Bioinformatic analysis of autophagy-related genes for prognosis, malignancy and immune phenotype of colorectal cancer

Zhiming Jin (jzmgyp@hotmail.com)
Shanghai Jiao Tong University Affiliated Sixth People's Hospital

Juanjuan Tu
Shanghai Jiao Tong University Affiliated Sixth People's Hospital

Chengsheng Ding
Shanghai Jiao Tong University Affiliated Sixth People's Hospital

Zezhi Shan
Fudan University Shanghai Cancer Center

Mengcheng Li
Shanghai Jiao Tong University Affiliated Sixth People's Hospital

Yuping Gao
XinHua Hospital

Research Article

Keywords: Autophagy, Colorectal Neoplasms, Prognosis, Survival analysis

Posted Date: April 21st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1566232/v1

Abstract

Background: Autophagy plays a non-negligible role in carcinogenesis by maintaining a proper cancer microenvironment to provide nutritional supplements. This study aimed to explore the association between autophagy-related genes (ARGs) and colorectal cancer (CRC).

Material and Methods: Public data profiles were downloaded from The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO) databases: GSE39582 set. In the TCGA cohort, 7 ARGs were selected to construct the prognostic risk score model for the prognosis of CRC patients. Based on the median value of risk scores, patients were divided into low- and high- risk score groups (p=4.393e−06). And the prognostic value of the model was further validated in the GEO cohort and the pooled cohort.

Results: In addition to prognosis, autophagy was also related to malignancy, chemotherapy resistance, as well as immune type of CRC. Furthermore, a nomogram, including the prognostic risk score model and clinical characteristics, was developed for quantifying the overall survival probability of CRC patients. Finally, some drugs/compounds including semagacestat, pentostatin, DNMDP, as well as bexarotene, were predicted as potential drugs for CRC treatment based on CTRP and PRISM database.

Conclusion: These would assist CRC patients with improving their prognosis.

Background

Colorectal cancer (CRC) is one of the most common malignant tumors worldwide. The mortality of CRC ranks second among 36 cancer types in the world, with nearly 900,000 deaths each year [1]. Its incidence is the third highest in men, second highest in women among all types of cancers[2]. At the same time, the incidence of CRC also varies geographically, which is higher in the developed countries as compared to that in the developing countries[3]. With the rapidly progress in developing countries, Arnold M et al.[4] expect that the incidence of CRC will increase to 2.5 million new cases in 2035. There are many risk factors closely related to CRC, including aging, unhealthy lifestyle, and dietary changes. Through the nationwide screening program and colonoscopy, the incidence of CRC can be stabilized or even reduced[5]. However, the mortality and postoperative recurrence rates of CRC are not significantly declined., 25%-40% of postoperative patients experience tumor recurrence with an unsatisfactory prognosis, even with chemotherapy[6]. The survival time of CRC patients is closely related to pathological staging. The majority of CRC patients in early stage have a satisfactory prognosis with a 5-year survival rate of > 90%[7]. As for patients with advanced cancer, their 5-year survival rate is even lower than 10%[7, 8]. Unfortunately, there is still no accurate and effective technological means to evaluate the prognosis of postoperative patients with CRC. The current mainstream clinical indicators such as pathological staging and AJCC-TNM staging cannot accurately determine the prognosis of patients. This brings a barrier to individualized treatment of patients. Hence, there is an urgent need for a method that can more accurately assess the prognosis of CRC patients.

Autophagy is a tightly coordinated process that misfolded proteins and dysfunctional organelles are sequestered in the autophagosomal vesicles and eventually transported to lysosomes for degradation[9]. Therefore, it plays an important role in maintaining intracellular homeostasis. Dysregulated autophagy is associated with many diseases such as obesity, type 2 diabetes and nonalcoholic fatty liver[10, 11]. The most important of these is its role in cancer. Autophagy plays a dual role in cancer. On the one hand, autophagy can prevent cancer development. On the other hand, autophagy helps maintain a proper cancer microenvironment to provide nutritional supplements under adverse conditions such as starvation and hypoxia[12]. And enhanced autophagy promotes tumor cell survival and growth with autophagic flux when tumor is established[13, 14]. The role of autophagy in tumor development depends on the type of cells and tumor microenvironment. In terms of drug treatment, autophagy plays a non-negligible role in promoting drug resistance[15]. For example, in the treatment of gastrointestinal stromal tumor (GIST) cells, autophagy was induced to resist to Imatinib™. However, inhibition of autophagy by the lysosomotropic agent chloroquine (CQ) could overcome resistance to Imatinib™ and cause tumor cell apoptosis[16].

At present, the pathological stage and AJCC-TNM stage of patients are used as tools to guide treatment and predict the overall survival in CRC. However, even for CRC patients who are at the same tumor stage, there are many significant differences in sensitivity to treatment and prognosis because of the high heterogeneity of CRC. Therefore, finding new biomarkers or signatures to predict the overall survival of CRC patients has been our main goal. Recently, in other cancers, some studies have proven that gene signature is capable of predicting the prognosis of cancer patients. Hu et al.[17] have confirmed that autophagy-related gene expression signature can predict prognosis in prostate cancer patients .

There are many studies demonstrating that prognostic signatures based on miRNA or lncRNA have abilities to predict the prognosis of CRC patients[18–20]. In the current study, we attempted to explore the relationship between autophagy and colorectal cancer. We downloaded gene expression profiles of CRC from TCGA and GEO databases. In the TCGA cohort, a prognostic risk score, which contained 7 ARGs, was constructed for predicting OS of CRC patients and validated in the GEO cohort and the pooled cohort. Besides, we developed a nomogram as a tool to evaluate overall survival probability and to manage patients better.

Results

1 Differentially expressed ARGs between CRC and adjacent normal colorectum tissue samples

A total of 568 CRC and 44 normal colorectum tissue samples with RNA-seq data in the TCGA cohort were involved in the current study. We selected 222 identical ARGs from 232 ARGs and RNA-seq data of TCGA cohort. Among the 222 identical ARGs, 32 differentially expressed ARGs were identified between CRC and adjacent normal colorectum tissue samples with thresholds of |log2 fold change|> 1 and FDR < 0.05. The result was shown in a heat map (Fig. 1A) Then, according to differentially expressed ARGs, gene ontology

(GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed. Autophagy, mitochondrial outer membrane, and ubiquitin protein ligase binding were the top enrichment GO terms for biological processes, cellular components and molecular functions, respectively. As for KEGG pathway enrichment analysis, apoptosis, p53 signaling pathway and human cytomegalovirus infection were the top enrichment pathways and autophagy related pathways were also statistically significant (Fig. 1B-C).

2 Development and validation of a prognostic risk score model

In the TCGA cohort, we merged the expression matrixes of 222 ARGs with the corresponding overall survival time and overall survival status for 541 CRC patients on the basis of patients' ID. The univariate Cox regression analysis was utilized to select out a total of 8 ARGs as candidate genes that were significantly associated with OS (Fig. 1D). And the differential expression levels of these candidate genes between CRC and adjacent normal colorectum tissue samples were shown in Fig. 1E. Subsequently, we excluded 26 CRC patients whose OS time was 0, and a total of 7 optimal genes were identified for the construction of the prognostic risk score model in the LASSO Cox regression analysis (Fig. 1F, G). The optimal genes are MAPK9, CDKN2A, SERPINA1, DAPK1, ULK3, EDEM1 and ULK1. The prognostic risk score = (-0.211578695097183* expression value of MAPK9) + (0.0426772420693879* expression value of CDKN2A) + (-0.14956314324999* expression value of SERPINA1) + (0.348009461063805* expression value of DAPK1) + (0.250327783852896* expression value of ULK3) + (-0.432074832788178* expression value of EDEM1) + (0.226888644181894* expression value of ULK1).

According to the median value of risk scores, we divided CRC patients into low- and high- risk score groups. Kaplan-Meier survival curves showed that high-risk group had a lower survival rate than low- risk group with P-value = 4.393e-06 (Fig. 2A). The ROC of the prognostic risk score model at 5-year was exhibited in Fig. 2B, with AUC of 0.678. To better demonstrate the association between risk scores and prognosis, the risk score of each CRC patient was ranked from low to high (Fig. 2C). The survival status and corresponding risk scores were plotted in Fig. 2D. Red dots represented dead patients and blue dots represented surviving patients. As shown in the figure, the red dot density of the low risk score group was significantly lower than that of the high risk score group. The differentially expressed heat map of optimal genes, between the low- and high- risk score groups was reported in Fig. 2E. For validation of the prognostic risk score model, we further compared the survival differences between low- and high- risk score groups in the GEO cohort. High-risk group also had a poorer prognostic than low-risk group with P-value of 4.707e − 02 (Supplement Fig. 1A) and the AUC of ROC at 5-year was shown in Supplement Fig. 1B. The distribution of risk score of each patient, survival status and corresponding risk scores, as well as the differentially expressed heat map of optimal genes were plotted in Supplement Fig. 1C-E. As expected, the prognostic risk score models also exhibited reliable ability to predict prognosis in the pooled cohort. Survival curve revealed that patients in the high risk score group had worse OS compared to those in the low risk score group with AUC of 0.638 (p = 7.7e − 05; Supplement Fig. 2A, B). Supplement Fig. 2C-E

respectively, demonstrated the distribution of the risk score of each patient, survival status and corresponding risk scores, as well as the differentially expressed heat maps of optimal genes in the pooled cohort. In addition to this, we explored the relationship of risk scores and clinical characteristics and found that patients with advanced cancer had higher risk scores than early stage patients (Fig. 3A-D). The expression levels of CEA were positively correlated with the risk scores of patients (Fig. 3E). Figure 3F-H demonstrated the differences of PD1, PD-L1, CTLA-4 between low- and high- risk score groups, respectively. However, there were no significant differences of risk scores in terms of patients' age and gender (Supplement Fig. 3A, B).

3 Chemotherapy in the GEO cohort

According to the above criteria, 139 patients were defined to non-chemotherapy resistance group and 111 patients were defined to chemotherapy resistance group. Then, we compared the difference of risk scores between non-chemotherapy resistance and chemotherapy resistance groups, and we found that the median value of risk scores in the chemotherapy resistance group was higher than that in non-chemotherapy resistance group with P-value of 0.03 (Fig. 3I). This result meant that patients with low risk scores could get more benefits from chemotherapy than those with high risk scores. Therefore, it is particularly important for patients with high risk scores to find potentially effective treatments.

4 PCA in the TCGA and GEO cohorts

In both TCGA and GEO cohorts, the expression matrixes of all ARGs were analyzed using PCA respectively. The results of PCA were shown in Fig. 4A, B. After that, we performed PCA on the expression matrixes of the optimal genes that identified by LASSO Cox regression analysis (Fig. 4C, D). As shown in the figures, compared with all ARGs, patients with different risk scores could be distinguished better using the optimal genes.

5 The results of enrichment analysis

The enrichment analyses of differentially expressed genes between low- and high- risk score groups were conducted with GO and KEGG pathway enrichment analyses. The results of GO and KEGG pathway enrichment analyses were demonstrated in Fig. 5A, B. Many immune-related biological processes and cancer- related pathways were the top enrichment terms and pathways, which indicated that there might be cancer-related immune differences between low- and high- risk score groups. Meanwhile, the results of GSEA showed that some cancer-related pathways, such as NOTCH signaling pathway and HEDGEHOG signaling pathway were enriched in high- risk score group (Fig. 5C, D). In addition, the p53 signaling pathway was upregulated in the low risk score group (Fig. 5E).

6 The difference of immune cell infiltration between low- and high- risk score groups

CIBERSORT algorithm was applied to quantify the relative abundance of 22 immune cells in each patient according to the RNA-seq profiles. Figure 5F presents the differences of immune cell infiltration between low- and high- risk score groups.

As shown in figure, the abundance of plasma cells, resting memory CD4 T cells, activated dendritic cells, Eosinophils and neutrophils were higher in the low risk score group compared to the high risk score group. However, regulatory T cells (Tregs) and Macrophages M0 exhibited higher infiltration in patients from the high risk score group. Based on the above results, patients with low risk scores were more likely to benefit from immunotherapy than those with high risk scores.

7 PPI network of differentially expressed ARGs between low- and high- risk score groups

We utilized the STRING online database to construct the PPI network of differentially expressed ARGs between low- and high- risk score groups, the result was shown in Fig. 6A and Supplement Table 1. Then, Cytoscape software was performed to process the data of PPI network further. Figure 6B presents the interaction of differentially expressed ARGs, red presents ARGs that are highly expressed in high- risk score group and blue presents ARGs that are highly expressed in low risk score group. Finally, we found the top 10 hub ARGs in the PPI network using cytoHubba. As shown in the Fig. 6C, the top 10 hub AGRs, degree ranked from high to low, are BECN1, ATG5, ATG16L1, PIK3R4, MAP1LC3A, MAP1LC3B, ULK1, ATG13, MAPK3 and ATG3.

8 Development of a nomogram for predicting overall survival

To predict the OS of CRC patients, we constructed a nomogram that integrated age, gender, pathological stage, neoadjuvant chemotherapy, the level of CEA and a prognostic risk score model (Fig. 7A). And the calibration curves at 1-year, 2-year, 3-year and 5-year demonstrated that the nomogram had a good ability to predict the OS of CRC patients (Fig. 7B-E). The results of univariate and multivariate Cox regression analyses proved that the prognostic risk score model, age and pathological stage could be regarded as independent prognostic indicators (Fig. 7F, G). The ROC curves demonstrated that the prognostic risk score model (AUC = 0.702) presented a good prognostic value than that of age (AUC = 0.623; Fig. 7H). Meanwhile, univariate and multivariate Cox regression analyses of GEO and pooled cohorts confirmed that the prognostic risk score model could act as an independent prognostic indicator (Supplement Fig. 4A-D).

9 Identification of potential drug targets for CRC patients

Based on the gene expression profiles and drug sensitivity profiles of CCLs in the CTRP and PRISM database, “pRRophetic” R package was applied to estimate the response of CRC patients to drugs/compounds via the built-in ridge regression model. The estimated AUC value represented the sensitivity of patients to drugs and compounds, lower AUC value indicated higher sensitivity to drugs. To further identify potential drugs for patients with different risk scores, Spearman correlation analysis was performed to quantify the correlation coefficient between AUC values and risk scores. The AUC value of CTRP-derived drugs, including DNMDP, brefeldin A, azacitidine, gossypol, PF-543 and L-685458, was positively related with risk scores; and the AUC value of CTRP-derived compounds, including semagacestat, epigallocatechin-3-monogallate, SMER-3, NSC48300, CHIR-99021, CHIR-99021, was negatively related with risk scores (Fig. 8A). In addition, the analysis in PRISM database yielded the top 6 compounds with positive correlation coefficient (bexarotene, zaleplon, propranolol, clofoctol, 3-amino-benzamide, selumetinib) and the top 6 compounds with negative correlation coefficient (pentostatin, afobazole, regorafenib, tofogliflozin, mepivacaine, tedizolid-phosphate) (Fig. 8B). Compounds with negative correlation coefficient had lower estimated AUC values in the high risk score group, which indicated that these compounds could act as potential drugs for CRC patients with high risk scores. In addition to immunotherapy, patients with low risk scores also benefited from the compounds with positive correlation coefficient, because these compounds had lower estimated AUC values in the low risk score group.

10 The results of HPA

We compared the expression level of the optimal genes between CRC tissue and normal colorectum tissue samples by immunohistochemistry in the HPA database. The results were shown in Supplement Fig. 5, and ULK1 gene was not included because no information was available in the HPA database.

Discussion

Some studies have validated that autophagy is involved in some diseases, such as cancer, and plays a role in promoting advanced cancer growth [27, 28]. Beyond that, Zhou et al. demonstrated that 5 core autophagy genes (CAPN10, DAPK2, DNAJB9, GNAI3, PPP1R15A) could help to predict the recurrence of colorectal cancer in early phase[29].

In the current study, we obtained the expression profiles of ARGs from public online databases and focused on exploring the relationship between ARGs and the overall survival of CRC patients. Firstly, 32 differentially expressed ARGs were identified between CRC and normal colorectum tissue samples. GO and KEGG pathway enrichment analyses were utilized to process these ARGs, and the results indicated that many differentially expressed ARGs played an important role in autophagy and cancer-related pathways. Then, we utilized the univariate Cox regression analysis to screen out 8 ARGs from CRC

patients of TCGA cohort. Further, the LASSO Cox regression analysis was performed to identify 7 ARGs (MAPK9, CDKN2A, SERPINA1, DAPK1, ULK3, EDEM1, ULK1) as the optimal genes for constructing the prognostic risk score model, which could be used as an independent indicator to predict the prognosis of CRC patients. And the results from the GEO cohort and the pooled cohort illustrated that the prognostic risk score model had a general applicability in CRC patients from other cohorts. Subsequently, we found that patients' risk scores calculated by a specific formula were associated with clinical characteristics and the expression levels of immune checkpoints (PD1, PD-L1, CTLA-4). The advanced-stage patients had higher risk scores than the patients at the early clinical stage. We further classified patients into low- and high- risk score groups according to the median value of risk scores, patients with high risk score had poorer overall survival than those in the low risk score group. To look for the potential mechanisms behind it, we compared the differences of enrichment analyses between low- and high- risk score groups. Many cancer-related pathways were enriched in the high risk score group. And interestingly, we found that patients with higher risk score seemed to be more likely to be resistant to chemotherapy according to the results of GEO cohort. A large number of studies on other cancer types and responses to other cancer therapies have confirmed that autophagy is both caused by the treatment used and produces resistance to the treatment[15, 30]. However, the mechanism of treatment-induced autophagy is very complicated and not fully understood. P53 signaling pathway activation induced by DNA damage may illustrate that how conventional genotoxic drugs, such as radiotherapy or cisplatin, induce autophagy through p53-mediated autophagy regulators such as DRAM (damage-regulated autophagy modulator)[31]. Furthermore, Simon et al.[32] have demonstrated that p53 is proapoptotic, and can also inhibit autophagy according to specific conditions. The role of p53 in these responses is consistent with our results.

Taking these results together, it was essential for patients with a high risk scores that received reinforced follow-up and active therapeutic intervention, while exploring new potential drugs/compounds for CRC patients was particularly important. Therefore, we developed a nomogram, which integrated age, gender, pathological stage, neoadjuvant chemotherapy, the level of CEA and a prognostic risk score model for predicting the 2-, 3- and 5-year overall survival probabilities of CRC patients. The results were satisfactory, which demonstrated that the prognostic risk score model was extensively applicable in CRC patients. Finally, we predicted some drugs/compounds that might be effective for patients with a variety of risk scores via analysis of CTRP and PRISM database. Drugs with negative coefficients, such as semagacestat, pentostatin and so on, might play a significant role in treating patients with high risk scores; and those patients with low risk scores probably benefited from the therapy containing drugs with positive coefficients, represented by DNMDP and bexarotene. This provides novel insights for drug therapy strategies for colorectal cancer in the future.

In conclusion, we explored the association between ARGs and CRC. This study reveals that some specific ARGs play a non-negligible role in the classification of CRC, prognosis, as well as prediction of potential drugs. However, there are still some limitations to this study. First, it is a retrospective study with data from the TCGA and GEO databases, which is short of validation in prospective clinical trials. Second, the threshold value that classifies patients into low- and high- risk score groups may vary depending on the method, so other methods are needed for study validation in the future. Third, the information of clinical

pathological characteristics such as pathological grading was not complete, which may affect the accuracy of the current study.

Conclusion

In conclusion, we analyzed the autophagy-related gene expression matrixes of TCGA and GEO databases. A total of 7 ARGs were screened out to develop a prognostic risk score model that had the ability of predict the overall survival of CRC patients accurately. Beyond that, we constructed a nomogram, which integrated the prognostic risk score model and clinical characteristics, to quantify the overall survival probability in CRC patients. According to these results, the prognostic risk score model can be regarded as a potential prognostic marker in CRC and nomogram can play a significant role in the screening of high-risk patients who require reinforced follow-up and positive therapeutic intervention. Some drugs/compounds were identified as potential drugs for CRC treatment. Further validation is needed prior to clinical application.

Material And Methods

1 Data acquisition

The RNA sequencing (RNA-seq) data of colorectal cancer (CRC) patients with HTSeq-FPKM workflow type were downloaded from the TCGA database (https://tcga-data.nci.nih.gov/tcga/), which contains 568 CRC and 44 adjacent normal colorectum tissue samples. Meanwhile, we also obtained the clinical information of 548 CRC patients using GDC data portal (https://portal.gdc.cancer.gov/), such as overall survival time, survival state, pathological stages. The microarray data and corresponding clinical information of GSE39582 were downloaded from GEO database (https://www.ncbi.nlm.nih.gov/geo/). The microarray data was generated by the Affymetrix Human Genome U133 Plus 2.0 Array and the annotation of gene symbols was based on the corresponding probe in the GPL570 platform. For subsequent analysis, Perl script based on the JAVA8 platform was applied to process and normalize the microarray data, while the probe ID was converted into the corresponding gene symbol based on the annotation file. When more than one probes were matched to the same symbol ID, mean value was taken. Furthermore, patients in TCGA database and GEO database were combined for pooled validation. ComBat method was performed to remove the batch effect between TCGA and GEO databases via “sva” R package.

Besides, in the current study, a total of 232 genes corresponding to autophagy were acquired from the Human Autophagy Database (http://www.autophagy.lu/index .html). Perl script was utilized to screen out the overlapping genes from 232 autophagy-related genes (ARGs) and the RNA expression matrix of TCGA cohort.

2 Enrichment analysis of the differentially expressed ARGs

In the TCGA cohort, the differentially expressed genes (DEGs) of ARGs between CRC and adjacent normal colorectum tissue samples were retained using the “limma” package in R, which with |log2 fold change|> 1 and the false discovery rate (FDR) < 0.05. Then, gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed to reveal the major biological features and cell functional pathways of DEGs using “clusterProfiler” package in R. Adjusted P-value < 0.05 was considered statistically significant. Next, the visual processing of the annotation results was performed by “enrichplot” and “ggplot2” packages in R.

3 Establishment and validation of a prognostic risk score model

In the TCGA cohort, the data analysis of the expression level of each ARG and CRC patients’ overall survival were performed using univariate Cox regression analysis, ARGs with $p < 0.05$ were considered as candidate genes. Then, we utilized the LASSO Cox regression analysis to narrow candidate genes for avoiding the interference of overfitting. Each candidate gene was screened based on the LASSO Cox regression using “glmnet” package in R, the candidate genes with nonzero regression coefficients were selected as optimal genes for the construction of the prognostic risk score model. To determine the penalty parameter (λ) of the model, the tenfold cross-validation was performed. The formula applied to calculate the prognostic risk score of each patient was as follows:

Risk score = $\sum_1^i (\mathrm{Coef}i * \mathrm{ExpGene}i)$ (1)The “Coef” represents nonzero regression coefficients from the LASSO Cox regression analysis and “ExpGene” is the expression value of optimal genes of prognostic risk score model. The risk score of each patient in GEO cohort and the pooled cohort was also calculated using the formula mentioned above. The median value of risk scores was regarded as a threshold that could classify all CRC patients into low- and high- risk score groups in TCGA cohort. Next, the comparison of the survival difference between low- and high- risk score groups was conducted by Kaplan-Meier analysis with the log-rank test. Besides, “survivalROC” package in R was utilized to plot the receiver operating characteristic (ROC) curve for assessing the predictive accuracy of the prognostic risk score model. Finally, we evaluated the potential relationship of risk scores and clinical characteristics. We also compared the expression levels of immune checkpoints (PD1, PD-L1, CTLA-4) between low- and high- risk score groups. When comparing the two groups, Wilcoxon rank-sum test was utilized and when comparing three or more groups, Kruskal-Wallis (K-W) test was utilized.

4 The relationship between chemotherapy and risk scores in CRC patients

In the GEO cohort, we selected two groups of patients with chemotherapy resistance and non-chemotherapy resistance. According to the study of Mini et al.[21]. The patients with disease recurrence within 3 years from chemotherapy were classified into the chemotherapy resistance group and patients

without disease recurrence within 5 years were considered as non-chemotherapy resistant. Then we obtained the corresponding risk score of each patient in the two groups according to patients' ID. The difference of risk scores between chemotherapy resistance and nonchemotherapy groups was explored with the Wilcox-test, which P-value < 0.05 was considered as statistically significance.

5 The results of principal component analysis (PCA) before and after modeling

In both TCGA and GEO cohorts, PCA was utilized to analyze the expression matrixes of all ARGs for each patient with "limma" package in R, respectively. Subsequently, the expression matrixes of the optimal genes from LASSO Cox regression analysis were also processed by PCA. To display the results of PCA, "ggpolt2" R package was performed and each patient was drawn as a point on two-dimensional diagrams on the basis of the first two principal components.

6 Enrichment analysis between low- and high- risk score groups

We utilized "limma" R package to screen out the differentially expressed genes (DEGs) from all genes between low- and high- risk score groups, which with thresholds of |log2 fold change|> 1 and FDR < 0.05. Then, we acquired the corresponding Entrez ID of each DEG using "org.Hs.eg.db" package in R. The GO and KEGG pathway enrichment analyses were performed on the DEGs, respectively with "clusterProfiler" package. Next, we visualized the results of enrichment analyses using "enrichplot" and "ggplot2" packages. Finally, we conducted gene set enrichment analysis (GSEA) based on GSEA software (version: 4.0.3) between low- and high- risk score groups using "c2.cp.kegg.v7.1.symbols.gmt gene sets" as a reference gene set. And enrichment pathways with FDR < 0.25 were considered statistically significant.

7 Evaluation of immune cell infiltration via CIBERSORT algorithm

To quantify the immune cell infiltration of each patient, CIBERSORT algorithm (R script v1.03) was applied to estimate the relative abundance of different immune cells in a mixed cell population. The analysis of CIBERSORT algorithm was based on a known reference set, which contained a set of gene expression features of 22 leukocyte subtypes-LM22, such as CD8 T cells, plasma cells, NK cells, as well as dendritic cells. Then, we compared the differences of immune cell infiltration between low- and high-risk score groups with the Wilcoxon rank-sum test and P-value < 0.05 was considered statistically significant. Finally, the results of the comparison were shown in box plots using "ggpubr" package in R.

8 Protein-protein interaction (PPI) network

In the current study, we utilized the “limma” package to select differentially expressed genes from autophagy-related genes between low- and high- risk score groups with |log2 fold change|> 1 and a false discovery rate < 0.05. The STRING online database (version: 11.0; https://string-db.org/) was used to produce PPI network data of differentially expressed genes with an interaction score > 0.95 (very high confidence). Subsequently, Cytoscape software (version: 3.7.2) was utilized to visualize the PPI network data. The Cytohubba (version: 0.1) is a plug-in of Cytoscape that can predict and explore important nodes and subnetworks in a given network by several topological algorithms. Therefore, we utilized cytoHubba to identify the top 10 hub genes from PPI network according to the degree rank of genes in a given network.

9 Construction of a nomogram for predicting overall survival (OS)

To predict the OS of CRC patients, we utilized the “rms” package in R to construct a nomogram that integrated age, gender, pathological stage, neoadjuvant chemotherapy, the level of CEA and a prognostic risk score model based on the TCGA cohort. The calibration curves at 1-year, 2-year, 3-year, and 5-year were plotted to evaluate the accuracy of OS prediction in the nomogram. Univariate and multivariate Cox regression analyses were performed by nomogram to validate the independent prognostic value of the risk score model, which P-value < 0.05 was considered as statistically significance. Besides, the effectiveness of OS prediction of the risk score model was compared with other clinical characteristics of nomogram through the Online ROC Curves (http://www.houshixu.cn:3838/sample-apps/ROC/).

10 Prediction of potential drugs for CRC

According to the patient's risk score, some compounds/drugs were predicted as potential treatment for CRC patients. Based on previous study[22], the expression profile data and somatic mutation data of human cancer cell lines (CCLs) were obtained from the Broad Institute Cancer Cell Line Encyclopedia (CCLE) project[23]. And we acquired the CCLs drug sensitivity data from the Cancer Therapeutics Response Portal (CTRP) (https://portals.broadinstitute.org/ctrp) and PRISM Repurposing dataset (https://depmap.org/portal/prism/). CTRP and PRISM have sensitivity data for 481 compounds in 835 CCL and 1448 compounds in 482 CCL, respectively. The area under the dose–response curve (AUC) represents the sensitivity to compounds/drugs, and lower AUC means higher sensitivity. Compounds/drugs lacking of more than 20% data were excluded. Before predicting drug sensitivity, k-nearest neighbor (k-NN) imputation was used to identify and replace missing AUC values.

11 The Human Protein Atlas (HPA)

HPA (version: 19.3; https://www.proteinatlas.org/)[24–26] is a program that aims to map all human proteins in cells, tissues and organs using the integration of various omics technologies, including

antibody-based imaging, mass spectrometry based proteomics, transcriptomics and systems biology. The tissue atlas of HPA can show the expression level of human proteins based on immunohistochemistry, which contains 44 different tissue types. Therefore, we utilized HPA to explore the expression levels of each optimal gene based on immunohistochemistry in colorectal cancer and normal tissue samples.

12 Statistical analysis

Wilcoxon rank-sum test was utilized to compare the difference in two groups, whereas the Kruskal-Wallis test was performed to compare three or more groups. Kaplan-Meier analysis was applied to evaluate the survival differences between low- and high- risk score groups. Univariate Cox regression analysis and multivariate Cox regression analysis were performed to determine independent indicators for predicting OS in colorectal cancer. The ROC curves were plotted to access the predictive effectiveness of the prognostic risk score mode and nomogram. All statistical analysis was conducted in R 4.0.0 ($p < 0.05$).

Declarations

Ethics approval and consent to participate

The patient data in this work were acquired from the publicly available datasets whose informed consent of patients were complete.

Consent for publication

All authors have agreed to publish this work.

Availability of data and material

In this study, gene expression data were collected from public databases (TCGA, GEO, CCLE).

Competing interests

All authors have no conflicts of interest to declare.

Funding

This work was supported by funding from Science and Technology Commission of Shanghai Municipality (No.16140900302)

Authors' contributions

J.T. ,C.D. and Z.S. collected and analyzed research data. M.L. was responsible for the interpretation of the results, J.T. and C.D. drafted the manuscript, Z.J. and Y.G. performed in the study design and revised the manuscript.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018; 68(6):394-424.
2. Dekker E, Tanis PJ, Vleugels JLA, Kasi PM, Wallace MB. Colorectal cancer. Lancet (London, England). 2019; 394(10207):1467-80.
3. Brenner H, Chen C. The colorectal cancer epidemic: challenges and opportunities for primary, secondary and tertiary prevention. British journal of cancer. 2018; 119(7):785-92.
4. Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global patterns and trends in colorectal cancer incidence and mortality. Gut. 2017; 66(4):683-91.
5. Ait Ouakrim D, Pizot C, Boniol M, et al. Trends in colorectal cancer mortality in Europe: retrospective analysis of the WHO mortality database. BMJ. 2015; 351:h4970.
6. Tjandra JJ, Chan MKY. Follow-up after curative resection of colorectal cancer: a meta-analysis. Dis Colon Rectum. 2007; 50(11):1783-99.
7. O'Connell JB, Maggard MA, Ko CY. Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging. Journal of the National Cancer Institute. 2004; 96(19):1420-25.
8. Siegel R, Desantis C, Jemal A. Colorectal cancer statistics, 2014. CA: a cancer journal for clinicians. 2014; 64(2):104-17.
9. Behrends C, Sowa ME, Gygi SP, Harper JW. Network organization of the human autophagy system. Nature. 2010; 466(7302):68-76.
10. Sarparanta J, García-Macia M, Singh R. Autophagy and Mitochondria in Obesity and Type 2 Diabetes. Curr Diabetes Rev. 2017; 13(4):352-69.
11. Mao Y, Yu F, Wang J, Guo C, Fan X. Autophagy: a new target for nonalcoholic fatty liver disease therapy. Hepat Med. 2016; 8:27-37.
12. Bermúdez M, Aguilar-Medina M, Lizárraga-Verdugo E, et al. LncRNAs as Regulators of Autophagy and Drug Resistance in Colorectal Cancer. Front Oncol. 2019; 9:1008.
13. Amaravadi R, Kimmelman AC, White E. Recent insights into the function of autophagy in cancer. Genes & development. 2016; 30(17):1913-30.
14. White E. Deconvoluting the context-dependent role for autophagy in cancer. Nature reviews Cancer. 2012; 12(6):401-10.
15. Levy JMM, Towers CG, Thorburn A. Targeting autophagy in cancer. Nature reviews Cancer. 2017; 17(9):528-42.
16. Gupta A, Roy S, Lazar AJ, et al. Autophagy inhibition and antimalarials promote cell death in gastrointestinal stromal tumor (GIST). Proceedings of the National Academy of Sciences of the United States of America. 2010; 107(32):14333-8.

17. Hu D, Jiang L, Luo S, et al. Development of an autophagy-related gene expression signature for prognosis prediction in prostate cancer patients. J Transl Med. 2020; 18(1):160.
18. Kandimalla R, Gao F, Matsuyama T, et al. Genome-wide Discovery and Identification of a Novel miRNA Signature for Recurrence Prediction in Stage II and III Colorectal Cancer. Clinical cancer research : an official journal of the American Association for Cancer Research. 2018; 24(16):3867-77.
19. Yang G, Zhang Y, Yang J. A Five-microRNA Signature as Prognostic Biomarker in Colorectal Cancer by Bioinformatics Analysis. Frontiers in oncology. 2019; 9:1207.
20. Zhang L, Chen S, Wang B, et al. An eight-long noncoding RNA expression signature for colorectal cancer patients' prognosis. Journal of cellular biochemistry. 2019; 120(4):5636-43.
21. Mini E, Lapucci A, Perrone G, et al. RNA sequencing reveals PNN and KCNQ1OT1 as predictive biomarkers of clinical outcome in stage III colorectal cancer patients treated with adjuvant chemotherapy. International journal of cancer. 2019; 145(9):2580-93.
22. Yang C, Huang X, Li Y, Chen J, Lv Y, Dai S. Prognosis and personalized treatment prediction in TP53-mutant hepatocellular carcinoma: an in silico strategy towards precision oncology. Brief Bioinform. 2020.
23. Ghandi M, Huang FW, Jané-Valbuena J, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019; 569(7757):503-08.
24. Uhlén M, Fagerberg L, Hallström BM, et al. Proteomics. Tissue-based map of the human proteome. Science (New York, NY). 2015; 347(6220):1260419.
25. Thul PJ, Åkesson L, Wiking M, et al. A subcellular map of the human proteome. Science (New York, NY). 2017; 356(6340).
26. Uhlen M, Zhang C, Lee S, et al. A pathology atlas of the human cancer transcriptome. Science (New York, NY). 2017; 357(6352).
27. Kaur J, Debnath J. Autophagy at the crossroads of catabolism and anabolism. Nat Rev Mol Cell Biol. 2015; 16(8):461-72.
28. Onorati AV, Dyczynski M, Ojha R, Amaravadi RK. Targeting autophagy in cancer. Cancer. 2018; 124(16):3307-18.
29. Zhou Z, Mo S, Dai W, et al. Development and Validation of an Autophagy Score Signature for the Prediction of Post-operative Survival in Colorectal Cancer. Frontiers in oncology. 2019; 9:878.
30. Galluzzi L, Bravo-San Pedro JM, Levine B, Green DR, Kroemer G. Pharmacological modulation of autophagy: therapeutic potential and persisting obstacles. Nat Rev Drug Discov. 2017; 16(7):487-511.
31. Crighton D, Wilkinson S, O'Prey J, et al. DRAM, a p53-induced modulator of autophagy, is critical for apoptosis. Cell. 2006; 126(1):121-34.
32. Simon HU, Friis R, Tait SW, Ryan KM. Retrograde signaling from autophagy modulates stress responses. Science signaling. 2017; 10(468).

Supplementary

Supplement Table 1 is not available with this version

Figures

Figure 1

Differentially expressed ARGs selection in the TCGA cohort. (**A**) The heat map of 32 differentially expressed ARGs. (**B**) The result of GO enrichment analysis of differentially expressed ARGs. (**C**) The result of KEGG enrichment analysis of differentially expressed ARGs. (**D**) Forrest plot of 8 ARGs related to prognosis after univariate Cox regression analysis. (**E**) The differential expression levels of candidate genes between normal and CRC tissue samples. (**F**) LASSO Cox regression analysis of 8 ARGs related to prognosis selected 7 genes to construct a prognostic risk score model for predicting prognosis. (**G**) LASSO coefficient profiles of the 8 ARGs with $p < 0.05$ as assessed by the univariate Cox regression analysis.

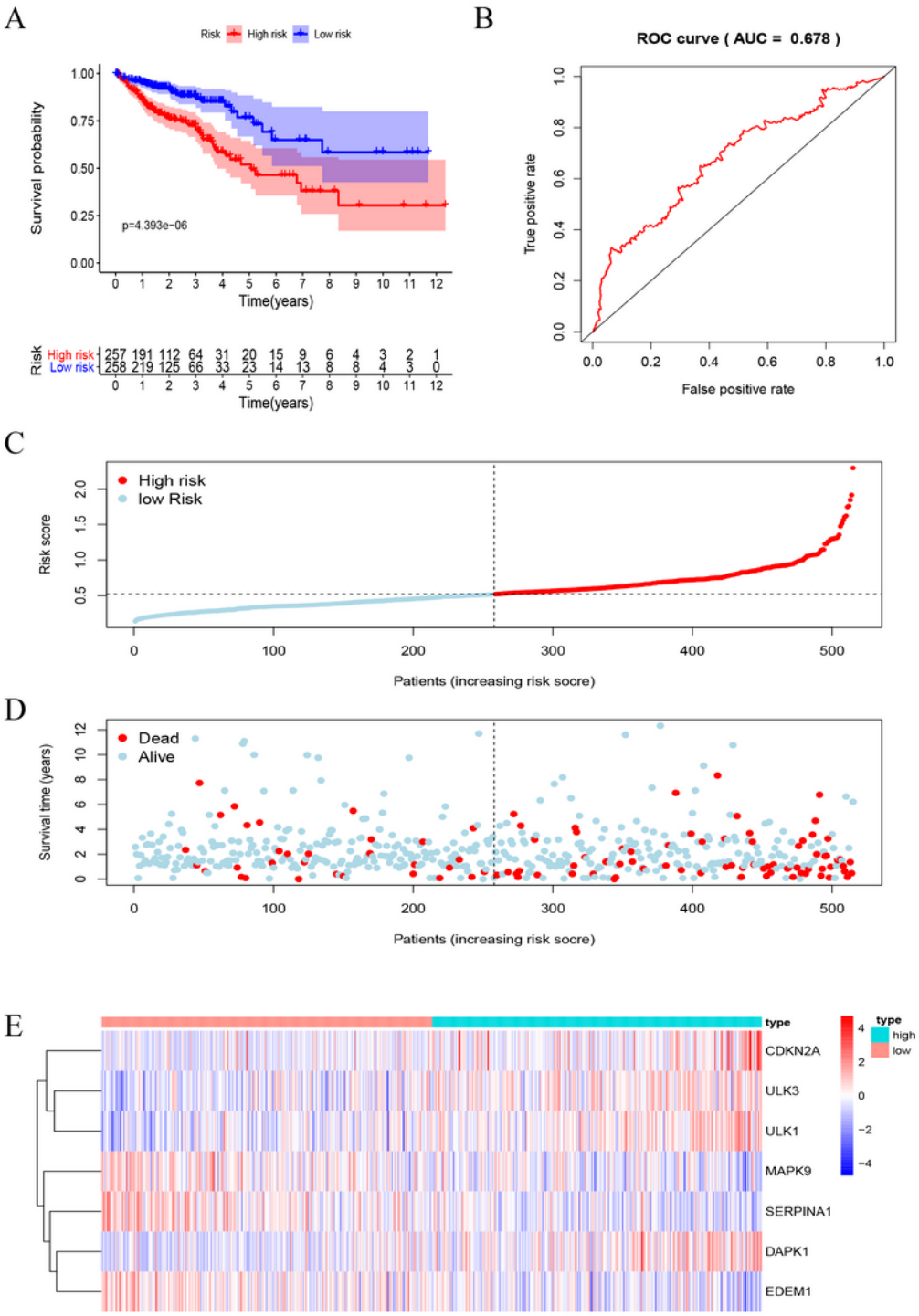

Figure 2

Prognostic value of prognostic risk score model in the TCGA cohort. (**A**) Kaplan–Meier curve of TCGA cohort. (**B**) ROC analysis of the sensitivity and specificity for OS prediction by the prognostic risk score model. (**C**) Distribution of risk scores ranked from low to high. (**D**) Comparison of survival status between low- and high- risk score groups. (**E**) Heat map of 7 genes selected by LASSO Cox regression analysis.

Figure 3

The associations of risk score with clinical characteristics, the expression level of immune checkpoints, as well as chemotherapy resistance of patients. (**A**) The association between risk scores and pathological stages in the TCGA cohort. (**B**) The association between risk scores and AJCC-T stages in the TCGA cohort. (**C**) The association between risk scores and AJCC-N stages in the TCGA cohort. (**D**) The association between risk scores and AJCC-M stages in the TCGA cohort. (**E**) The association between risk scores and the expression level of CEA in the TCGA cohort. (**F**) The association between risk scores and expression level of PD-1 in the TCGA cohort. (**G**) The association between risk scores and expression level of PD-L1 in the TCGA cohort. (**H**) The association between risk scores and expression level of CTLA-4 in the TCGA cohort. (**I**) The association between risk scores and chemotherapy resistance in the GEO cohort.

Figure 4

The results of PCA. (**A**) The result of PCA before LASSO selection in the TCGA cohort. (**B**) The result of PCA after LASSO selection in the TCGA cohort. (**C**) The result of PCA before LASSO selection in the GEO cohort. (**D**) The result of PCA before LASSO selection in the GEO cohort.

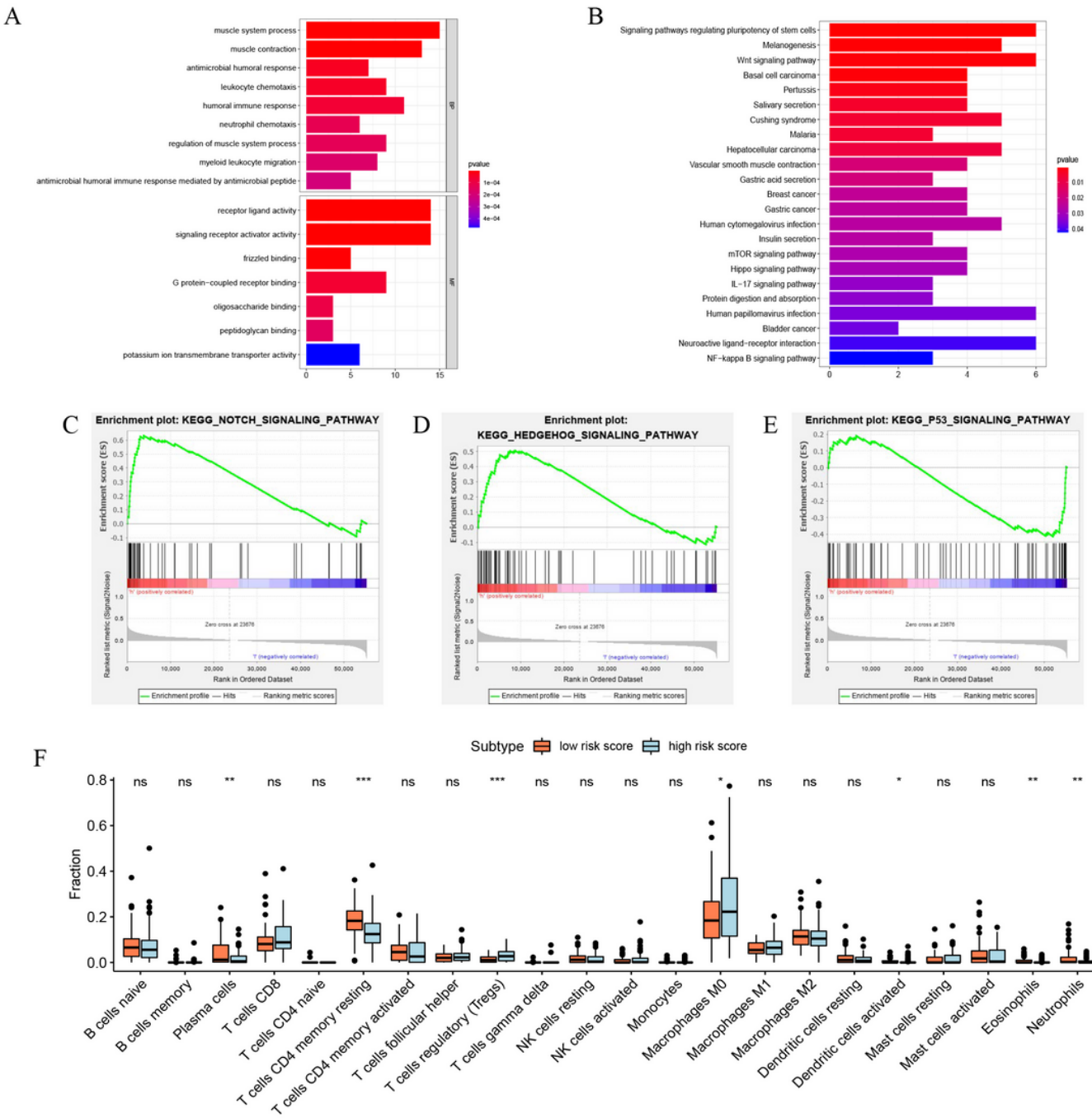

Figure 5

Comparison of enrichment analysis and immune cell infiltration between low- and high- risk score groups. (**A**) The GO enrichment analysis of genes with |log2 fold change|> 1 and FDR< 0.05. (**B**) Kegg enrichment analysis of genes with |log2 fold change|> 1 and FDR< 0.05. (**C**, **D** and **E**) The results of GSEA. (**F**) The comparison of immune cell infiltration between low- and high- risk score groups.

Figure 6

Protein-protein interaction (PPI) network. (**A**) PPI network of DEGs between low- and high- risk score groups. (**B**) PPI network processed by Cytoscape; red: ARGs that expressed highly in the high risk score group, blue: ARGs expressed highly in the low risk score group. (**C**) Top 10 hub genes selected by cytoHubba.

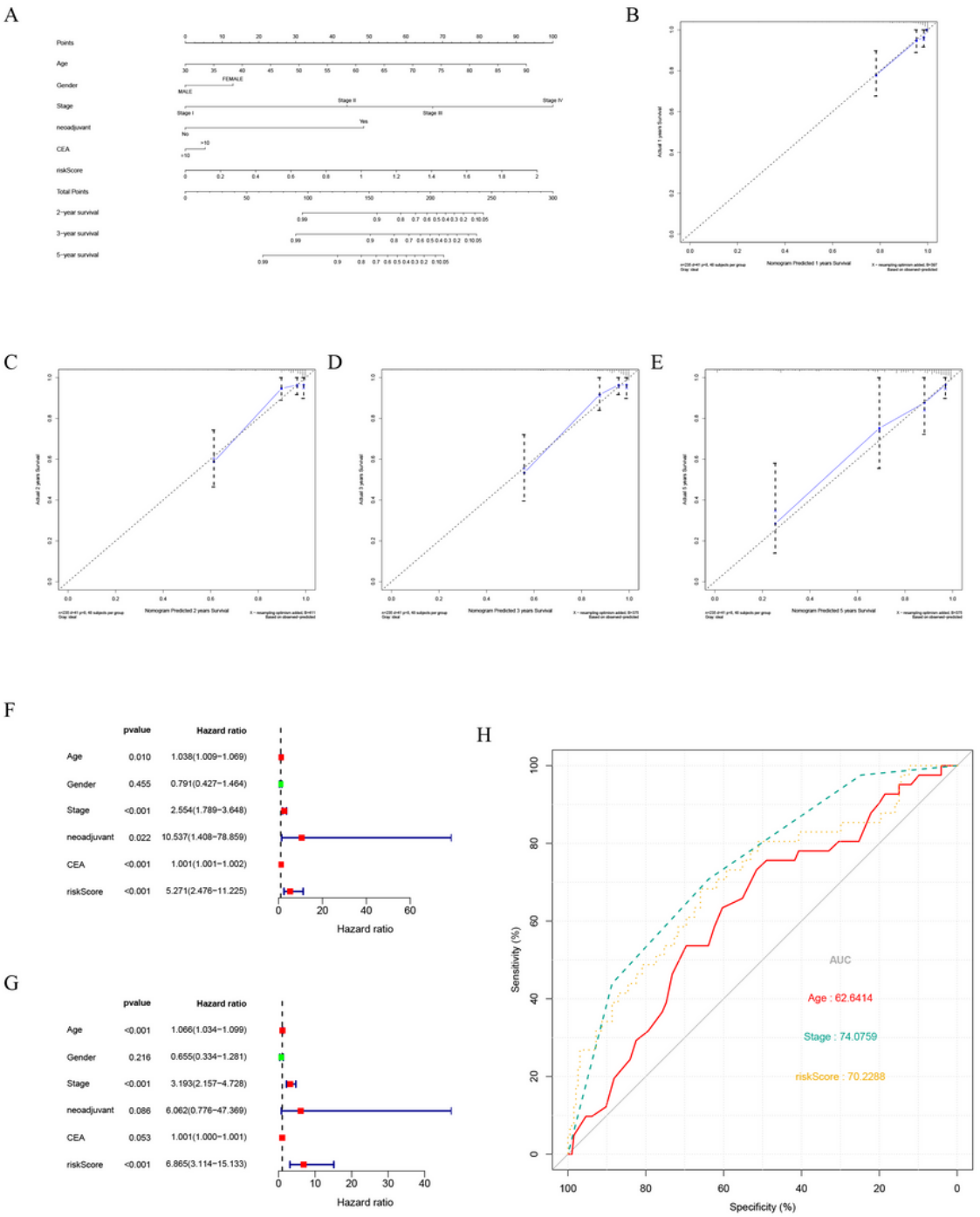

Figure 7

Development of a nomogram. (**A**) Nomogram predicting OS of patients in the TCGA cohort. (**B**, **C**, **D** and **E**) The calibration plots of the nomogram prediction for 1-, 2-, 3- and 5-year survival, respectively. The x-axis is the nomogram-predicted survival and the y-axis is actual survival. (**F**) Univariate Cox regression analysis of nomogram. (**G**) Multivariate Cox regression analysis of nomogram. (**H**) ROC curves of independent prognostic indicators.

Figure 8

Prediction of drugs for CRC treatment. (**A**) Top six drugs with positive and negative coefficients predicted in the CTRP database. (**B**) Top six drugs with positive and negative coefficients predicted in the PRISM database.

Supplementary Files

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

- SupplementFigure1.jpg
- SupplementFigure2.jpg
- SupplementFigure3.jpg
- SupplementFigure4.jpg
- SupplementFigure5.jpg