Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Transcriptomic and machine-learning analyses reveal dephosphorylation-related prognostic genes in colorectal cancer

**Qiao Deng**
The Third Affiliated Hospital of CQMU (FangDa Hospital)
**Lv Hu**
9355495840qq.com

The Third Affiliated Hospital of CQMU (FangDa Hospital)

**Research Article**

**Keywords:** colorectal cancer, dephosphorylation, prognostic model, machine learning, tumor immune microenvironment

**Posted Date:** November 20th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11528084/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Transcriptomic and machine-learning analyses reveal dephosphorylation-related prognostic genes in colorectal cancer

Qiao Deng[1],

[1]The Third Affiliated Hospital of CQMU (FangDa Hospital)., ChongQing, 401120 , China

[*]**Correspondence:** Hu Lv;email: 935549584@qq.com

**ORCID:0009-0002-8419-8693**

**Author information**

**Qiao Deng, E-mail: 310501148@qq.com**

## Abstract

**Background:** Protein dephosphorylation is a reversible post-translational event that influences signaling networks, cellular fate, and adaptation to therapy. However, whether dephosphorylation-related genes (DRGs) have prognostic utility in colorectal cancer (CRC) remains unclear.

**Methods:** The Cancer Genome Atlas colon adenocarcinoma (TCGA-COAD) and TCGA-rectum adenocarcinoma (READ) were combined as the training cohort, and GSE17537 as the validation cohort. Prognostic DRGs were screened by differential-expression analysis and univariate Cox regression, followed by proportional-hazards assessment and least absolute shrinkage and selection operator (LASSO) regression. A risk model was evaluated by Kaplan-Meier analysis and time-dependent receiver operating characteristic (ROC) curves. Further analyses comprised gene set enrichment analysis (GSEA), immune infiltration analysis, tumor immune dysfunction and exclusion (TIDE) scoring, reverse transcription quantitative

polymerase chain reaction (RT-qPCR), and drug-sensitivity prediction.

**Results:** SNAI1, TCF7L1, and TIMP1 formed the prognostic signature, and survival differed significantly between high- and low-risk patients. The training cohort yielded 1-, 2-, and 3-year areas under the curve (AUCs) of 0.604, 0.626, and 0.658, and validation-cohort values of 0.687, 0.764, and 0.652. In high-risk tumors, extracellular-matrix and cell-adhesion programs were preferentially enriched, alongside broad immune-infiltration differences. The low-risk group showed greater predicted immunotherapy response potential and distinct drug sensitivity. Experimental validation confirmed higher SNAI1 expression in tumor tissues and higher TCF7L1 expression in adjacent non-tumor tissues, whereas TIMP1 did not differ significantly between groups.

**Conclusion:** This three-gene dephosphorylation-related signature stratified colorectal cancer prognosis and linked dephosphorylation signaling to extracellular-matrix remodeling, immune features, and treatment sensitivity; further validation is warranted

**Keywords:** colorectal cancer; dephosphorylation; prognostic model; machine learning; tumor immune microenvironment

## 1. Introduction

Colorectal cancer (CRC) represents a major global cancer burden and is among the malignancies most frequently diagnosed in both sexes, with mortality remaining substantial [1]. Its progression involves disruption of normal intestinal epithelial differentiation together with excessive proliferation, evasion of cell death, and early

acquisition of metastatic capacity [2]. Current management incorporates surgery, chemotherapy, radiotherapy, targeted agents, and immunotherapy; nevertheless, advanced CRC continues to have an unfavorable prognosis because invasion, distant dissemination, and therapeutic resistance remain common [2, 3]. Such clinical behavior arises from the interplay of genomic abnormalities, epigenetic changes, and altered signaling. In metastatic disease, epithelial plasticity, stromal reorganization, and immune escape can jointly shape recurrence and survival, underscoring the value of biomarkers with biological relevance [2]. The availability of large-scale molecular datasets now permits survival patterns to be examined in relation to these alterations. Prognostic models therefore have greater interpretive value when considered together with the biological programs that maintain invasion, metastatic spread, and adaptation to treatment rather than being viewed only as statistical classifiers [2]. Identifying prognostically informative genes and defining the processes linked to CRC progression consequently remain important for more accurate risk assessment and treatment planning.

Phosphatases chiefly catalyze protein dephosphorylation, a reversible mode of post-translational regulation. Through this activity, dephosphorylation participates in control of the cell cycle, intracellular signaling, metabolic balance, stress responses, and programmed cell death [4]. Evidence accumulated in recent years further implicates phosphatase-regulated signaling in tumor initiation, maintenance of stem-like traits, resistance to therapy, and remodeling of antitumor immunity [4]. These regulatory effects are increasingly relevant to CRC. Phosphatase-centered pathways

can influence both oncogenic plasticity and the immune microenvironment, raising the possibility that DRG-based classification could provide information complementary to transcriptome-based prognostic evaluation [5]. Specific observations support this concept: dephosphorylation of ACOX1 has been linked to palmitate-associated CRC progression, whereas circGPRC5A promotes CRC progression through the UBA1/PPP1CA/YAP axis [6, 7]. PP6 has additionally been documented to promote stemness-associated programs in CRC cells [5]. Even so, a systematic evaluation of dephosphorylation-related genes (DRGs) as prognostic factors in CRC is still lacking, and both their clinical relevance and biological meaning remain incompletely clarified. Here, a curated DRG collection was analyzed together with publicly available CRC transcriptomic cohorts. Univariate Cox analysis combined with machine-learning-based selection first filtered the prognostic candidates, after which a risk model was built and validated in an independent dataset. The model's biological context was then characterized by examining immune infiltration, gene-set enrichment, and predicted drug sensitivity. The purpose was not to assert a fully resolved dephosphorylation mechanism, but to derive a potentially useful prognostic signature and define testable associations between DRGs and CRC progression. The resulting observations are intended to support subsequent mechanistic validation and future approaches to risk-adapted management.

## 2. Materials and methods

### 2.1 Data collection

CRC expression datasets were sourced from The Cancer Genome Atlas (TCGA) database (https://www.cancergenome.nih.gov/) and the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). TCGA_COAD included 455 CRC tissue samples and 41 normal controls, and 387 of the tumor samples also had survival information together with accompanying clinical data. TCGA_READ provided a further 163 CRC tissues and 10 normal controls, and 137 of those CRC cases had both survival and clinical information. TCGA_COAD and TCGA_READ were then combined to establish the training dataset, designated TCGA_CRC. After integration, TCGA_CRC contained 618 tumor tissues and 51 normal tissues, while 524 of the tumor samples had the survival and clinical variables required for subsequent prognostic analyses. GSE17537, generated on the GPL570 platform, was downloaded separately from GEO and used as the validation dataset; it consisted of 55 CRC tissue samples for which survival information was available. Meanwhile, the Molecular Signatures Database (MsigDB) supplied 261 DRGs, which are listed in Supplementary Table 1. The TCGA, GEO, and MsigDB resources were accessed on September 10, 2025.

### 2.2 Identification of candidate genes associated with CRC and dephosphorylation

Using DESeq2 (v1.40.2) [8], tumor-versus-normal differential expression was assessed across all TCGA-CRC samples. Genes with adjusted P (Padj) < 0.05 and |log2fold change (FC)| > 1.5 were defined as DEGs [9]. The resulting DEG distribution was

summarized in a volcano plot drawn with ggplot2 (v3.5.2) [10]. For annotation purposes, the upregulated and downregulated gene sets were each sorted by the absolute magnitude of log2FC, and the 10 genes at either extreme were labeled on the plot. For a sample-level depiction of these leading expression changes, a heatmap was generated with ComplexHeatmap (v2.16.0) [10], displaying the 10 most upregulated and 10 most downregulated DEGs. Overlap analysis was then applied to define dephosphorylation-associated CRC candidates. In particular, a Venn diagram was constructed with ggvenn (v0.1.10) [11] between the DEG set and the 261 DRGs, and genes common to both sets were retained as candidate genes for downstream analysis.

### 2.3 Enrichment analysis and protein-protein interaction (PPI) network of candidate genes

Enrichment analysis was run with clusterProfiler (v4.8.3) [12] to characterize the biological functions represented by the candidate genes, covering both Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG). Terms or pathways were deemed significantly enriched at $P < 0.05$ [13]. Across the three GO components—biological process (BP), cellular component (CC), and molecular function (MF)—GO results were assessed separately, and both GO terms and KEGG pathways were then ordered by increasing P value. Using ggplot2 (v3.5.2) [10], the top five entries for each GO component and the top five KEGG pathways were then displayed. Protein-level relationships among the same candidate genes were independently evaluated via the STRING database (https://string-db.org/). Candidate genes were submitted to STRING; only interactions with a confidence score above 0.9

were retained, and the resulting PPI network was then visualized with ggplot2 (v3.5.2) [10].

### 2.4 Screening of prognostic genes

The candidate-gene set was then narrowed according to its relationship with patient survival. In tumor samples from the training cohort with available survival data, univariate Cox regression was conducted using survival (v3.8-3) [14]. Genes with a hazard ratio (HR) $\neq 1$ and $P < 0.05$ were kept [15], and the estimated effects were shown in a forest plot produced with forestplot (v3.1.3) [16]. Because validity of a Cox model depends on the proportional hazards assumption, each gene passing the initial screen was additionally examined with the cox.zph function in survival (v3.8-3) [14]. Only genes with a PH-test P value > 0.05 [17] were considered candidate prognostic genes. These candidates were then subjected to LASSO regression using glmnet (v4.1-8) [18]. The optimal regularization parameter ($\lambda$) was determined by ten-fold cross-validation, and genes with non-zero contribution at that $\lambda$ were retained as the final prognostic genes.

### 2.5 Evaluation and validation of a prognostic model

The designated prognostic formula was applied to each CRC tumor sample with survival information in the training cohort to derive a risk score: $\text{Risk Score} = \sum_{i=1}^{n} \text{Coefficient of gene (i)} \times \text{Expression of gene (i)}$. For gene i, "Expression" is its measured expression level, while "Coefficient" is the matching coefficient supplied by the prognostic model. The optimal risk-score cutoff classified samples into a high-risk group (HRG) or low-risk group (LRG). All eligible tumors were sorted by ascending

risk score to illustrate the risk–outcome relationship, and both the ranked risk-score plot and the accompanying survival-status scatter plot were generated with ggplot2 (v3.5.2) [10]. Prognostic separation between HRG and LRG was then examined with Kaplan-Meier (KM) analysis generated through survminer (v0.5.0) [19], using $P < 0.05$ as the threshold for a statistically significant survival difference [20]. Independent assessment of time-dependent model discrimination relied on timeROC (v0.4) [21], which generated receiver operating characteristic (ROC) curves for 1-, 2-, and 3-year survival [22]. These area under the curve (AUC) values were compared against the predefined AUC > 0.6 threshold [23]. Finally, the full workflow—risk stratification, KM evaluation, and time-dependent ROC analysis—was rerun in the GSE17537 validation set, enabling the model's reproducibility and reliability to be examined outside the training cohort.

**2.6 Gene set enrichment analysis (GSEA)**

To identify biological pathways with systematic differences between HRG and LRG tumors in the training cohort, GSEA was conducted. The analysis began with an HRG-versus-LRG differential-expression comparison using DESeq2 (v1.40.2) [8], from which log2FC values were obtained for gene ranking. Genes were ordered from the largest to the smallest log2FC, and this ranked list was supplied to clusterProfiler (v4.8.3) [12] for GSEA. The MSigDB file "c2.cp.kegg_legacy.v2025.1.Hs.symbols.gmt" served as the reference collection. Significant enrichment required Padj < 0.05 and |NES| > 1 [24]. The significant pathways were subsequently ordered by NES. For presentation, the five pathways with

the strongest positive enrichment and the five with the strongest negative enrichment in HRG were selected and visualized with enrichplot (v1.20.0) [25] together with ggplot2 (v3.5.2) [10].

**2.7 Immune infiltration analysis**

The training cohort's immune composition was estimated for 28 immune-cell populations [26]. Using GSVA (v1.48.2) [27], the ssGSEA procedure [28] was run on each sample separately to produce immune-cell enrichment scores for HRG and LRG. A heatmap was first generated from these scores using pheatmap (v1.0.12) [29]. HRG and LRG scores underwent Wilcoxon rank-sum testing to identify cell populations whose inferred abundance differed between risk groups; $P < 0.05$ served as the threshold for differentially infiltrated immune cells (DIICs). The between-group distributions of the identified DIICs were illustrated with boxplots generated by ggplot2 (v3.5.2) [10]. Subsequently, correlation structure was evaluated at two levels—DIIC–DIIC relationships and associations between each prognostic gene and the DIICs. Spearman correlations were obtained via psych (v2.5.3) [30], with $P < 0.05$ and $|cor| > 0.3$ applied as thresholds [31]. Significant correlations were summarized graphically using corrplot (v0.92) [29].

**2.8 Tumor immune dysfunction and exclusion (TIDE)**

Differences in predicted immunotherapy-related features between the two risk groups were estimated with the TIDE platform. For every training-set CRC tumor with survival information, the TIDE platform provided an overall TIDE score as well as dysfunction, exclusion, and microsatellite instability (MSI) scores. The four measures

were each compared between HRG and LRG using the Wilcoxon rank-sum test, with $P < 0.05$ considered statistically significant. Distributions of the four scores were visualized with ggplot2 (v3.5.2) [10]. Beyond continuous TIDE-related scores, the same TIDE platform provided the predicted immunotherapy-response classification for each eligible CRC sample. Response proportions were then calculated for HRG and LRG and visualized with ggplot2 (v3.5.2) [10] to compare the predicted responder distribution between the risk strata.

**2.9 Drug-sensitivity prediction**

To assess predicted drug sensitivity, HRG and LRG were compared across 198 compounds from the Genomics of Drug Sensitivity in Cancer (GDSC) database. In every training-set CRC tumor with survival data available, oncoPredict [32] estimated the IC50 for each compound. Here, greater predicted sensitivity to a given drug was inferred from a lower predicted IC50. Predicted IC50 values for HRG and LRG were compared using the Wilcoxon rank-sum test, and $P < 0.05$ indicated a significant difference. After the resulting P values were ordered from smallest to largest, the 10 compounds with the most significant risk-group differences were defined as key candidate drugs. Their predicted IC50 distributions in HRG and LRG were displayed with ggplot2 (v3.5.2) [10]. To examine whether expression of the prognostic genes tracked with predicted drug response, Spearman correlation analysis was performed with psych (v2.5.3) [30]. Associations were considered relevant when $|cor| > 0.3$ and $P < 0.05$. The complete gene-drug correlation pattern was then displayed as a heatmap generated with corrplot (v0.92) [29].

### 2.10 Single-gene GSEA

In the training cohort, single-gene GSEA was carried out to investigate pathways linked to each prognostic gene. For each prognostic gene, Spearman correlations with all remaining genes were computed via psych (v2.5.3) [29], after which genes were ordered by decreasing correlation coefficient. The reference gene set used was the MSigDB collection "c2.cp.kegg_legacy.v2025.1.Hs.symbols.gmt". Ranked lists were analyzed with clusterProfiler (v4.8.3) [33], with Padj < 0.05 and |NES| > 1 defining significant enrichment [24]. For each prognostic gene, the five pathways with the largest positive enrichment and the five with the largest negative enrichment were selected according to NES and visualized using enrichplot (v1.20.0) [24] together with ggplot2 (v3.5.2) [10].

### 2.11 Drug prediction

The prognostic genes were submitted to the Drug Signatures Database (DSigDB) to identify enriched drug signatures. A network was then built from the drug-gene associations obtained in this analysis and visualized with ggplot2 (v3.5.2) [10].

### 2.12 Reverse transcription quantitative polymerase chain reaction (RT-qPCR)

RT-qPCR was applied to experimentally assess prognostic gene expression in clinical tissues. From CRC patients treated at The Third Affiliated Hospital of CQMU (FangDa Hospital), five matched pairs of CRC tumor and paracancerous tissue were obtained. All participants gave informed consent before tissue collection, and the Ethics Committee of the Third Affiliated Hospital of Chongqing Medical University (Fangda Hospital) approved the study protocol (2026伦审临研第（51）号). Tissue samples

were subjected to total RNA extraction with the TRIzol reagent kit supplied by Vazyme Biotech Co., Ltd., and all extraction steps followed the manufacturer's protocol. The extracted RNA was subsequently converted into cDNA using the HP All-in-one qRT Master Mix II RT203-Ver.1 kit manufactured by Kunming Yungen Biotechnology Co., Ltd., and reverse transcription followed the supplier's instructions. Quantitative PCR reactions used Servicebio® 2×Universal Blue SYBR Green qPCR Master Mix (Servicebio, Wuhan, China). Primer sequences for amplification are provided in Table 1. The $2^{-\Delta\Delta Ct}$ method, previously described and validated , was used to compute relative expression of each prognostic gene. After RT-qPCR data acquisition, the raw measurements were exported to Microsoft Excel for initial organization. GraphPad Prism (v10.1.2) was subsequently employed to statistically analyze and graphically visualize the organized data.

### 2.13 Statistical analysis

R (v4.3.3) was used for all statistical analyses. Between-group differences were evaluated by the Wilcoxon rank-sum test, with $P < 0.05$ denoting significance. In the RT-qPCR experiment, t-tests compared expression between CRC tumor tissues and adjacent non-tumor tissues at the same significance level ($P < 0.05$).

## 3. Results

### 3.1 The 39 candidate genes related to CRC and dephosphorylation

A total of 3,191 DEGs were identified by differential-expression analysis comparing TCGA-CRC tumor and normal samples (Supplementary Table 2). Of these genes, 1,525 showed higher expression and 1,666 showed lower expression in the tumor group

(Fig.1A). The 10 most strongly upregulated genes were MAGEA3, MAGEA6, KLK6, MAGEA12, NOTUM, FEZF1, PRSS56, ZIC5, LY6G6D, and CST1. By contrast, OTOP2, MS4A10, AQP8, CA1, TMIGD1, OTOP3, G6PC, KRT24, GUCA2B, and INSL5 represented the 10 most strongly downregulated genes (Fig.1B). When the complete DEG list was intersected with the DRG collection, 39 genes were shared by both sets and were therefore defined as candidate genes (Fig.1C; Supplementary Table 3).

**3.2 Candidate genes might be involved in the pathological progression of CRC by regulating calcium signaling-related pathways**

To determine which biological processes and signaling pathways the 39 candidate genes represented, functional enrichment was assessed. With $P < 0.05$ as the significance threshold, GO analysis returned 827 enriched terms: 663 BP, 69 CC, and 95 MF terms (Supplementary Table 4). KEGG analysis further yielded 96 significantly enriched pathways (Supplementary Table 5). Fig.2A summarizes the five most significant terms within each GO component together with the five most significant KEGG pathways. The leading BP results included calcineurin-mediated signal transduction and calcium ion-mediated signal transduction. Enriched CC terms included membrane rafts and β-catenin-TCF complexes, whereas MF enrichment included scaffold protein binding and ionotropic glutamate receptor activity. Adherens junctions and the calcium signaling pathway were also prominent among the KEGG results. Taken together, the enrichment profile connected the candidate genes with calcium-dependent signaling, cell-junction organization, and defined

molecular-binding functions, suggesting that these processes may accompany CRC development and progression. Protein-protein connectivity among the candidate proteins was then evaluated. PPI analysis, which included 20 of the 39 genes, yielded a network of 31 interactions (Fig.2B). This interconnected subset indicates that several candidate proteins might contribute to shared or coordinated biological processes and provides a basis for subsequent studies of their combined regulatory roles in CRC.

### 3.3 Identification of 3 prognostic genes associated with CRC prognosis

Using tumor samples with survival information from the training cohort, we next screened the 39 candidate genes for prognostic relevance. Univariate Cox regression initially identified seven genes that met the predefined criteria of $HR \neq 1$ and $P < 0.05$ (Fig.3A). Since Cox estimates depend on the proportional hazards assumption, these seven genes underwent the PH assumption test, with $P > 0.05$ set as the acceptance threshold. All genes passing the univariate Cox screen also met the PH criterion (Supplementary table 6), so the seven were retained as candidate prognostic genes. To further reduce model complexity, LASSO regression was then applied to this narrowed gene set. Ten-fold cross-validation determined lambda.min = 0.0601 as the optimal regularization parameter. At this value, three genes—TCF7L1, SNAI1, and TIMP1—remained in the model and were designated the final prognostic genes (Fig.3B-C).

### 3.4 A prognostic model based on 3 prognostic genes demonstrated prognostic discrimination in CRC

TCF7L1, SNAI1, and TIMP1 were combined to build a prognostic risk model, and its

formula yielded a risk score for each CRC tumor with survival data in both training and validation cohorts. Samples were allocated to risk groups using the cohort-specific optimal cutoff: 0.3544314 for the training set and 1.182279 for the validation set. Of the 524 eligible training tumors, 389 fell into HRG and 135 into LRG. In the 55-sample validation cohort, 26 cases fell into HRG and 29 into LRG. Examination of the risk distributions and survival status showed that deaths accounted for a larger proportion of HRG than LRG in both cohorts (Fig.4A-B). KM curves further demonstrated significant survival separation between the risk strata. Survival was poorer among HRG patients in the training cohort ($P < 0.0001$), and the validation cohort showed the same pattern ($P = 0.028$) (Fig.4C-D). Discrimination was additionally evaluated by time-dependent ROC analysis. The training cohort yielded 1-, 2-, and 3-year AUCs of 0.604, 0.626, and 0.658, respectively (Fig.4E), while the validation cohort produced corresponding AUCs of 0.687, 0.764, and 0.652 (Fig.4F). Across the two datasets, these values indicated that the three-gene model achieved modest-to-moderate prognostic discrimination at the evaluated time points.

### 3.5 Cell-junction-related pathways were enriched in the HRG and associated with poor prognosis in CRC

Pathway activity in training-set tumors with survival data was compared between HRG and LRG by GSEA. In total, 39 pathways differed significantly in enrichment between the two groups (Supplementary Table 7). Ranking significant pathways by NES revealed the top positively enriched programs in HRG, namely extracellular matrix (ECM)-receptor interaction, cell adhesion molecules (CAMs), and hematopoietic cell

lineage (Fig.5A). Conversely, ascorbate and aldarate metabolism, retinol metabolism, and butanoate metabolism were among the pathways with the strongest negative enrichment in HRG (Fig.5B). Thus, the high-risk phenotype was characterized by this enrichment pattern, with stronger activity of matrix- and cell-junction-related programs and weaker enrichment of several metabolic pathways. Since HRG also showed poorer survival, these pathway-level differences offer testable biological clues for subsequent exploration of mechanisms underlying high-risk CRC.

### 3.6 Differential infiltration of 21 immune cell types correlated with risk stratification in CRC

In the training cohort, the two risk strata showed broad differences in immune infiltration patterns. Fig.6A presents a heatmap of ssGSEA scores for all 28 immune-cell populations in HRG and LRG. Wilcoxon rank-sum testing revealed 21 cell types with significant between-group differences (DIICs; $P < 0.05$). Inferred infiltration scores were higher in HRG for most DIICs, while Type 17 T helper cells were higher in LRG ($P < 0.05$) and Type 2 T helper cells were higher in LRG ($P < 0.001$) (Fig.6B). The model-defined risk groups were therefore linked to distinct immune-composition profiles. Nevertheless, ssGSEA estimates relative enrichment, not cellular function; accordingly, these differences should not be treated as direct evidence of antitumor or protumor activity. They instead provide a basis for examining how CRC risk status relates to remodeling of the immune microenvironment and to potential immunotherapeutic features. Among the DIICs, the strongest positive relationship occurred between macrophages and myeloid-derived suppressor cells

(MDSCs), as further revealed by correlation analysis (cor = 0.94, P < 0.001) (Fig.6C). This value reflects tight co-variation of inferred infiltration scores and does not demonstrate direct functional interaction or coordinated recruitment. At the gene–cell level, Natural killer T (NKT) cells had the strongest positive correlation with TIMP1 (cor = 0.55, P < 0.001) (Fig.6D; Supplementary Table 8). Likewise, this finding reflects an expression–infiltration association, not a regulatory relationship.

### 3.7 TIDE analysis predicted greater potential immunotherapy response in the LRG

Predicted immunotherapy-related characteristics of HRG and LRG were compared using TIDE analysis. Within the training cohort, HRG exhibited higher dysfunction, exclusion, and overall TIDE scores than LRG (P < 0.0001 for each comparison), whereas LRG showed a higher MSI score (P < 0.0001) (Fig.7A). Because lower TIDE-associated dysfunction and exclusion together with higher MSI are compatible with a more favorable predicted immune-response profile, these score differences suggested greater potential for immunotherapy response in LRG. TIDE's categorical response output was consistent with this pattern: LRG contained a higher proportion of samples predicted as potential immunotherapy responders than HRG (Fig.7B). Thus, these findings reflect a model-based difference in predicted response potential, not directly observed treatment outcomes.

### 3.8 The 150 chemotherapeutic drugs with significant differences between the HRG and LRG

HRG and LRG were compared in the training cohort to further investigate

drug-response prediction. From the GDSC resource, 198 compounds were included in the sensitivity analysis (Supplementary Table 9). Based on predicted IC50 values, 150 compounds showed significant differences between the two risk groups at $P < 0.05$ (Supplementary Table 10). Fig.8A presents the predicted IC50 distributions of the key compounds showing the strongest group differences. For these drugs, HRG showed significantly higher IC50 values ($P < 0.0001$), which pointed to lower predicted sensitivity in high-risk tumors and greater predicted sensitivity in LRG. Relationships between prognostic-gene expression and the key compounds were also assessed. In this analysis, the strongest significant positive correlation was observed between TCF7L1 and Erlotinib_1168 ($cor = 0.57$, $P < 0.001$) (Fig.8B; Supplementary Table 11). This correlation suggests that TCF7L1 expression may be informative for predicted Erlotinib_1168 sensitivity in CRC, although the computational association does not establish an actual clinical response to the drug.

### 3.9 Prognostic genes were suggested to jointly affect CRC progression by coordinately regulating core biological processes such as extracellular matrix remodeling

Single-gene GSEA was used to characterize pathway associations separately for SNAI1, TCF7L1, and TIMP1. SNAI1 was significantly associated with 71 KEGG pathways (Supplementary Table 12). Positive enrichment included ECM-receptor interaction, cytokine-cytokine receptor interaction, and cell adhesion molecules (CAMs), while significant negative enrichment was observed for fatty acid metabolism, butanoate metabolism, and propanoate metabolism (Fig.9A-B). TCF7L1 showed significant

enrichment in 85 KEGG pathways (Supplementary Table 13). For this gene, ECM-receptor interaction, CAMs, and neuroactive ligand-receptor interaction were among the positively enriched pathways, while DNA replication, nucleotide excision repair, and aminoacyl-tRNA biosynthesis were among the negatively enriched pathways (Fig.9C-D). TIMP1 was associated with 69 significantly enriched KEGG pathways (Supplementary Table 14). For this gene, positive enrichment covered CAMs and cytokine-cytokine receptor interaction, while negative enrichment was observed for aminoacyl-tRNA biosynthesis, cell cycle, and propanoate metabolism (Fig.9E-F). Comparison across the three analyses identified 39 KEGG pathways shared by SNAI1, TCF7L1, and TIMP1. Common pathways comprised ECM-receptor interaction, Fc gamma R-mediated phagocytosis, cytokine-cytokine receptor interaction, chemokine signaling pathway, and MAPK signaling pathway (Supplementary Table 15). Beyond their 39 shared pathways, all three genes showed similar positive associations with ECM-receptor interaction and cell-adhesion-related programs. Taken together, these convergent enrichment patterns suggest that SNAI1, TCF7L1, and TIMP1 may contribute to CRC progression through coordinated participation in extracellular-matrix remodeling and intercellular signaling processes.

**3.10 Drug prediction**

Drug-signature enrichment through DSigDB identified 125 candidate drug signatures associated with TIMP1, 13 associated with TCF7L1, and 59 associated with SNAI1. Comparison of these sets showed that 14 candidate drugs were shared by SNAI1 and TIMP1, three were shared by TIMP1 and TCF7L1, and one candidate drug was

common to TCF7L1 and SNAI1 (Fig.10).

### 3.11 RT-qPCR

According to RT-qPCR, control tissues had higher TCF7L1 expression ($P<0.001$), while CRC tissues showed higher SNAI1 expression ($P<0.001$). TIMP1 was also somewhat higher in CRC tissues, although this difference was not statistically significant ($P > 0.05$) (Fig.11).

## 4. Discussion

Marked biological heterogeneity in CRC means that outcome reflects several interacting features, including metastatic competence, stemness-associated programs, and responsiveness to treatment [2]. Three DRGs with prognostic relevance—SNAI1, TCF7L1, and TIMP1—emerged from the present analysis and were combined into a model that stratified patients into risk groups showing modest prognostic discrimination. When considered alongside the enrichment, immune, and drug-sensitivity results, the signature appears to represent multiple clinically relevant dimensions of CRC biology rather than one isolated signaling route [4]. Importantly, these observations indicate an association between dephosphorylation-related networks and malignant CRC behavior; they do not establish a single, fully characterized phosphatase pathway [5]. Because the study is based primarily on transcriptomic associations and computational inference, the three genes should presently be viewed as plausible biomarkers and candidate regulators rather than experimentally confirmed mechanistic drivers.

Of the three model genes, SNAI1 has the clearest established connection with epithelial-mesenchymal transition (EMT), invasive behavior, stemness, and resistance to therapy. According to recent studies, epithelial SNAI1 promotes CRC progression via EMT and Wnt/beta-catenin signaling, and SNAI1-linked programs may foster chemoresistance and cancer-stem-cell phenotypes[34, 35]. This biological context is compatible with the association observed here between elevated SNAI1 and unfavorable prognosis. Nevertheless, our data do not demonstrate that dephosphorylation is the principal upstream determinant of SNAI1 activity in CRC. A more defensible explanation is that dephosphorylation-related signaling could influence SNAI1 abundance, localization, or transcriptional activity and thereby reinforce EMT- and stemness-associated states. SNAI1 could thus serve as a downstream convergence node where DRG-associated signaling acts on invasion, metastasis, and treatment resistance[34, 35]; however, whether specific phosphatases regulate SNAI1 still requires experimental testing. Taken together with prior evidence, SNAI1 may function as an important link among EMT, maintenance of stem-like features, and 5-FU resistance.

TCF7L1 offers a second potential connection between the DRG signature and CRC progression. As a Wnt-associated transcription factor, it participates in epithelial differentiation, stem-cell behavior, and migratory capacity. The migration-promoting activity of TCF7L1 may therefore provide one biological route linking the model with metastatic phenotypes [36]. A recent report indicated that TCF7L1 facilitates CRC-cell migration via GAS1 repression [36], consistent with our finding that higher TCF7L1

expression accompanied poorer prognosis. Even so, the present data do not support labeling TCF7L1 as a confirmed dephosphorylation-dependent driver. A more cautious possibility is that dephosphorylation-associated networks alter the abundance, stability, localization, or activity of TCF7L1 and thereby influence Wnt-related transcription. This interpretation preserves the observed biological connection without implying a direct regulatory mechanism. Future functional studies will be needed to test whether defined phosphatases modify TCF7L1 localization, DNA-binding behavior, or its interaction with beta-catenin in CRC cells.

Through TIMP1, the biological scope of the signature is broadened beyond transcriptional control to encompass extracellular-matrix remodeling and cellular stress adaptation. In CRC, TIMP1 has been associated with poor survival, altered matrix organization, pro-survival signaling, and treatment resistance. Recent evidence also indicates that enhancer-driven TIMP1 expression can increase ferroptosis resistance through the SPI1/TIMP1 axis [37], thereby connecting TIMP1 with both invasive behavior and therapeutic vulnerability. Our observations are compatible with TIMP1 marking a more aggressive disease state, but they do not demonstrate direct regulation of TIMP1 by dephosphorylation. At this stage, it is more appropriate to propose that dephosphorylation-related pathways interact with the upstream signaling environment that maintains TIMP1 expression or function, rather than to describe TIMP1 as a validated dephosphorylation-dependent oncogene. Beyond extracellular remodeling, TIMP1 may also contribute to the high-risk phenotype through its relationship with ferroptosis sensitivity and treatment vulnerability [37].

The three-gene model should also be considered from a systems-level perspective. Its prognostic behavior overlaps with patterns reported for other CRC gene signatures, suggesting that different molecular starting points can converge on shared aspects of tumor aggressiveness. This does not mean that a dephosphorylation-related signature replaces existing prognostic models. Instead, it may contribute an additional mechanistic viewpoint by linking transcriptomic risk to the balance of post-translational regulatory processes. Such a viewpoint is particularly relevant if dephosphorylation participates in stemness maintenance, signaling plasticity, and adaptation to therapy [4, 5]. At the same time, similar prognostic performance across different signatures should not be interpreted as evidence that the signatures are biologically interchangeable. Distinct models may emphasize different layers of the same disease process, and the present model specifically highlights a layer centered on post-translational regulatory balance and its association with malignant CRC behavior.

The immune findings also require interpretation beyond simple cell-abundance estimates. Higher inferred infiltration of macrophages, mast cells, and natural killer T cells in HRG should not be equated with a stronger antitumor response. In CRC, immune-rich tumors may still exhibit dysfunctional or suppressive microenvironments depending on the activation states and spatial arrangement of the infiltrating cells [2]. Accordingly, the observed group differences are better regarded as evidence that immune composition changes with risk status, rather than as proof of enhanced protective immunity in HRG. Antitumor activity is likely to depend more on

immune-cell function, localization, and metabolic fitness than on estimated abundance alone [38, 39].

The apparent immunotherapy advantage of LRG should be interpreted with the same restraint. The TIDE results represent computational predictions and therefore do not constitute direct evidence that low-risk patients will respond better to immunotherapy in clinical practice. TIDE can be valuable for generating hypotheses about dysfunction and immune exclusion, but it cannot replace treatment-response observations from prospective cohorts. One plausible interpretation is that LRG retains a microenvironment with less pronounced suppressive features and is therefore more permissive for immunotherapy, whereas HRG may contain stronger programs associated with immune resistance [4]; [40]. This framing maintains the potential translational relevance of the result without extending the conclusion beyond the available data. This analysis is therefore best interpreted as a hypothesis-generating foundation for subsequent immunotherapy stratification studies, rather than as proof of actual clinical benefit [41, 42].

The model's translational value lies in providing a compact framework that links prognosis with candidate biological and therapeutic features, rather than in explaining CRC pathogenesis comprehensively. Integrating gene-expression risk, immune context, and predicted treatment sensitivity can prioritize experiments on phosphatase-related pathways and future multi-marker strategies. If the proposed phosphatase-substrate relationships are confirmed, such a signature could support prognostic assessment and selection of patients or models for mechanism-focused validation.

Several limitations warrant consideration. Because the analysis depends mainly on public transcriptomic datasets and computational inference, experimental evidence has not yet established causal links between DRGs and CRC progression. In addition, direct phosphatase-dependent mechanisms involving SNAI1, TCF7L1, or TIMP1 were not examined in vitro or in vivo. The predicted differences in immune behavior and drug response also lack confirmation in independent treatment cohorts. Thus, validation is still needed both at the phosphatase-substrate level and in external therapeutic datasets. Future studies should test the signature in independent clinical cohorts, define the relevant phosphatase-substrate pairs, and determine how dephosphorylation-associated signaling contributes to EMT, stemness, and ferroptosis resistance in CRC.

**Abbreviations**

| **Abbreviation** | **Full term** |
|---|---|
| CRC | colorectal cancer |
| DRGs | dephosphorylation-related genes |
| DEGs | differentially expressed genes |
| GO | Gene Ontology |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| PPI | protein-protein interaction |
| HRG | high-risk group |

| | |
|---|---|
| LRG | low-risk group |
| GSEA | gene set enrichment analysis |
| ssGSEA | single-sample Gene Set Enrichment Analysis |
| DIICs | differentially infiltrated immune cells |
| TIDE | Tumor Immune Dysfunction and Exclusion |
| MSI | microsatellite instability |
| GDSC | Genomics of Drug Sensitivity in Cancer |
| RT-qPCR | reverse transcription quantitative polymerase chain reaction |
| KM | Kaplan-Meier |
| ROC | receiver operating characteristic |
| AUC | area under the curve |
| EMT | epithelial-mesenchymal transition |

**References**


1. Huang, X., et al., *Prognostic value of fatty acid metabolism-related genes in colorectal cancer and their potential implications for immunotherapy.* Front Immunol, 2023. **14**: p. 1301452.
2. Cañellas-Socias, A., E. Sancho, and E. Batlle, *Mechanisms of metastatic colorectal cancer.* Nat Rev Gastroenterol Hepatol, 2024. **21**(9): p. 609–625.
3. Yang, F., et al., *PLIN2 promotes colorectal cancer progression through CD36-mediated epithelial-mesenchymal transition.* Cell Death Dis, 2025. **16**(1): p. 510.
4. Zhao, M., et al., *Protein tyrosine phosphatases: emerging role in cancer therapy resistance.* Cancer Commun (Lond), 2024. **44**(6): p. 637–653.
5. Fujiwara, N., et al., *Protein phosphatase 6 promotes stemness of colorectal cancer cells.* Cancer Sci, 2024. **115**(9): p. 3067–3078.
6. Zhang, Q., et al., *Reprogramming of palmitic acid induced by dephosphorylation of ACOX1 promotes β-catenin palmitoylation to drive colorectal cancer progression.* Cell Discov, 2023. **9**(1): p. 26.

7. Chen, Z., et al., *CircGPRC5A enhances colorectal cancer progress by stabilizing PPP1CA and inducing YAP dephosphorylation.* J Exp Clin Cancer Res, 2023. **42**(1): p. 334.
8. Liu, S., et al., *Three Differential Expression Analysis Methods for RNA Sequencing: limma, EdgeR, DESeq2.* J Vis Exp, 2021(175).
9. Ahmed, S., et al., *Identification of the Hub Genes Involved in Chikungunya Viral Infection.* Cureus, 2024. **16**(4): p. e57603.
10. Yuan, S.M., et al., *C6 and KLRG2 are pyroptosis subtype-related prognostic biomarkers and correlated with tumor-infiltrating lymphocytes in lung adenocarcinoma.* Sci Rep, 2024. **14**(1): p. 24861.
11. Zhou, W., et al., *Identification and mechanism analysis of biomarkers related to butyrate metabolism in COVID-19 patients.* Ann Med, 2025. **57**(1): p. 2477301.
12. Song, H., et al., *Combining mendelian randomization analysis and network toxicology strategy to identify causality and underlying mechanisms of environmental pollutants with glioblastoma: A study of Methyl-4-hydroxybenzoate.* Ecotoxicol Environ Saf, 2024. **287**: p. 117311.
13. Zhang, Y., et al., *Identification of atrial fibrillation-related genes through transcriptome data analysis and Mendelian randomization.* Front Cardiovasc Med, 2024. **11**: p. 1414974.
14. Lei, J., et al., *Clinicopathological characteristics of pheochromocytoma/paraganglioma and screening of prognostic markers.* J Surg Oncol, 2023. **128**(4): p. 510–518.
15. Li, Z., et al., *Differentiation-related genes in tumor-associated macrophages as potential prognostic biomarkers in non-small cell lung cancer.* Front Immunol, 2023. **14**: p. 1123840.
16. Maffeo, D., et al., *The Evidence Base for Circulating Tumor DNA-Methylation in Non-Small Cell Lung Cancer: A Systematic Review and Meta-Analysis.* Cancers (Basel), 2024. **16**(21).
17. Yan, T., et al., *Survival prediction for patients with glioblastoma multiforme using a Cox proportional hazards denoising autoencoder network.* Front Comput Neurosci, 2022. **16**: p. 916511.
18. Shi, H., et al., *Identifying and Validating GSTM5 as an Immunogenic Gene in Diabetic Foot Ulcer Using Bioinformatics and Machine Learning.* J Inflamm Res, 2023. **16**: p. 6241–6256.
19. Shi, Y., et al., *Crosstalk of ferroptosis regulators and tumor immunity in pancreatic adenocarcinoma: novel perspective to mRNA vaccines and personalized immunotherapy.* Apoptosis, 2023. **28**(9-10): p. 1423–1435.
20. Huang, A., et al., *Novel hypoxia- and lactate metabolism-related molecular subtyping and prognostic signature for colorectal cancer.* J Transl Med, 2024. **22**(1): p. 587.
21. Yan, J., et al., *Exploring Programmed Cell Death-Related Biomarkers and Disease Therapy Strategy in Nasopharyngeal Carcinoma Using Transcriptomics.* Front Biosci (Landmark Ed), 2024. **29**(7): p. 240.

22. Yu, S., et al., *A senescence-based prognostic gene signature for colorectal cancer and identification of the role of SPP1-positive macrophages in tumor senescence.* Front Immunol, 2023. **14**: p. 1175490.
23. Chen, J., et al., *Construction of novel hypoxia-related gene model for prognosis and tumor microenvironment in endometrial carcinoma.* Front Endocrinol (Lausanne), 2022. **13**: p. 1075431.
24. Liu, L., et al., *Machine learning-based solution reveals cuproptosis features in inflammatory bowel disease.* Front Immunol, 2023. **14**: p. 1136991.
25. Wang, L., et al., *Cuproptosis related genes associated with Jab1 shapes tumor microenvironment and pharmacological profile in nasopharyngeal carcinoma.* Front Immunol, 2022. **13**: p. 989286.
26. Chen, J., et al., *Identification of hub genes and their correlation with infiltration of immune cells in MYCN positive neuroblastoma based on WGCNA and LASSO algorithm.* Front Immunol, 2022. **13**: p. 1016683.
27. Zhou, J., et al., *Identification of aging-related biomarkers and immune infiltration characteristics in osteoarthritis based on bioinformatics analysis and machine learning.* Front Immunol, 2023. **14**: p. 1168780.
28. Liang, L., et al., *Integration of scRNA-Seq and Bulk RNA-Seq to Analyse the Heterogeneity of Ovarian Cancer Immune Cells and Establish a Molecular Risk Model.* Front Oncol, 2021. **11**: p. 711020.
29. Zhang, X., et al., *Single-cell RNA and transcriptome sequencing profiles identify immune-associated key genes in the development of diabetic kidney disease.* Front Immunol, 2023. **14**: p. 1030198.
30. Zhang, L., et al., *Dysregulated Circulating Apoptosis- and Autophagy-Related lncRNAs as Diagnostic Markers in Coronary Artery Disease.* Biomed Res Int, 2021. **2021**: p. 5517786.
31. Xie, Q., et al., *Identification and verification of immune-related genes for diagnosing the progression of atherosclerosis and metabolic syndrome.* BMC Cardiovasc Disord, 2024. **24**(1): p. 405.
32. Maeser, D., R.F. Gruener, and R.S. Huang, *oncoPredict: an R package for predicting in vivo or cancer patient drug response and biomarkers from cell line screening data.* Brief Bioinform, 2021. **22**(6).
33. Liu, X., et al., *Potential mechanisms of traditional Chinese medicine in treating insomnia: A network pharmacology, GEO validation, and molecular-docking study.* Medicine (Baltimore), 2024. **103**(18): p. e38052.
34. Qing, F., et al., *Intestinal epithelial SNAI1 promotes the occurrence of colorectal cancer by enhancing EMT and Wnt/β-catenin signaling.* Med Oncol, 2023. **41**(1): p. 34.
35. Janakiraman, H., et al., *Targeting SNAI1-Mediated Colorectal Cancer Chemoresistance and Stemness by Sphingosine Kinase 2 Inhibition.* World J Oncol, 2024. **15**(5): p. 744–757.
36. King, C.M., et al., *TCF7L1 regulates colorectal cancer cell migration by repressing GAS1 expression.* Sci Rep, 2024. **14**(1): p. 12477.

37. Li, M., Q.Y. Ni, and S.Y. Yu, *Integration of single-cell transcriptomics and epigenetic analysis reveals enhancer-controlled TIMP1 as a regulator of ferroptosis in colorectal cancer.* Genes Genomics, 2024. **46**(1): p. 121–133.
38. N, M.I., et al., *Colorectal Cancer and Its Microenvironment: A Brief Review.* Med J Islam Repub Iran, 2024. **38**: p. 108.
39. Nicolini, A. and P. Ferrari, *Involvement of tumor immune microenvironment metabolic reprogramming in colorectal cancer progression, immune escape, and response to immunotherapy.* Front Immunol, 2024. **15**: p. 1353787.
40. Salmond, R.J., *Targeting Protein Tyrosine Phosphatases to Improve Cancer Immunotherapies.* Cells, 2024. **13**(3).
41. González-Montero, J., C.I. Rojas, and M. Burotto, *Predictors of response to immunotherapy in colorectal cancer.* Oncologist, 2024. **29**(10): p. 824–832.
42. Kaviyarasan, V., et al., *Advancements in immunotherapy for colorectal cancer treatment: a comprehensive review of strategies, challenges, and future prospective.* Int J Colorectal Dis, 2024. **40**(1): p. 1.

**Declarations**

**Funding**

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

**Conflict of Interest**

The authors have no relevant financial or non-financial interests to disclose.

**Author contributions**

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Qiao Deng. The first draft of the manuscript was written by Qiao Deng and

all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

**Acknowledgments**

We would like to express our sincere gratitude to all individuals and organizations who supported and assisted us throughout this research. Special thanks to the following authors: Lv Hu. In conclusion, we extend our thanks to everyone who has supported and assisted us along the way. Without your support, this research would not have been possible.

**Data Availability**

All datasets used in this study are publicly available. CRC expression data were obtained from TCGA GDC portal: TCGA-COAD (https://portal.gdc.cancer.gov/projects/TCGA-COAD) and TCGA-READ (https://portal.gdc.cancer.gov/projects/TCGA-READ). The validation cohort GSE17537 (GPL570) is available at GEO: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17537. The 261 DRGs were retrieved from MsigDB (https://www.gsea-msigdb.org/gsea/msigdb). Other resources: STRING (https://string-db.org/), GDSC (https://www.cancerrxgene.org/), TIDE (http://tide.dfci.harvard.edu/), DSigDB (https://dsigdb.tanlab.org/).

**Ethics approval**

This study was approved by the Ethics Committee of the Third Affiliated Hospital of Chongqing Medical

University (Fangda Hospital) (2026 Ethics Review (Clinical Research) No. (51)). It was conducted in accordance with the Declaration of Helsinki and the 2023 Measures for Ethical Review of Life Science and Medical Research Involving Human Subjects issued by the National Health Commission of China.

**Consent to participate**

Written informed consent was obtained from all individual participants included in the study prior to the collection and use of their tumor specimens

**Consent to publish**

Not applicable.

**Figures**

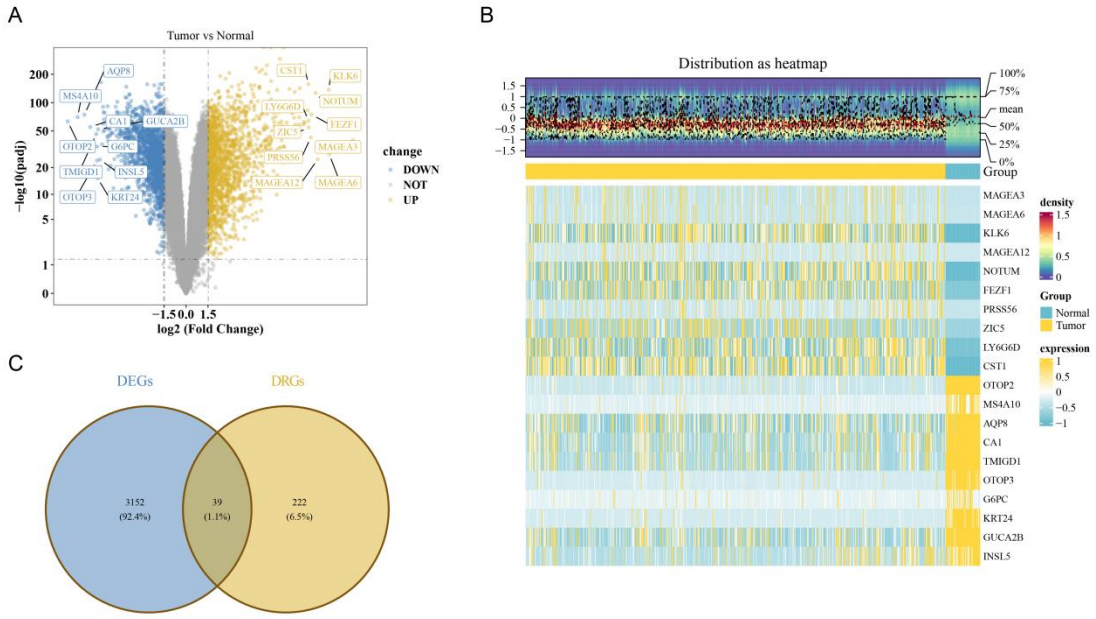


**Fig.1 Identification of differentially expressed genes (DEGs) and candidate genes.** **A** The volcano diagram showed the DEGs between tumor and normal groups. **B** The heatmap diagram for the top 10 upregulated and top 10 downregulated genes in the tumor group. **C** The Venn plot displayed the overlap of DEGs and dephosphorylation-related genes (DRGs).

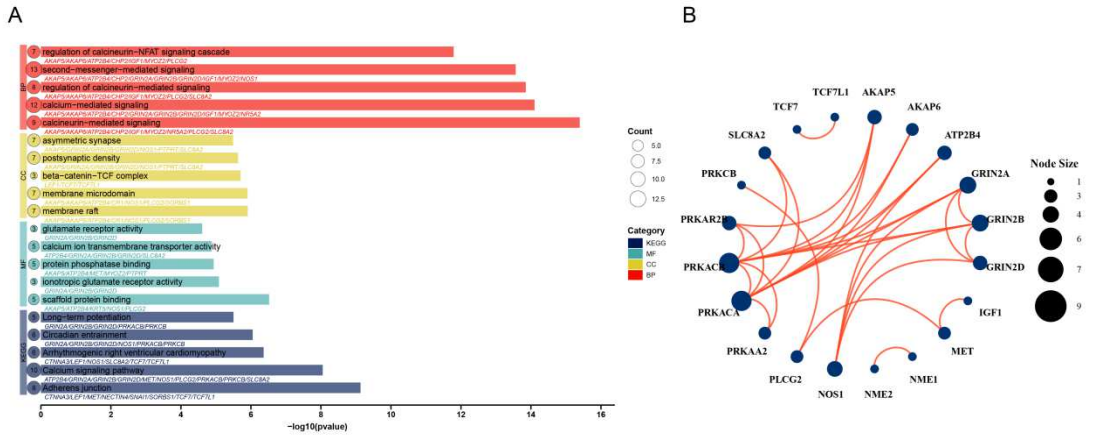


**Fig.2 Enrichment analysis and protein-protein interaction (PPI) network of candidate genes.** **A** The top 5 significantly enriched Gene Ontology (GO) terms in biological process (BP), cellular component (CC), and molecular function (MF) as well as the top 5 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways for candidate genes. **B** PPI network.

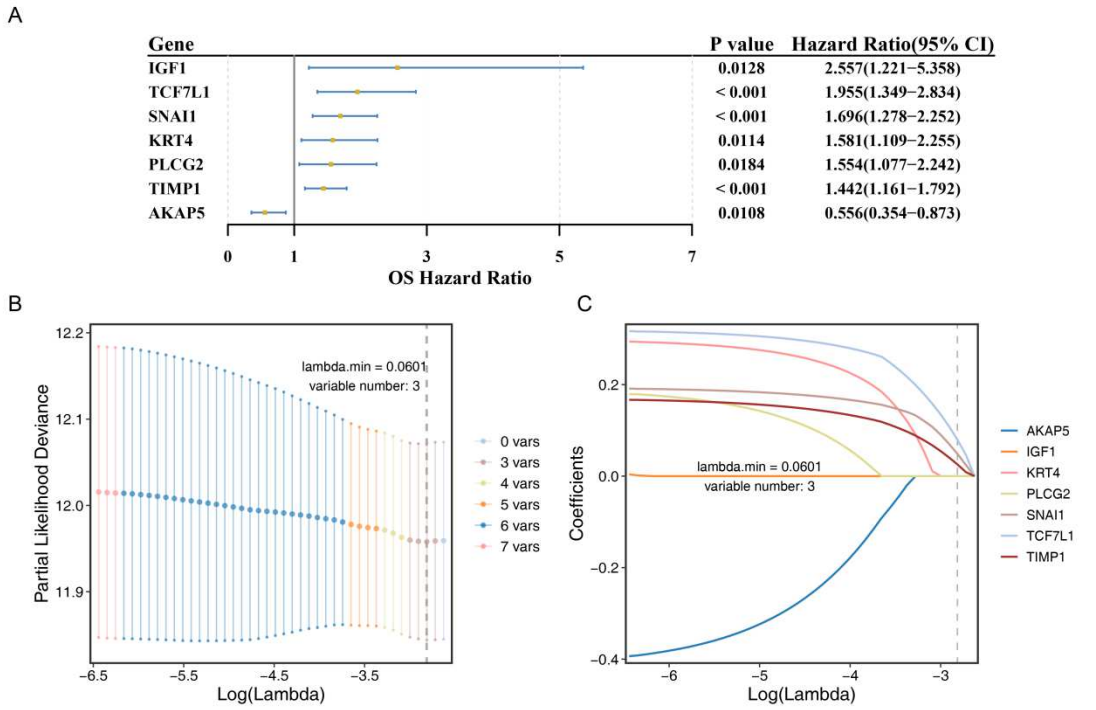


**Fig.3** Identification of prognostic genes. **A** The forest plot of univariate Cox Regression analysis to screen genes meeting HR≠1 and P<0.05 from candidate genes. **B** The λ selection diagram for 10-fold cross-validation of LASSO regression model. **C** The LASSO coefficient profiles of 7 candidate prognostic genes in 10-fold cross-validation.

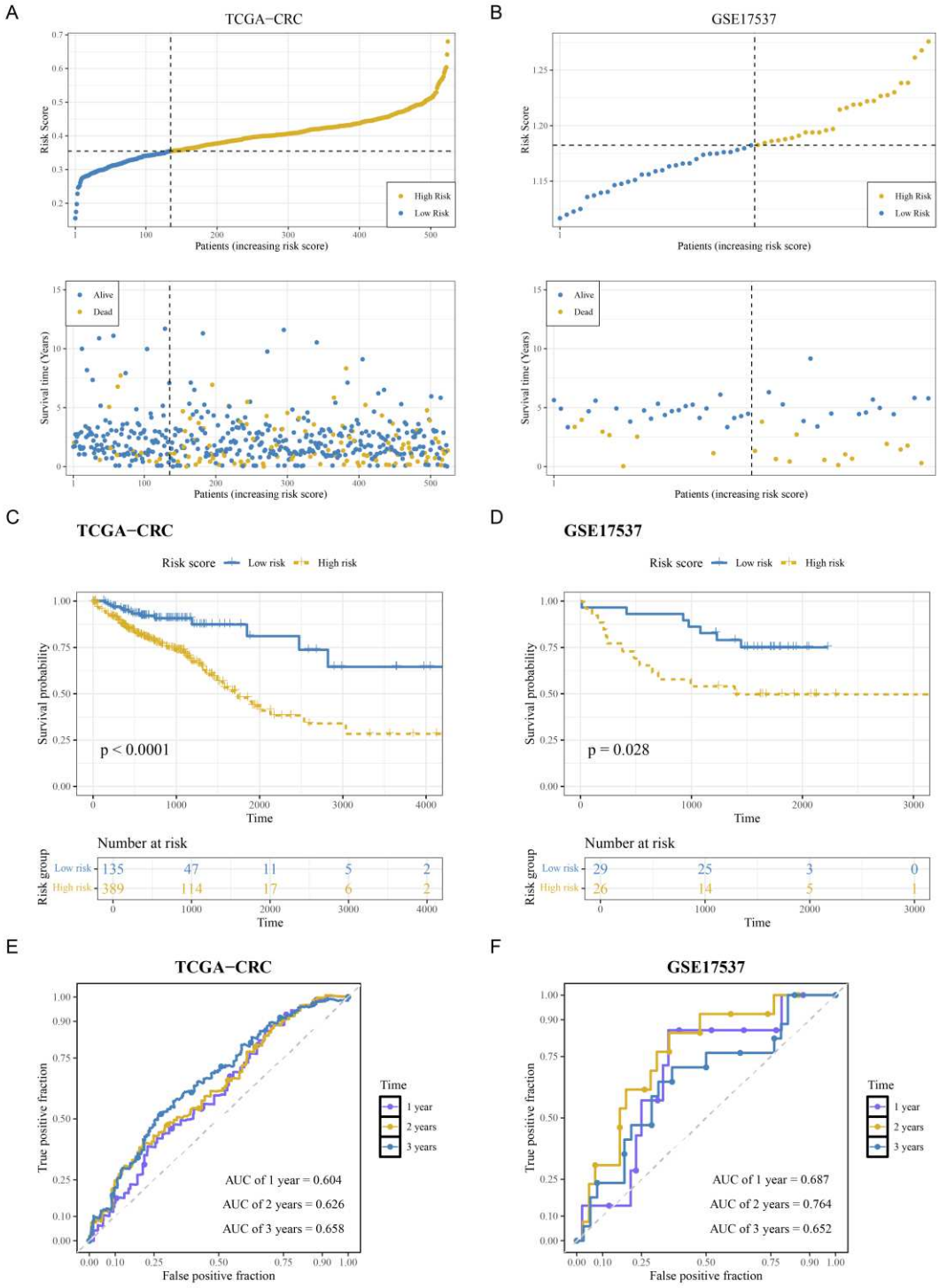


**Fig.4** Evaluation and validation of the prognostic model. Distribution of risk scores and survival status of CRC patients in both **A** training set (n = 524) and **B** validation set (n = 55). The ranked dot plot showed patients stratified into two groups according to the optimal cut-off value of risk scores (top). The scatter plot graphed riskscores corresponding to survival time and status (bottom).The Kaplan–Meier (KM) survival curves for overall survival (OS) in high-risk group (HRG) and low-risk group (LRG)

of the **C** training set and **D** validation set. The time-dependent receiver operating characteristic (ROC) analysis for 1-year, 2-year, and 3-year OS prediction in **E** training set and **F** validation set.

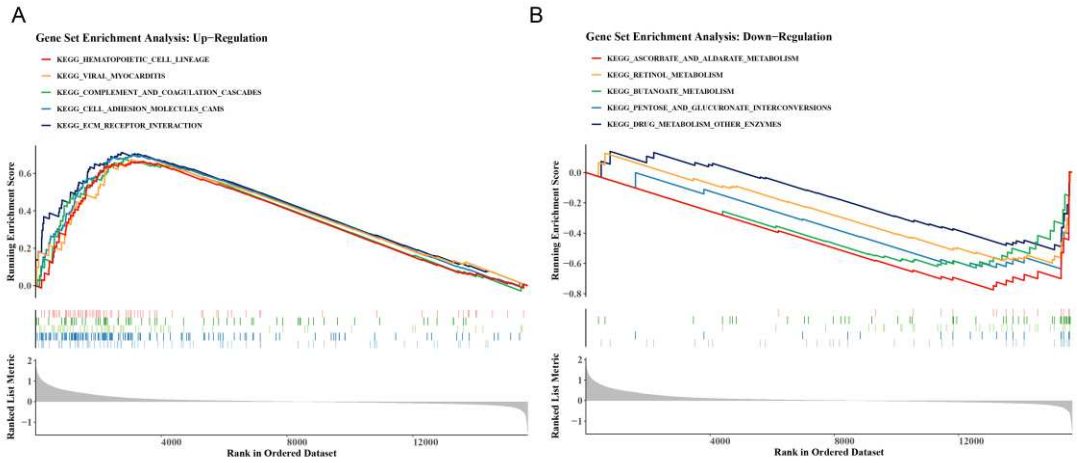


**Fig.5 Gene set enrichment analysis (GSEA) was used to analyze the signaling pathways enrichment in the HRG and LRG. A** The top 5 significantly upregulated KEGG pathways in the HRG. **B** The top 5 significantly downregulated KEGG pathways in the HRG.

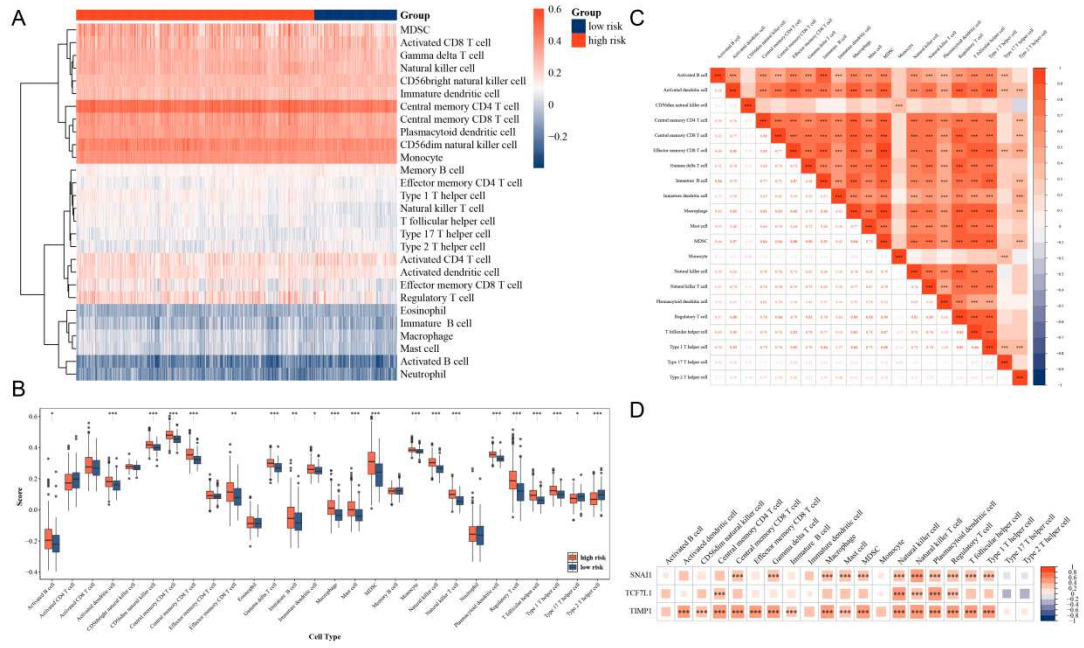


**Fig.6 Immune infiltration analysis. A** The heatmap generated based on the single sample Gene Set Enrichment Analysis (ssGSEA) algorithm showed the infiltration scores of 28 types of immune cells in the HRG and LRG of CRC. **B** The box plot of

differences in the infiltration scores of 28 immune cells in HRG and LRG. Correlation analysis between **C** differentially infiltrating immune cells (DIICs), and **D** between prognostic genes and DIICs. *P < 0.05, **P < 0.01, ***P < 0.001.

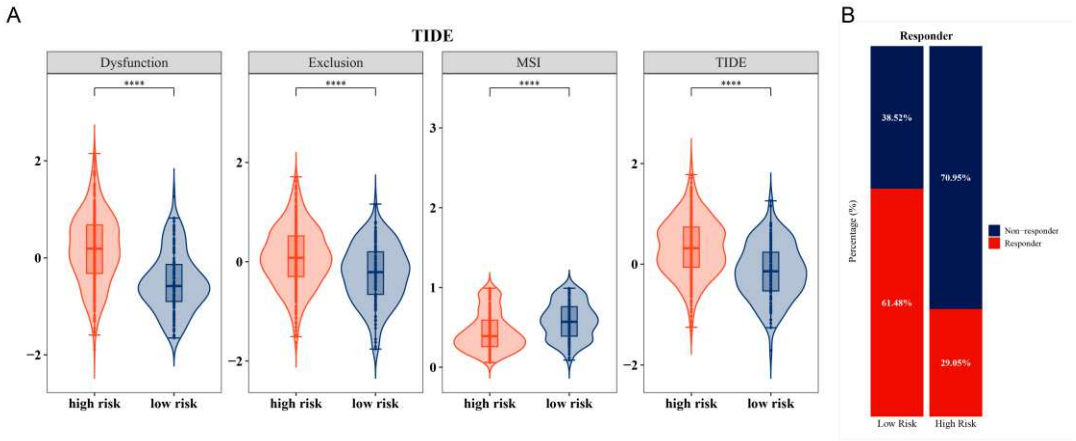


**Fig.7 Tumor immune dysfunction and exclusion (TIDE).** **A** The dysfunction scores, exclusion scores, TIDE scores, and MSI scores of samples in the HRG and LRG. P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001. **B** Differences in immunotherapy response between HRG and LRG patients.

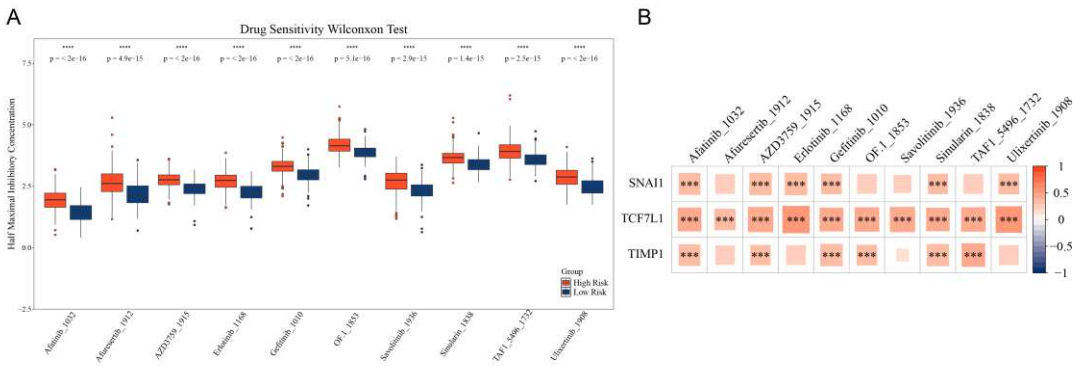


**Fig.8 Chemotherapeutic drug sensitivity analysis.** **A** The distribution of half maximal inhibitory concentration (IC50) values for the key chemotherapeutic drugs between the HRG and LRG. *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001. **B** Correlation analysis between key chemotherapeutic drugs and prognostic genes.

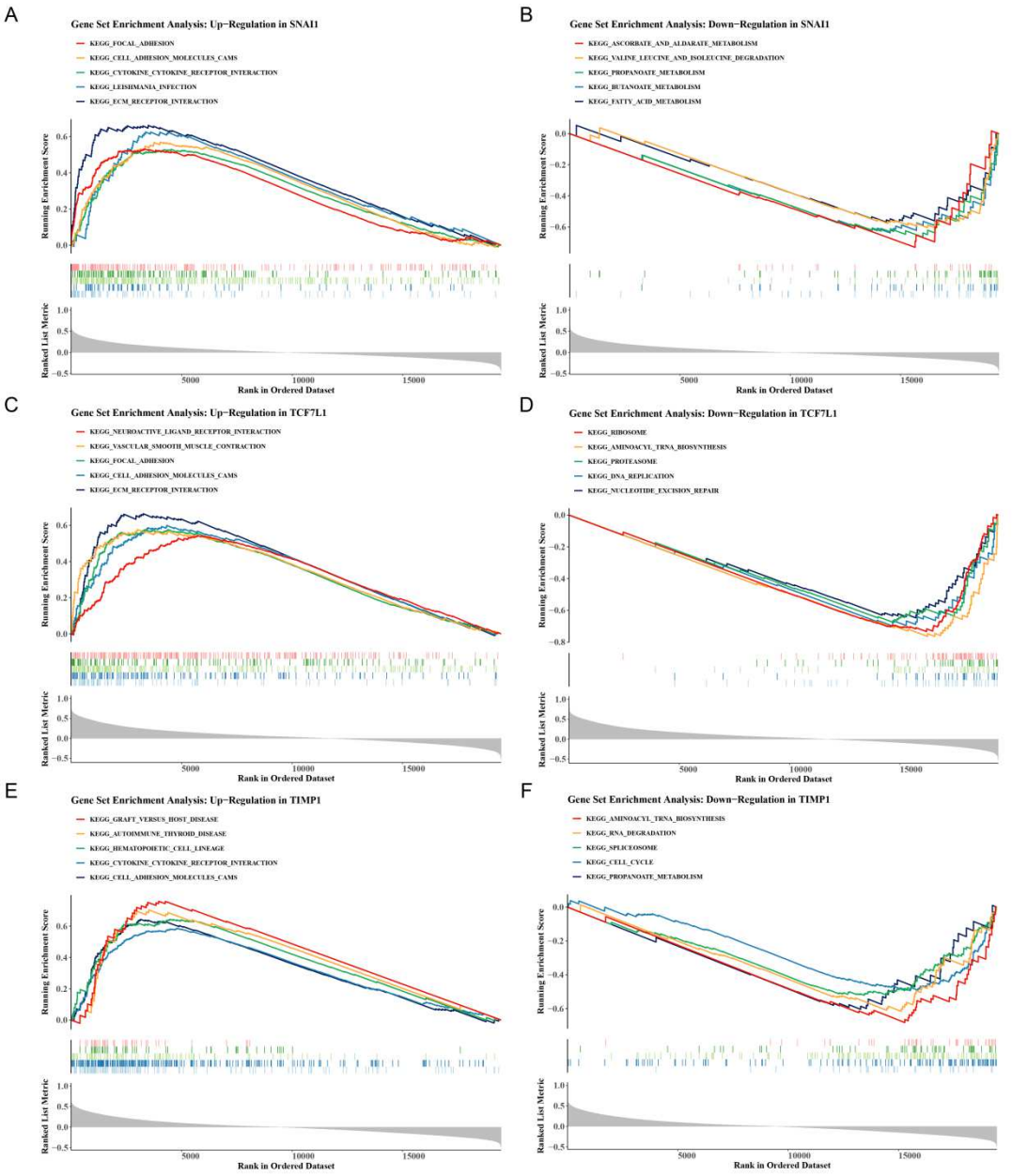


**Fig.9 Single-gene GSEA.** **A** Top 5 KEGG pathways in which SNAI1 was significantly upregulated. **B** Top 5 KEGG pathways in which SNAI1 was significantly downregulated. **C** Top 5 KEGG pathways in which TCF7L1 was significantly upregulated. **D** Top 5 KEGG pathways in which TCF7L1 was significantly downregulated. **E** Top 5 KEGG pathways in which TIMP1 was significantly upregulated. **F** Top 5 KEGG pathways in which TIMP1 was significantly downregulated.

**Fig.10 The interaction network between prognostic genes and potential drugs**.

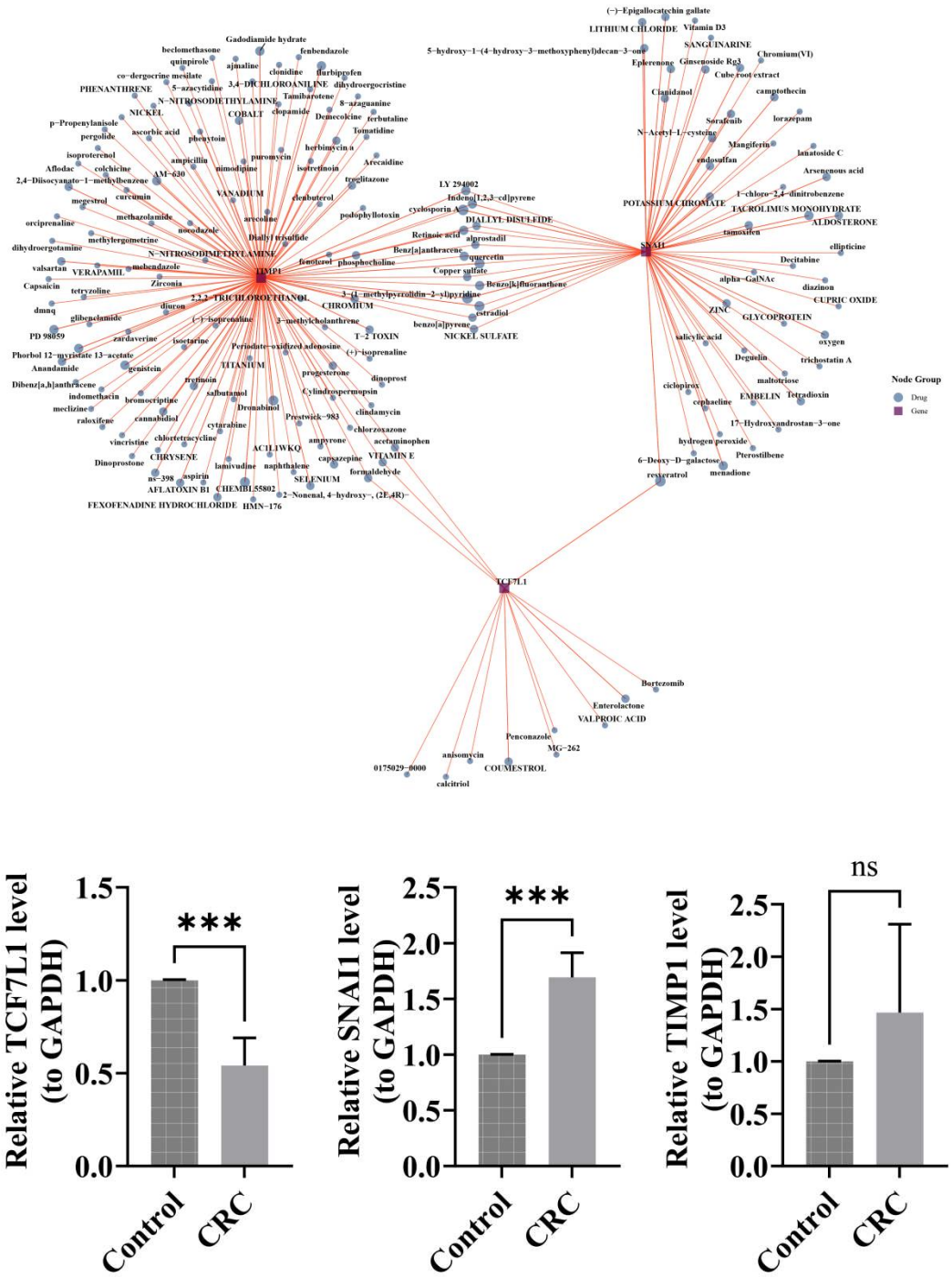


**Fig.11** RT-qPCR validation of SNAI1, TCF7L1, and TIMP1 expression in colorectal cancer (CRC) tissues and adjacent non-tumor tissues. ***P < 0.001; ns, not significant.

**Tables**

**Supplementary Table 1** The 261 DRGs collected in this research.

**Supplementary Table 2** The 3,191 DEGs identified in this study.

**Supplementary Table 3** The 39 candidate genes related to CRC and dephosphorylation identified in this study.

**Supplementary Table 4** The 827 GO terms significantly enriched in candidate genes.

**Supplementary Table 5** The 96 KEGG pathways significantly enriched in candidate genes.

**Supplementary Table 6** The results of univariate Cox regression and proportional hazards (PH) assumption test for candidate prognostic genes.

**Supplementary Table 7** The 39 KEGG pathways with significant differences between the HRG and LRG.

**Supplementary Table 8** The correlation coefficients between prognostic genes and DIICs.

**Supplementary Table 9** The 198 common chemotherapeutic drugs.

**Supplementary Table 10** The 150 chemotherapeutic drugs with significantly different IC50 values between HRG and LRG.

**Supplementary Table 11** The correlation coefficients between prognostic genes and key chemotherapeutic drugs.

**Supplementary Table 12** The 71 KEGG pathways significantly enriched with SNAI1.

**Supplementary Table 13** The 85 KEGG pathways significantly enriched with TCF7L1.

**Supplementary Table 14** The 69 KEGG pathways significantly enriched with TIMP1.

**Supplementary Table 15** The 39 KEGG pathways co-enriched with prognostic genes.

**Table 1** PCR primer information.

Table 1 PCR primer information

| Primer | Sequence |
|---|---|
| TCF7L1 F | CCTCTCCCTCACCACCAAAC |
| TCF7L1 R | GCTTAGTGGGCAGACTTGGT |
| SNAI1 F | CTCGGACCTTCTCCCGAATG |
| SNAI1 R | AAAGTCCTGTGGGGCTGATG |
| TIMP1 F | CATCCGGTTCGTCTACACCC |
| TIMP1 R | GTGCATTCCTCACAGCCAAC |
| Internal reference gene GAPDH F | ATGGGCAGCCGTTAGGAAAG |
| Internal reference gene GAPDH R | AGGAAAAGCATCACCCGGAG |

## Supplementary Files

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

- SupplementaryTable1.csv
- SupplementaryTable6.csv
- SupplementaryTable11.csv
- SupplementaryTable3.csv
- SupplementaryTable5.xlsx
- SupplementaryTable8.csv
- SupplementaryTable7.csv
- SupplementaryTable10.csv
- SupplementaryTable9.xls
- SupplementaryTable14.csv
- SupplementaryTable13.csv
- SupplementaryTable15.xls
- SupplementaryTable12.csv
- SupplementaryTable4.xlsx
- SupplementaryTable2.csv