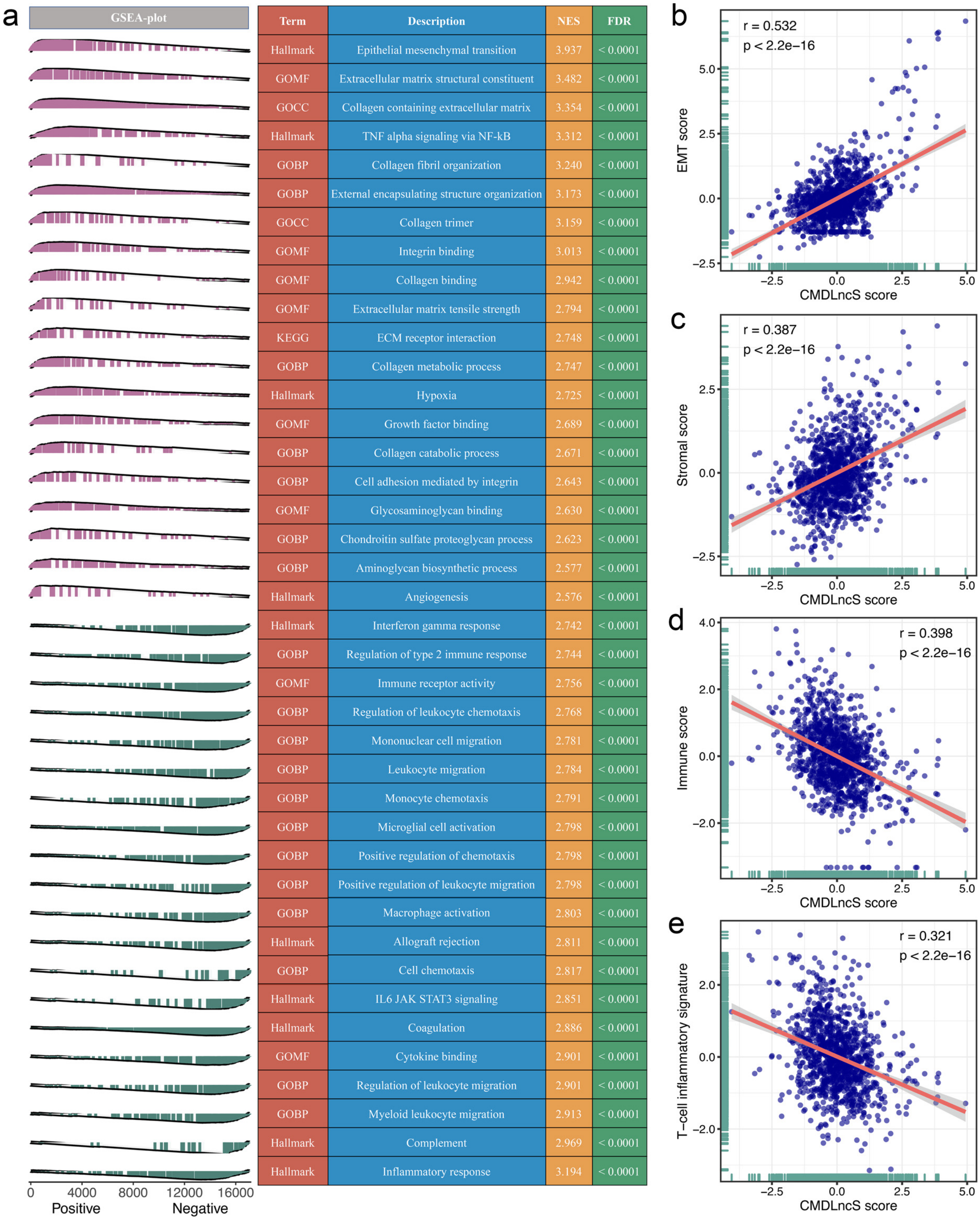
Cohort	AUC	SPE	SEN
TCGA-CRC	0.843	0.571	1.000
GSE19860	0.792	0.840	0.800
GSE28702	0.795	0.829	0.738
GSE69657	0.824	0.706	0.846
GSE72970	0.758	0.885	0.524
GSE62080	0.741	0.750	0.778
GSE45404	0.761	0.783	0.737
Meta-Cohort	0.764	0.769	0.656

Cohort	AUC	SPE	SEN
GSE19860	0.771	0.571	1.000
GSE19862	0.755	0.857	0.714
GSE72970	0.703	0.750	0.667
Meta-Cohort	0.884	0.645	1.000

Figure 7. Predictive value of fluorouracil-based ACT and bevacizumab benefits. a. Distributions of CMDLncS between responders and nonresponders of fluorouracil-based ACT. b. ROC curves of CMDLncS to predict the benefits of fluorouracil-based ACT. c. Distributions of CMDLncS between responders and nonresponders of bevacizumab. d. ROC curves of CMDLncS to predict the benefits of bevacizumab. e. Distribution of CMDLncS between responders and nonresponders regarding fluorouracil-based ACT and bevacizumab in our in-house cohort, respectively. f. ROC curves regarding fluorouracil-based ACT and bevacizumab in our in-house cohort, respectively. T-test: $^{ns}P > 0.05$, $^{*}P < 0.05$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

Discussion

The limitations of the current TNM staging system hamper the capacity to offer optimal clinical care to patients, as the clinical management of stage II/III CRC is primarily determined by clinicopathological staging, without regard to molecular biological characteristics.[29] LncRNAs, as a novel class of noncoding RNA, have profound impacts on tumorigenesis, progression, recurrence, metastasis, and drug sensitivity in tumors [15,26–28,30]. Our study systematically established links between lncRNA profiles and recurrence, prognosis, and benefits from fluorouracil-based ACT and bevacizumab



Term	Description	NES	FDR
Hallmark	Epithelial mesenchymal transition	3.937	< 0.0001
GOMF	Extracellular matrix structural constituent	3.482	< 0.0001
GOCC	Collagen containing extracellular matrix	3.354	< 0.0001
Hallmark	TNF alpha signaling via NF-kB	3.312	< 0.0001
GOBP	Collagen fibril organization	3.240	< 0.0001
GOBP	External encapsulating structure organization	3.173	< 0.0001
GOCC	Collagen trimer	3.159	< 0.0001
GOMF	Integrin binding	3.013	< 0.0001
GOMF	Collagen binding	2.942	< 0.0001
GOMF	Extracellular matrix tensile strength	2.794	< 0.0001
KEGG	ECM receptor interaction	2.748	< 0.0001
GOBP	Collagen metabolic process	2.747	< 0.0001
Hallmark	Hypoxia	2.725	< 0.0001
GOMF	Growth factor binding	2.689	< 0.0001
GOBP	Collagen catabolic process	2.671	< 0.0001
GOBP	Cell adhesion mediated by integrin	2.643	< 0.0001
GOMF	Glycosaminoglycan binding	2.630	< 0.0001
GOBP	Chondroitin sulfate proteoglycan process	2.623	< 0.0001
GOBP	Aminoglycan biosynthetic process	2.577	< 0.0001
Hallmark	Angiogenesis	2.576	< 0.0001
Hallmark	Interferon gamma response	2.742	< 0.0001
GOBP	Regulation of type 2 immune response	2.744	< 0.0001
GOMF	Immune receptor activity	2.756	< 0.0001
GOBP	Regulation of leukocyte chemotaxis	2.768	< 0.0001
GOBP	Mononuclear cell migration	2.781	< 0.0001
GOBP	Leukocyte migration	2.784	< 0.0001
GOBP	Monocyte chemotaxis	2.791	< 0.0001
GOBP	Microglial cell activation	2.798	< 0.0001
GOBP	Positive regulation of chemotaxis	2.798	< 0.0001
GOBP	Positive regulation of leukocyte migration	2.798	< 0.0001
GOBP	Macrophage activation	2.803	< 0.0001
Hallmark	Allograft rejection	2.811	< 0.0001
GOBP	Cell chemotaxis	2.817	< 0.0001
Hallmark	IL6 JAK STAT3 signaling	2.851	< 0.0001
Hallmark	Coagulation	2.886	< 0.0001
GOMF	Cytokine binding	2.901	< 0.0001
GOBP	Regulation of leukocyte migration	2.901	< 0.0001
GOBP	Myeloid leukocyte migration	2.913	< 0.0001
Hallmark	Complement	2.969	< 0.0001
Hallmark	Inflammatory response	3.194	< 0.0001

Figure 8. Biological mechanisms underlying CMDLncS. a. Top 20 pathways that were positively and negatively correlated with CMDLncS. b−e. Correlations of CMDLncS with EMT (b), stromal (c), immune (d), and TIS scores (e).

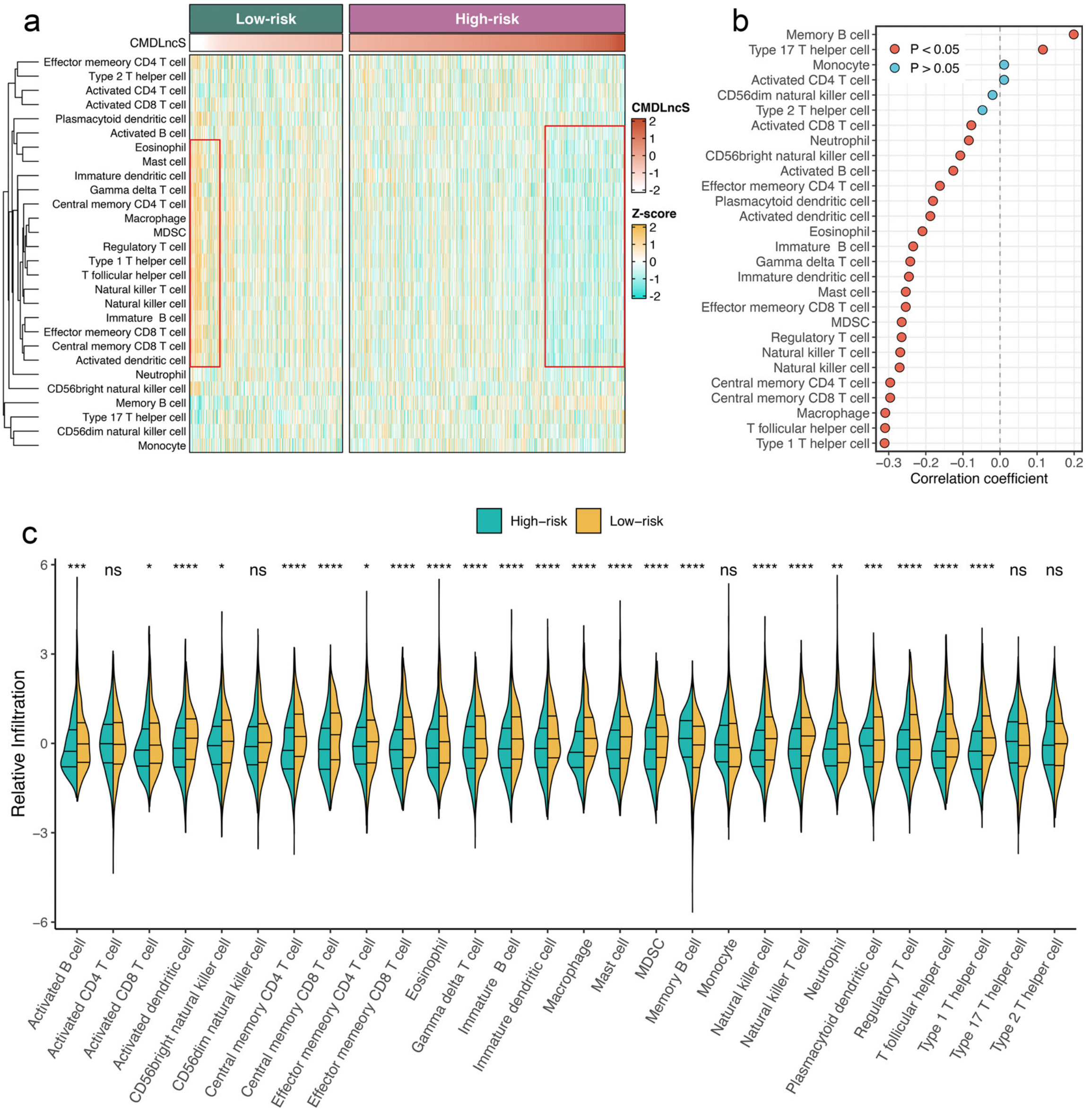


Figure 9. Immune landscape of CMDLncS. a. Heatmap of 28 immune cells infiltration. b. Relationship between CMDLncS and immune cell infiltrations. c. Distributions of 28 immune cells infiltration between high- and low-risk groups. T-test: $^{ns}P > 0.05$, $^{*}P < 0.05$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

in stage II/III CRC, providing an attractive platform for detecting high-risk patients and further improving the clinical outcomes in stage II/III CRC.

In this study, a total of 27 SRRLs that were stably associated with recurrence were identified in stage II/III CRC using eight independent cohorts, which offered a resource for developing biomarkers. With advancements in high-throughput sequencing techniques and computational biology, numerous predictive gene expression signatures have been proposed according to various machine learning approaches.[25] However, two questions worth considering are why a particular algorithm should be used and which solution is the optimal one. The selection of algorithms by researchers may rely largely on their own preferences and bias. Thus, to generate a consensus signature, we collected 10 prevalent algorithms and then combined them into 76 combinations. The algorithm combinations can also further reduce the dimension of variables, making the signature more simplified and feasible. Ultimately, the optimal model (CMDLncS) was determined by a combination of Lasso and stepwise Cox with the highest average C-index (0.777), which was also the only model with a C-index above 0.7 in seven independent validation cohorts. Our CMDLncS model was proven to be an independent risk factor and maintained robust performance in all cohorts. However, the power of CMDLncS in clinical practice is not yet known. To test the clinical interpretation of CMDLncS, another validation based on qRT-PCR results from 173 frozen CRC tissues, further verified our prior findings and assessed its feasibility in different centers. Therefore, our CMDLncS model has great potential for clinical application in patients with stage II/III CRC.

Clinical traits (e.g., AJCC stage) and emerging molecular alterations (e.g., microsatellite state, *KRAS* mutation) are still cornerstones for assessing prognosis and determining treatment options for stage II/III CRC.[2,7] For example, mutations in *KRAS* and *BRAF* tend to suggest a worse prognosis in stage II CRC, while high microsatellite instability (MSI-H) is a protective prognostic indicator and has predictive value for guiding fluorouracil-based ACT in stage II/III CRC.[2] Thus, our study compared the predictive superiority of CMDLncS with common clinical traits and molecular features in predicting the recurrence risk of stage II/III CRC after radical surgery. In eight public cohorts and our clinical in-house cohort, CMDLncS was not only able to work independently of these factors, such as age, gender, T, N, AJCC stage, VI, TMB, NAL, microsatellite state, ACT, and *TP53, KRAS*, or *BRAF* mutations, but also presented dramatically superior performance in evaluating the recurrence risk according to the C-index assessment. This indicated that our CMDLncS model could be a promising surrogate for evaluating the recurrence risk of stage II/III CRC in clinical practice.

Additionally, we enrolled 109 published signatures composed of various functional gene combinations. Among these signatures, very few have been translated into clinical settings and even fewer have been rigorously validated.[2] For example, univariate Cox regression analysis exhibited that except for CMDLncS, no signature maintained complete significance across all datasets. Compared with these signatures, our CMDLncS model also possessed relatively superior performance in each cohort. Most signatures were usually powerful within their own training cohort but weak in some external cohorts, such as ColoFinder and Chu.[31,32] The poor generalization ability may arise from model overfitting, though our signature was dimensionally reduced by two algorithms and fitted based on SRRLs that were stably associated with recurrence, thus having a better extrapolation possibility.

CMDLncS quantifies the recurrence risk at the individual patient level and could stratify potential “high-risk” or “low-risk” patients. Thus, reasonable interventions of patients with different levels of CMDLncS are currently essential. Indeed, our CMDLncS model also has great implications in predicting drug benefits. A high CMDLncS indicated sensitivity to fluorouracil-based ACT alone and resistance to bevacizumab. ROC analysis suggested that our CMDLncS model afforded greater accuracy in predicting fluorouracil-based ACT and bevacizumab benefits. These findings have far-reaching meaning for selecting treatment strategies in the era of precision medicine. For example, current guidelines recommend that ACT is not required for a subset of stage II patients without high-risk clinical traits,[2] but when these patients display a high CMDLncS, using additional ACT could be essential, and bevacizumab may not be necessary due to potential resistance. Thus, the CMDLncS system could also serve as a powerful tool for optimizing decision-making for stage II/III CRC patients.

Afterwards, our study also explored the latent biological mechanisms underlying CMDLncS. Patients with a high CMDLncS demonstrated stroma and EMT activation, which may explain their high rates of recurrence and mortality as well as the sensitivity of fluorouracil-based ACT.[33,34] As previously reported, stromal cells could promote progression and chemoresistance via enhancing cell stemness and EMT in CRC.[33] In contrast, patients with a low CMDLncS were mainly associated with immune activation, which paralleled their prognostic outcomes. The superior immune infiltrations also indicated that these patients might have stronger potential for immunotherapy. As is well-known, immune checkpoint inhibitors (ICIs) have emerged as a revolutionary modality of cancer immunotherapy that function by targeting immune checkpoints, such as *PD-1, PD-L1*, and *CD40*.[35] Our study showed that patients with a low CMDLncS presented significantly higher *CD27, CD40, CD70, ENTPD1, HAVCR2, PD-L1, PD-L2*, and *TNFRSF4* expression levels, which supported the conclusion that these patients may benefit more from current immunotherapy. Of note, patients with a low CMDLncS displayed a higher *HHLA2* expression level, which is a newly discovered immune checkpoint molecule that is positively associated with a high mortality rate and negatively related to CD8+ T cell infiltration in CRC patients.[36] Recently, *HHLA2* was reported to be broadly expressed in patients with *PD-1* negative human tumors,[37] indicating that *HHLA2* might be a promising therapeutic target for patients who do not respond to *PD-1/PD-L1* pathway inhibitors, similar to latent patients with a high CMDLncS. Thus, targeting *HHLA2* as an immune stimulator may become a valuable approach to improve the clinical outcomes of patients with a high CMDLncS.

Although the implications of CMDLncS in stage II/III CRC are profound, some limitations should be acknowledged. First, all the samples from this study were retrospective, and future verification of CMDLncS should be conducted in a prospective multi-center study. Second, some clinical and molecular traits on public datasets were quite inadequate, which thus had concealed the potential associations between CMDLncS and some traits. Third, the functions of most lncRNAs from CMDLncS in stage II/III CRC remain to be elucidated, and further *in vivo* and *in vitro* experiments are needed to reveal their roles. Finally, combining mRNA and lncRNA expression, may provide a more powerful signature, which is worth further exploration.

Taken together, our study comprehensively explored the clinical significance of lncRNAs in stage II/III CRC and systematically identified a consensus lncRNA signature (termed CMDLncS) that could independently predict the recurrence and prognosis of stage II/III CRC patients. We compared CMDLncS with common clinical traits, molecular features, and 109 published signatures to further verify its robustness and translation. Our CMDLncS model also had great implications in predicting the benefits of fluorouracil-based ACT and bevacizumab. Overall, CMDLncS could serve as a promising tool to optimize decision-making and surveillance protocols for stage II/III CRC patients.

Contributors

Zaoqu Liu made the conceptualization, data curation, formal analysis, investigation, methodology, resources, software, validation, visualization, and writing – original draft. Xinwei Han and Zhenqiang Sun made the conceptualization, funding acquisition, project administration, supervision, and writing – review & editing. Chunguang Guo, Qin Dang, Libo Wang, Long Liu, Siyuan Weng, Hui Xu, and Taoyuan Lu made the writing – review & editing. Zaoqu Liu, Zhenqiang Sun, Xinwei Han, and Qin Dang have verified the underlying data. All authors read and approved the final version of the manuscript.

Data sharing statement

Public data used in this work can be acquired from the TCGA Research Network portal (https://portal.gdc.cancer.gov/) and Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/). The raw experimental data and analysis codes supporting the conclusions of this article will be made available by the corresponding author.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Acknowledgments

We are indebted and thankful to all participants for their valuable contributions. This study was supported by the National Natural Science Foundation of China (81972663); Henan Province Young and Middle−Aged Health Science and Technology Innovation Talent Project (YXKC2020037); and Henan Provincial Health Commission Joint Youth Project (SB201902014). We thank the contribution of the appropriate specimen donors and research groups.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ebiom.2021.103750.

References

1 Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71(3):209–49.

2 Koncina E, Haan S, Rauh S, Letellier E. Prognostic and predictive molecular biomarkers for colorectal cancer: updates and challenges. Cancers (Basel) 2020;12(2):319.

3 Punt CJ, Koopman M, Vermeulen L. From tumour heterogeneity to advances in precision treatment of colorectal cancer. Nat Rev Clin Oncol 2017;14(4):235–46.

4 Auclin E, Zaanan A, Vernerey D, Douard R, Gallois C, Laurent-Puig P, et al. Subgroups and prognostication in stage III colon cancer: future perspectives for adjuvant therapy. Ann Oncol 2017;28(5):958–68.

5 Varghese A. Chemotherapy for stage II colon cancer. Clin Colon Rectal Surg 2015;28(4):256–61.

6 Johnston PG. Stage II colorectal cancer: to treat or not to treat. Oncologist 2005;10(5):332–4.

7 Dienstmann R, Villacampa G, Sveen A, Mason MJ, Niedzwiecki D, Nesbakken A, et al. Relative contribution of clinicopathological variables, genomic markers, transcriptomic subtyping and microenvironment features for outcome prediction in stage II/III colorectal cancer. Ann Oncol 2019;30(10):1622–9.

8 Liu Z, Wang L, Guo C, Liu L, Jiao D, Sun Z, et al. TTN/OBSCN 'double-hit' predicts favourable prognosis, 'immune-hot' subtype and potentially better immunotherapeutic efficacy in colorectal cancer. J Cell Mol Med 2021;25(7):3239–51.

9 Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. Nat Med 2015;21(11):1350–6.

10 Isella C, Brundu F, Bellomo SE, Galimi F, Zanella E, Porporato R, et al. Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. Nat Commun 2017;8:15107.

11 Qian Y, Daza J, Itzel T, Betge J, Zhan T, Marme F, et al. Prognostic cancer gene expression signatures: current status and challenges. Cells 2021;10(3):648.

12 Pages F, Mlecnik B, Marliot F, Bindea G, Ou FS, Bifulco C, et al. International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. Lancet 2018;391(10135):2128–39.

13 Lin C, Yang L. Long noncoding RNA in cancer: wiring signaling circuitry. Trends Cell Biol 2018;28(4):287–301.

14 Stein LD. Human genome: end of the beginning. Nature 2004;431(7011):915–6.

15 Zhang Y, Xu M, Sun Y, Chen Y, Chi P, Xu Z, et al. Identification of LncRNAs associated with FOLFOX chemoresistance in mCRC and construction of a predictive model. Front Cell Dev Biol 2020;8:609832.

16 Liu C, Hu C, Li J, Jiang L, Zhao C. Identification of epithelial-mesenchymal transition-related lncRNAs that associated with the prognosis and immune microenvironment in colorectal cancer. Front Mol Biosci 2021;8:633951.

17 Kelley RK, Venook AP. Prognostic and predictive markers in stage II colon cancer: is there a role for gene expression profiling? Clin Colorectal Cancer 2011;10(2):73–80.

18 Sztupinszki Z, Gyorffy B. Colon cancer subtypes: concordance, effect on survival and selection of the most representative preclinical models. Sci Rep 2016;6:37169.

19 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 Med 2013;10(5):e1001453.

20 Liu Z, Zhang Y, Dang Q, Wu K, Jiao D, Li Z, et al. Genomic alteration characterization in colorectal cancer identifies a prognostic and metastasis biomarker: FAM83A|IDO1. Front Oncol 2021;11:632430.

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

22 Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun 2013;4:2612.

23 Ayers M, Lunceford J, Nebozhyn M, Murphy E, Loboda A, Kaufman DR, et al. IFN-gamma-related mRNA profile predicts clinical response to PD-1 blockade. J Clin Invest 2017;127(8):2930–40.

24 Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. Cell Rep 2017;18(1):248–62.

25 Ahluwalia P, Kolhe R, Gahlay GK. The clinical relevance of gene expression based prognostic signatures in colorectal cancer. Biochim Biophys Acta Rev Cancer 2021;1875(2):188513.

26 Grepin R, Guyot M, Dumond A, Durivault J, Ambrosetti D, Roussel JF, et al. The combination of bevacizumab/avastin and erlotinib/tarceva is relevant for the treatment of metastatic renal cell carcinoma: the role of a synonymous mutation of the EGFR receptor. Theranostics 2020;10(3):1107–21.

27 Li Y, Jiang T, Zhou W, Li J, Li X, Wang Q, et al. Pan-cancer characterization of immune-related lncRNAs identifies potential oncogenic biomarkers. Nat Commun 2020;11(1):1000.

28 Huan L, Guo T, Wu Y, Xu L, Huang S, Xu Y, et al. Hypoxia induced LUCAT1/PTBP1 axis modulates cancer cell viability and chemotherapy response. Mol Cancer 2020;19(1):11.

29 Weiser MR. AJCC 8th edition: colorectal cancer. Ann Surg Oncol 2018;25(6):1454–5.

30 Kim T, Croce CM. Long noncoding RNAs: undeciphered cellular codes encrypting keys of colorectal cancer pathogenesis. Cancer Lett 2018;417:89–95.

31 Shi M, He J. ColoFinder: a prognostic 9-gene signature improves prognosis for 871 stage II and III colorectal cancer patients. PeerJ 2016;4:e1804.

32 Chu Y, Liu Z, Liu J, Yu L, Zhang D, Pei F. Characterization of lncRNA-perturbed TLR-signaling network identifies novel lncRNA prognostic biomarkers in colorectal cancer. Front Cell Dev Biol 2020;8:503.

33 Hu JL, Wang W, Lan XL, Zeng ZC, Liang YS, Yan YR, et al. CAFs secreted exosomes promote metastasis and chemotherapy resistance by enhancing cell stemness and epithelial-mesenchymal transition in colorectal cancer. Mol Cancer 2019;18(1):91.

34 Matsuyama T, Ishikawa T, Takahashi N, Yamada Y, Yasuno M, Kawano T, et al. Transcriptomic expression profiling identifies ITGBL1, an epithelial to mesenchymal transition (EMT)-associated gene, is a promising recurrence prediction biomarker in colorectal cancer. Mol Cancer 2019;18(1):19.

35 Mahoney KM, Rennert PD, Freeman GJ. Combination cancer immunotherapy and new immunomodulatory targets. Nat Rev Drug Discov 2015;14(8):561–84.

36 Zhu Z, Dong W. Overexpression of HHLA2, a member of the B7 family, is associated with worse survival in human colorectal carcinoma. Onco Targets Ther 2018;11:1563–70.

37 Wang C, Feng H, Cheng X, Liu K, Cai D, Zhao R. Potential therapeutic targets of B7 family in colorectal cancer. Front Immunol 2020;11:681.