Development and Validation of a Novel 13-Genes Prognostic Model for Colon Cancer

Duo Yun

Inner Mongolia Yuanhe Mongolian Traditional Medicine Research Institute https://orcid.org/0000-0001-7999-5626

Zhirong Yang (doctor.yang@cumt.edu.cn)

Fujian Province The Fujian Pinghe County People's Hospital https://orcid.org/0000-0003-0045-5497

Research article

Keywords: colon cancer, survival, biomarker, prognostic genes

Posted Date: April 3rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-378434/v1

Abstract

Colon cancer is one of the most common malignant tumors in the world. The purpose of this study is to explore the prognostic value of genes in colon cancer. After analyzing gene expression profiles, differential expressed genes between 39 normal tissues and 398 tumor tissues were identified from The Cancer Genome Atlas database. We use Cox and lasso regression to find genes related to prognosis. Through analysis, 13 genes were found to predict the overall survival of colon cancer patients. In addition, the external comparing of gene expression and the single prognostic gene survival analysis were made. Finally, pathway enrichment and mutation status of each gene were also analyzed. After a series of bioinformatics analysis, we select 13 survival-related signature and established a prognostic risk model based on these genes. The prognostic risk model was developed to comprehensively predict the overall survival of colon cancer patients. The prognostic value of the 13-genes (CLDN23,HAND1,IL23A,KLHL35,SIX2,UPK2,HOXC11,KRT6B,SRCIN1,TNNI3,TYRO3,MIR6835,LINC02474) related risk score for each colon cancer patent was calculated to predict the survival. Furthermore, five genes (SIX2 MIR6835 LINC02474 CLDN23 HOXC11) were significantly associated with overall survival (OS). The KEGG pathway enrichment results suggested that most of the pathways are related to the occurrence, metabolism, proliferation and invasion of the tumor cells. It was found that the expression of 13-genes signature can be used as prognostic indicator for colon cancer patients. The 13-genes signature predictive model may help clinicians provide a prognosis and personalized treatment for colon cancer patients.

1. Introduction

Colon cancer is the main malignant tumor of the digestive tract and ranks the third among malignant tumors in terms of global incidence[1–3].Although the morbidity and mortality associated with colon cancer have gradually decreased in recent years, colon cancer is still the most common cause of cancer related death, and long-term survival rates have hardly improved[4–6]. The accurate prognosis of colon cancer patients is essential to provide effective personalized treatment. According to some reports, the prognosis of colon cancer patients is mainly determined by clinicopathological characteristics and tumor stage. However, due to the heterogeneity of the disease, it is difficult to determine the prognosis of patients only based on these traditional risk factors[7, 8].Therefore, in order to improve the prognosis of colon cancer patients, it is very important to better understand the pathogenesis of the disease and discover some prognosis-related biomarkers. However, nowadays most studies on the survival rate of colorectal cancer patients do not separately analyze the patients with colon and rectal cancer[9, 10].The prognosis of these two types of malignant tumors, metastatic potential and treatment methods are different[11, 12]. Thus, in this study, we built a prognostic model containing 13 genes applying Cox and the LASSO model (Least Absolute Shrinkage and Selection Operator), on which only focused colon cancer patients in TCGA-COAD dataset. We aimed to construct a highly reliable risk prediction prognostic model for colon cancer patients.

2. Material And Methods

2.1 Collection and processing of expression profile data

The gene expression data and clinical phenotype information of all colon cancer samples were downloaded from The Cancer Genome Atlas (TCGA) on September 3rd, 2020. The TCGA colon gene data from 437 samples include 398 tumor tissues and 39 normal tissues. In addition, the clinical data of 437 patients were also downloaded, including age, gender, tumor histological grade, survival time and survival status. Data sorting and ID transferring were done through Perl project.

2.2 Differential gene analysis

The Wilcox test method was used to analyze the difference between COAD samples and normal samples in the gene expression profile by limma package with R project (version 3.40.0). |Log2

fold change (FC)|≥1 and false discovery rate (FDR) < 0.05 were used to identify the differentially expressed genes (DEGs).

2.3 Establishment of genes prognostic risk model

All samples are randomly divided into training set and testing set, for the training set, the Cox regression analysis in R survival package (version 2.44–1.1) was used to analyze the regression coefficient and FDR value of each gene in relation to survival time. The genes with FDR value < 0.05 were initially considered as the genes related to prognosis. Subsequently, lasso regression analysis with glmnet and survival package was used to further reduce selected genes. In order to build the cox prognosis model, the calculation of risk score for each patient was defined as follows: Risk score = β gene1*expression (gene 1) + β gene2 * expression (gene 2) + ... + β gene n * expression (gene n), Where, β is the prognostic correlation coefficient beta estimated by Cox analysis which equals to log (Hazard Ratio), and expression represents the expression value of corresponding gene.

According to the risk score, the training set and the testing set patients were divided into high risk and low risk subgroups according to the median value of risk core. Then Kaplan-Meier survival curve analysis was performed to calculate the difference in survival prognosis time between samples of high risk and low risk group using the R survival package. Receiver operating characteristic (ROC) curves were constructed by using the R survival ROC package(version 1.0.3). The Area Under the Curve (AUC) was calculated to assess the predictive value of the models.

2.4 Pathway enrichment analysis

The Gene set enrichment analysis (GSEA) (http://software.broadinstitute.org/gsea/) was carried out to reveal the signaling pathways between high and low expression of each single gene based on TCGA-COAD dataset. Each single gene set"c2.cp.kegg.v7.0.symbols.gmt (Curated)" was utilized for GSEA. Of note, 1,000 permutations were used in the analysis of each parameter to calculate the enrichment scores. The P value < 0.05 and normalized enrichment score (NES) were implemented to determine vital

enrichment pathways among the subgroups. The multiple GESA pathway graph of each gene was constructed using the ggplot2, plyr, grid and gridExtra packages in R project.

2.5 Statistical analysis

R version 3.5.3 (http://www.R-project.org), was applied for statistical analysis and graph plotting. A 2-sided P-value < 0.05 was determined to be statistically significant for Kaplan-Meier survival analysis. The Area Under the ROC Curve (AUC) was calculated to evaluate the diagnosis performance of the identified signature. Univariate and multi-variate Cox regression analyses were conducted to determine the prognostic performance of risk factors.

3. Results

3.1 Identification of differential genes between normal tissues and colon cancer

In the TCGA-COAD dataset, there are 398 colon tumor samples and 39 normal samples. The gene differential expression profile of normal tissues and colon cancer of patients were calculated using the R project. A total of 1593 genes were identified as DEGs with FDR value < 0.05 and $|\log_2 FC| > 2$.

3.2 Construction of a 13-genes signature prognostic model in the training set

Among the 1593 genes, the DEGs in each patient was displayed. The patients were randomly divided into training set and testing set. According to clinical information, we combined the survival time of the training set with DEGs expression to analyze the genes that may affect the patients' clinical prognosis. Through Cox regression analysis of the training set, the single factor significance was defined as P value < 0.01, and prognosis-related genes were selected. The hazard ratio (HR) and the 95% confidence interval (CI) of each variable were shown in Table 1.There were a total of 48 genes related to prognosis. The candidate genes were further reduced to 15 genes through the lasso regression method to reduce the overfitting of the model (Fig. 1). Finally, through the following stepwise variable selection procedure, we established an optimal Cox proportional hazard model containing 13 key genes: MIR6835(microRNA 6835),TYRO3(TYRO3 protein tyrosine kinase),SRCIN1(SRC kinase signaling inhibitor 1),HOXC11(homeobox C11),CLDN23(claudin 23),UPK2(uroplakin 2),TNNI3(troponin I3),KLHL35(kelch like family member 35),KRT6B(keratin 6B),HAND1(heart and neural crest derivatives expressed 1),IL23A(interleukin 23 subunit alpha),LINC02474(long intergenic non-protein coding RNA 2474),SIX2(SIX homeobox 2)(Table 2). Thepredicted risk score can be calculated with the formula: 0.325257866*MIR6835 + 0.367503693*TYRO3 + 0.418341346*SRCIN1 + 0.134589791*HOXC11 + 0.123203974*CLDN23 + 0.205044971*UPK2 + 0.236533736*TNNI3 + 0.15943496*KLHL35 + 0.009257475*KRT6B + 0.083391046*HAND1 + 0.09208856*IL23A + 0.255146739*LINC02474 + 0.079143843*SIX2. The forest graph of risk genes was shown (Fig. 2A).The samples of the training set were separated into high-risk and low-risk subgroups according to the median value of risk scores. In

order to evaluate the diagnosis prediction value of the model, the time-dependent survival ROC curves were also established. At the training set, the predictive capability of the genes remains well, with an AUC of 0.823(Fig. 2B).The Kaplan-Meier survival curve also showed that the survival rate of the high-risk group was significantly reduced ($P < 0.01$,Fig. 2C).The risk distribution plot and heat map of gene expression in the training set are displayed (Fig. 2D–2F).

Table 1
The hazard ratio (HR) and the 95% confidence interval (CI) of 48 genes related to prognosis

Id	HR	HR.95L	HR.95H	P value
KLHL35	1.126807	1.071441	1.185034	3.41E-06
FOXD1	1.189012	1.095268	1.290778	3.60E-05
TYRO3	1.577604	1.243019	2.00225	0.000178
FGF19	1.092626	1.042115	1.145587	0.000244
HOXC11	1.290783	1.122404	1.484421	0.000345
SIX2	1.129378	1.056494	1.20729	0.000351
ZIC5	1.441737	1.174138	1.770324	0.000479
MIR6835	1.367411	1.144605	1.633587	0.000564
FXYD4	1.103671	1.041967	1.169029	0.000778
SCG2	1.144913	1.057974	1.238996	0.000783
PKIB	1.09021	1.035758	1.147526	0.000954
KREMEN2	1.316959	1.117901	1.551461	0.000991
CCDC78	1.310785	1.112871	1.543897	0.001193
IL23A	1.072684	1.028007	1.119302	0.001227
HAND1	1.064342	1.024794	1.105416	0.001248
UPK2	1.298285	1.107812	1.521508	0.001261
OTX1	1.62325	1.205182	2.186342	0.001431
LINC02474	1.290539	1.098895	1.515606	0.001872
FEZF1-AS1	1.219718	1.07338	1.386006	0.00232
LINC00941	1.481484	1.15003	1.908467	0.002352
AC022144.1	1.281687	1.09177	1.504641	0.002422
KRT6B	1.010332	1.003592	1.017117	0.002614
CXCL5	1.004519	1.001571	1.007475	0.002642
SLC7A11	1.122323	1.040905	1.210109	0.00267
KCNJ14	3.356977	1.513859	7.444087	0.002878
PTMAP4	1.019718	1.006701	1.032904	0.002894

Id	HR	HR.95L	HR.95H	P value
ERFE	1.2022	1.064423	1.357811	0.003024
TNNI3	1.350007	1.105976	1.647882	0.003176
NTN1	1.437744	1.127424	1.83348	0.003425
TP73	1.416373	1.11812	1.794183	0.003909
CLDN23	0.869051	0.789565	0.956539	0.004132
GNG8	1.095393	1.028951	1.166126	0.004319
TLX1	1.381873	1.098554	1.738259	0.005729
AC092112.1	1.085882	1.024198	1.151281	0.005758
SCARNA22	0.306391	0.131481	0.713983	0.006135
MNX1-AS1	1.15757	1.042099	1.285837	0.006351
PRAME	1.100824	1.027258	1.179658	0.006488
AC004080.1	1.096815	1.025262	1.173362	0.007258
LRMP	1.262019	1.064141	1.496693	0.007486
MDFI	1.093243	1.023913	1.167267	0.007655
NPTX1	1.232221	1.056334	1.437396	0.007875
AJUBA	1.178079	1.042487	1.331306	0.008616
SRCIN1	1.387746	1.084955	1.77504	0.009075
MYC	0.991024	0.984326	0.997768	0.009164
AC037487.3	1.433162	1.093202	1.878842	0.009188
VGF	1.023501	1.005659	1.04166	0.009628
TNNT1	1.052653	1.012541	1.094354	0.009634

Table 2
13 key genes were established by Cox proportional hazard model

id	coef	HR	HR.95L	HR.95H
MIR6835	0.325258	1.384388	1.109466	1.727434
TYRO3	0.367504	1.444125	1.053595	1.979411
SRCIN1	0.418341	1.519439	1.082098	2.133536
HOXC11	0.13459	1.144067	0.963941	1.357853
CLDN23	-0.1232	0.884083	0.786796	0.993401
UPK2	0.205045	1.22758	0.991952	1.51918
TNNI3	0.236534	1.26685	1.007453	1.593037
KLHL35	0.159435	1.172848	1.099408	1.251194
KRT6B	0.009257	1.0093	1.001587	1.017073
HAND1	0.083391	1.086967	1.03609	1.140342
IL23A	0.092089	1.096462	1.041242	1.15461
LINC02474	0.255147	1.290651	1.044759	1.594415
SIX2	0.079144	1.08236	1.005181	1.165465

3.3 External validation in testing set

The signature's predictive performance was further validated in the testing set. Using the same formula established in the training set, the risk scores of each patient in the testing set was calculated based on the relative expression levels of 13 genes. In the testing set, the AUC of the 13-gene signature used to predict patient survival was 0.702 (Fig. 3A). The Kaplan-Meier curve generated from the testing set showed that the survival outcome of low risk group patients was significantly better than that of high risk group patients(Fig. 3B), which indicated that it has moderate diagnostic performance. The risk distribution plot and heat map of gene expression in the testing set were displayed(Fig. 3C-E). These results suggested that the prognostic model has better sensitivity and specificity.

3.4 External comparing of gene expression in Oncomine datasets

However, HOXC11, KRT6B, SRCIN1, TNNI3, TYRO3 were relatively highly expressed in COAD tissue. Besides, the protein expression of CLDN23, HAND1, IL23A, KLHL35, SIX2, UPK2 was not significantly different between COAD tissue and normal lung tissue (Fig. 4).

3.5 Survival analysis of the prognostic 13 candidate genes signature in COAD

In order to further investigate the specific relationship between the 13 individual genes with OS in colon cancer patients, a comprehensive survival analysis was performed using the Kaplan-Meier method. The results showed that five genes (SIX2 MIR6835 LINC02474 CLDN23 HOXC11) were significantly associated with OS, while the other candidate genes were no significantly correlated with OS in COAD (Fig. 5).

3.6 Pathway enrichment analysis

To further investigate the correlated pathway of each candidate gene in COAD, KEGG pathway enrichment analysis of these genes was obtained with GSEA. The upregulated and downregulated pathways of each single gene were shown in GESA multiple pathway graph (Fig. 6).The GESA pathway results suggested that most pathways are related to the occurrence, metabolism, proliferation and invasion of the tumor cells.

3.7 Identifying the mutation status of candidate genes in colon cancer

The genetic alterations of 13 candidate genes were examined to determine their role in colon cancer genes in GEPIA database. Among the 110 COAD patients, a total of 30 (27.27%) patients had changes (cbioportalCPTAC-2 Prospective, Cell 2019 cohort) (Fig. 7). The alteration rates of CLDN23, TYRO3 and SRCIN1 in this dataset were 9%, 9% and 10% respectively. Frequent genetic alterations indicated that these genes play a vital role in the development of colon cancer.

4. Discussion

In this study, we identified a total of 13 genes as signature genes that is associated with the patients' survival from the TCGA colon cancer datasets through the training set. We further verified that these findings were consistent in the testing set. The prognostic risk scores of colon cancer patients for precise predictions was important. In the prognostic model, our results showed that patients' survival with high risk scores was decreased, compared to that with low risk scores.

Among the 13 genes that built the model, much of them are associated with cancer. For example, CLDN23 is down-regulated in colorectal cancer tumors, and its level down-regulation is related to the prognosis of colorectal cancer patients[13].Changes in UPKs levels in urothelial carcinoma are considered to be useful markers for diagnosis, detection and prediction of urothelial carcinoma[14, 15].TYRO3 plays

a role in promoting the survival and growth of cancer cells[16] and the knockdown of TYRO3 by siRNA prevents melanoma cell migration and invasion[17].Knockdown of KHLH35 in HEK293 cells promoted anchorage-independent growth, indicating that it may play a role in tumorigenesis[18, 19]. HAND1 gene has been found to be silent and hypermethylated in human gastric cancer[20],pancreatic cancer[21] and ovarian carcinomas[22].A study shows that SIX2 can promote breast cancer metastasis[23].IL-23 may be related to the progression of bladder cancer[24, 25].The KRT6B is associated with the increased risk of lung cancer[26, 27],breast carcinoma[28], and urothelial cancer[29]. When expressed outside the context, HOXC11 can abnormally promote proliferation, thereby promoting tumorigenesis[30].SRCIN1 inhibited the invasion of highly metastatic breast cancer cells by inhibiting corticosteroid-dependent cell movement[20, 25, 31].For the other three genes (MIR6835, LINC02474,TNNI3) among the predicted scoring model, their roles in human cancers have not yet been fully investigated. Strength of this study is that we analyzed KEGG pathway enrichment with GSEA and identified mutation status for each candidate gene in GEPIA database. We also made the external comparing of gene expression in Oncomine datasets and the OS of each gene based on TCGA-COAD dataset. However, there also are some limitations in this research. First, the differential expression of 13 genes was only identified from the TCGA colon dataset, and experimental verification was lacking. Although TCGA's DNA-seq data is of high quality, it still needs further experimental verification. In addition, further studies on the functions of these 13 genes in colon cancer are needed in vitro and in vivo.

5. Conclusion

In summary, our study revealed a 13 genes signature related to the prognosis of colon cancer patients. Besides, the 13genes predictive model may be useful for survival prediction and diagnosis in colon cancer patients, which may increase the effectiveness of personal treatment for colon cancer patients. Moreover, some gene signature displayed vital biological function, which can potentially be used for further biological research. Pre-clinical studies followed by clinical trials are needed to validate our findings in the future.

Declarations

Acknowledgement

We are grateful to the contributors of data to public databases including TCGA, Oncomine and cBioPortal database.

Authors’ contributions

Zhirong Yang designed and revised the current study. Duo Yun performed the analyses, calculations and wrote the first manuscript. All authors read and approved the final manuscript.

Funding

No funding.

Availability of data and materials

The datasets generated and/or analyzed during the current study were TCGA(https://portal.gdc.cancer.gov/).

Ethics approval and consent to participate

Not applicable. This article does not contain either human or animal experiments.

Informed consent

Written consent is not required for the current study.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Freeman HJJWJoGW: **Early stage colon cancer**. 2013(46):25-30.
2. Siegel R, DeSantis C, Virgo K, Stein K, Mariotto A, Smith T, Cooper D, Gansler T, Lerro C, Fedewa S *et al*: **Cancer treatment and survivorship statistics, 2012**. *CA: a cancer journal for clinicians* 2012, **62**(4):220-241.
3. Wei X, Zhang Y, Yang Z, Sha Y, Pan Y, Chen Y, Cai LJBR: **Analysis of the role of the interleukins in colon cancer**. 2020, **53**.
4. Daniel, Longley, and, Wendy, L., Allen, and, Patrick, Cancer GJBEBARo: **Drug resistance, predictive markers and pharmacogenomics in colorectal cancer**. 2006.
5. McLornan DP, Barrett HL, Cummins R, McDermott U, McDowell C, Conlon SJ, Coyle VM, Van Schaeybroeck S, Wilson R, Kay EWJCCRAOJotAAfCR: **Prognostic Significance of TRAIL Signaling Molecules in Stage II and III Colorectal Cancer**. 2010, **16**(13):3442.
6. Xiaona, Li, Fei, Han, Wenbin, Liu, Xiaoyan, Cancer SJBD: **PTBP1 promotes tumorigenesis by regulating apoptosis and cell cycle in colon cancer**. 2018.
7. Yang H, Liu H, Lin HC, Gan D, Wang ZJA: **Association of a novel seven-gene expression signature with the disease prognosis in colon cancer patients**. 2019, **11**(19):8710-8727.
8. Bert, Vogelstein, Nickolas, Papadopoulos, Victor, E., Velculescu, Shibin, Zhou, Science KJ: **Cancer Genome Landscapes**. 2013.

9. Ukegjini K, Zadnikar M, Warschkow R, Müller S, Schmied BM, Marti LJLAoS: **Baseline mortality-adjusted survival in colon cancer patients**. 2016.
10. Coleman MP, Forman D, Bryant H, Butler J, Rachet B, Maringe C, Nur U, Tracey E, Coory M, Hatcher J *et al*: **Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995-2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data**. *Lancet (London, England)* 2011, **377**(9760):127-138.
11. Mitry E, Guiu B, Cosconea S, Jooste V, Faivre J, Bouvier AMJG: **Epidemiology, management and prognosis of colorectal cancer with lung metastases: a 30-year population-based study**. 2010, **59**(10):1383-1388.
12. Iversen LH, Nrgaard M, Jepsen P, Jacobsen J, Christensen MM, Gandrup P, Madsen MR, Laurberg S, Wogelius P, Disease HTSJC: **Trends in colorectal cancer survival in northern Denmark: 1985–2004**. 2007, **9**(3).
13. Pitule P, Vycital O, Bruha J, Novak P, Liska VJAR: **Differential expression and prognostic role of selected genes in colorectal cancer patients**. 2013, **33**(11):4855-4865.
14. Munipalli SB, Yenugu SJG, Endocrinology C: **Uroplakin expression in the Male Reproductive Tract of Rat**. 2019, **281**:153-163.
15. Wu XR, Kong XP, Pellicer A, Kreibich G, Sun TTJKI: **Uroplakins in urothelial biology, function, and disease**. 2009, **75**(11):1153.
16. Crosier PS, Hall LR, Vitas MR, Lewis PM, Crosier KEJLL: **Identification of a novel receptor tyrosine kinase expressed in acute myeloid leukemic blasts**. 1995, **18**(5-6):443-449.
17. Shao H, Wang A, Lauffenburger DA, Wells AJIJoB, Biology C: **Tyro3-mediated phosphorylation of ACTN4 at tyrosines is FAK-dependent and decreases susceptibility to cleavage by m-Calpain**. 2018, **95**:73-84.
18. **Genome-wide analysis of DNA methylation identifies novel cancer-related genes in hepatocellular carcinoma %J Tumour Biol**. 2012, **33**(5):1307-1317.
19. Morris MR, Ricketts CJ, Gentle D, Mcronald F, Maher ERJO: **Genome-wide methylation analysis identifies epigenetically inactivated candidate tumour suppressor genes in renal cell carcinoma**. 2011, **30**(12):1390-1401.
20. Kaneda A, Kaminishi M, Yanagihara K, Sugimura T, Ushijima TJCR: **Identification of Silencing of Nine Genes in Human Gastric Cancers**. 2002, **62**(22):6645.
21. Hagihara, Atsushi, Miyamoto, Kazuaki, Furuta, Junichi, Hiraoka, Nobuyoshi, Wakazono, Oncogene KJ: **Identification of 27 5'CpG islands aberrantly methylated and 13 genes silenced in human pancreatic cancers**. 2004, **23**(53):8705-8710.
22. Takada T, Yagi Y, Maekita T, Imura M, Nakagawa S, Tsao S, Miyamoto K, Yoshino O, Yasugi T, Taketani YJCS: **Methylation-associated silencing of the Wnt antagonist SFRP1 gene in human ovarian cancers**. 2010, **95**(9):741-744.
23. Wang CA, Drasin D, Pham C, Jedlicka P, Zaberezhnyy V, Guney M, Li H, Nemenoff R, Costello JC, Tan ACJCR: **Homeoprotein Six2 promotes breast cancer metastasis via transcriptional and epigenetic**

control of E-cadherin expression. 2014, **74**(24):7357-7370.

24. Peng, Wang, Yao, Zhang, Jun, Jiang, Fengshuo, Jin, Zhongyi, Reports SJO: **IL?23 concentration? dependently regulates T24 cell proliferation, migration and invasion and is associated with prognosis in patients with bladder cancer**. 2018.
25. Wang P, Wang H, Li X, Liu Y, Zhu DJPO: **SRCIN1 Suppressed Osteosarcoma Cell Proliferation and Invasion**. 2016, **11**(8):e0155518.
26. Camilo R, Capelozzi VL, Siqueira SAC, Bernardi FDCJHP: **Expression of p63, keratin 5/6, keratin 7, and surfactant-A in non-small cell lung carcinomas**. 2006, **37**(5):542-546.
27. Zhang H, Huo M, Jia Y, Xu AJIJoC, Medicine E: **KRT6B, a key mediator of notch signaling in honokiol-induced human hepatoma cell apoptosis**. 2015, **8**(9):16880.
28. Brogi E, Murphy CG, Johnson ML, Conlin AK, Hsu M, Patil S, Akram M, Nehhozina T, Jhaveri KL, Hudis CAJAoO: **Breast carcinoma with brain metastases: clinical analysis and immunoprofile on tissue microarrays**. 2011, **22**(12):2597-2603.
29. Fichtenbaum EJ, Marsh WL, Zynger DLJAJoCP: **CK5, CK5/6, and Double-Stains CK7/CK5 and p53/CK5 Discriminate In Situ vs Invasive Urothelial Cancer in the Prostate**. 2012, **138**(2):190-197.
30. Eda AG, Taylor HSJBoR: **HOXC and HOXD Gene Expression in Human Endometrium: Lack of Redundancy with HOXA Paralogs1**. 2004(1):1.
31. Damiano L, Dévédec SEL, Stefano PD, Repetto D, Lalai R, Truong H, Xiong JL, Danen EH, Yan K, Verbeek FJJO: **p140Cap suppresses the invasive properties of highly metastatic MTLn3-EGFR cells via impaired cortactin phosphorylation**. 2012.

Figures

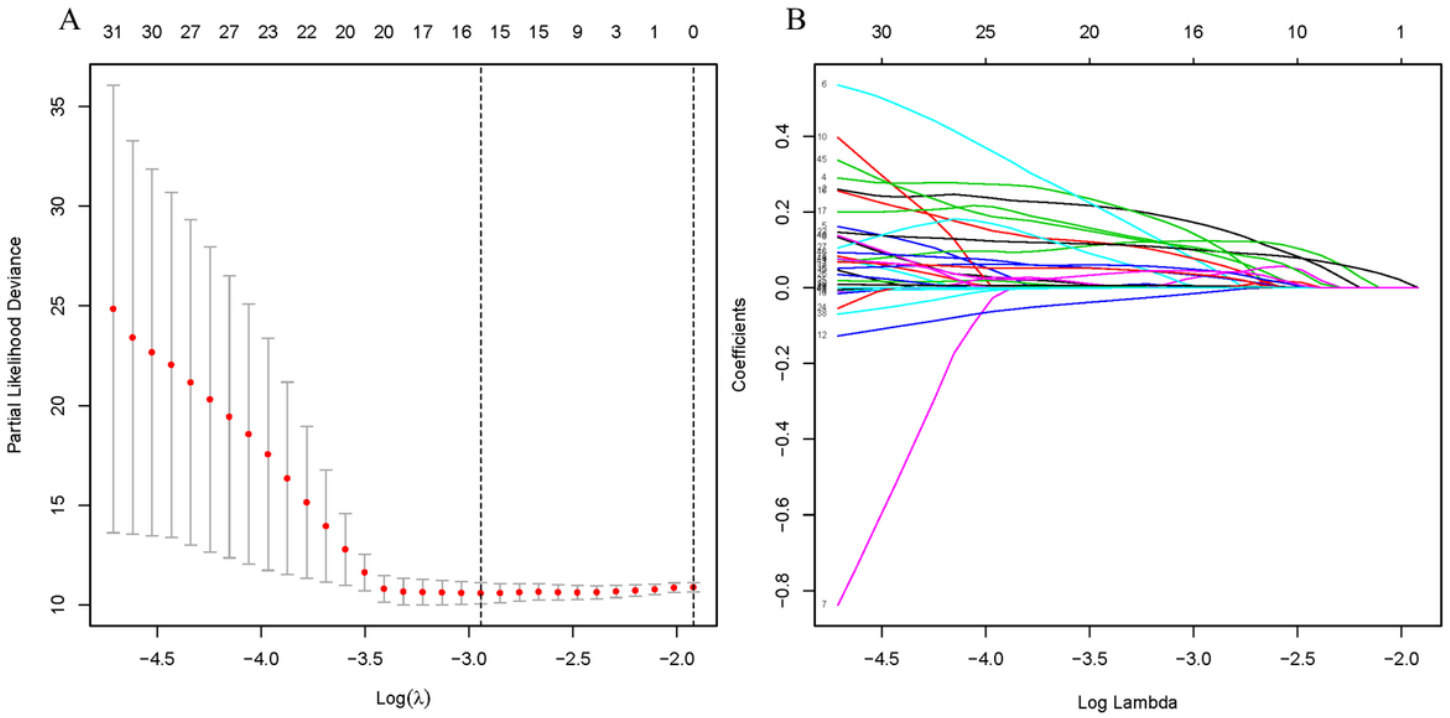

Figure 1

The Lasso regression reduce the 48 prognostic related genes to 15 prognostic related genes. (A) A coefficient profile plot was generated against the log (lambda) sequence. Selection of the optimal parameter (lambda) in the LASSO model;(B) LASSO coefficient profiles of the 15 candidates in the training set.

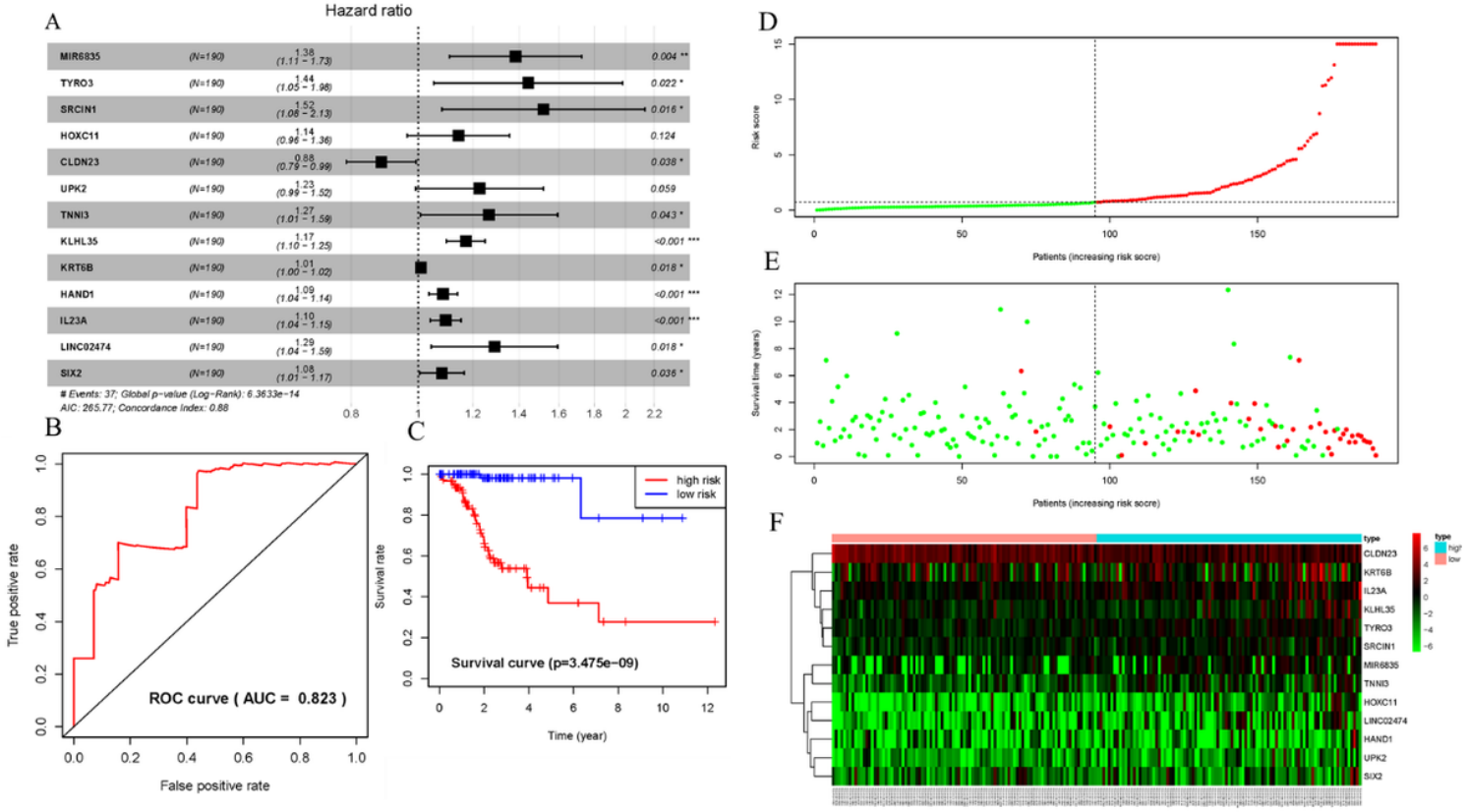

Figure 2

Predicted performance of the identified prognostic signature in the training set. (A)The forest graph of 13 prognostic genes, the hazard ratio and P value were shown.(B) ROC analysis showed that the model had a very high ability to discriminate controls from COAD samples validated in TCGA(AUC=0.823).(C) The Kaplan-Meier survival curve verifies the accuracy of the model in predicting the OS rates of COAD in the training set.(D)The distribution plot of risk scores. Patients are arranged from left to right in an increasing order of risk scores. (E)The survival status of each patient: The Y-axis represents the overall survival time. The color code: green for alive case, red for dead case. (F) The heatmap of training set's gene expression levels of the 13 genes.

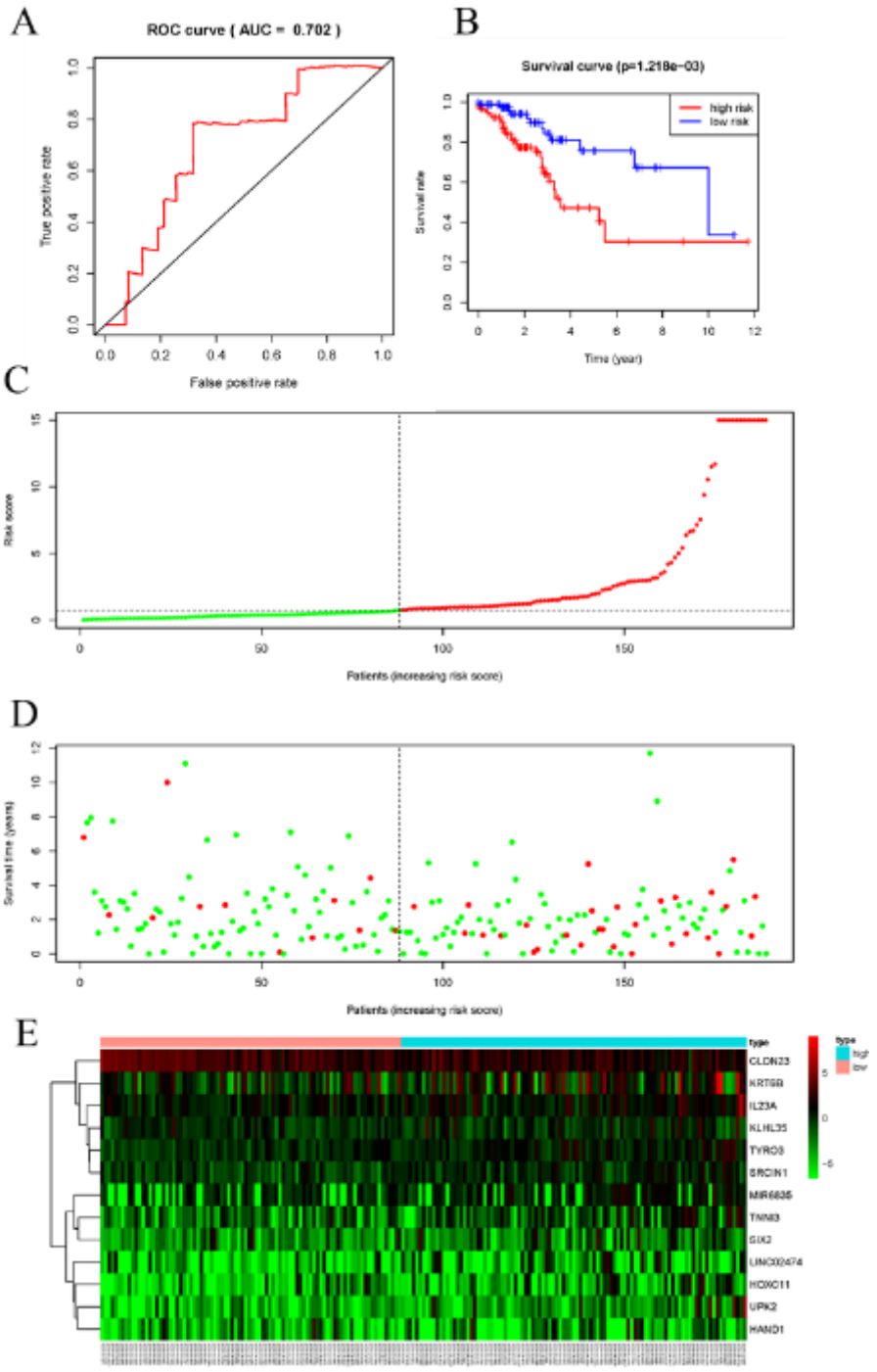

Figure 3

Predicted performance of the identified prognostic signature in the testing set. (A)The ROC curve for colon cancer patient relapse by the 13 genes. The AUC supports that the 13 genes best predicts in the training set.(B)The Kaplan-Meier survival analysis of the 13 prognostic genes in the TCGA colon cancer cohort; According to the median value, samples are divided into high-risk and low-risk group. (C)The distribution plot of risk score. Patients are arranged from left to right in an increasing order of risk score. (D)The survival status of each patient: The Y-axis represents the overall survival time. The color code: green for alive case, red for dead case.(E) The heatmap of training set's gene expression levels of the 13 genes.

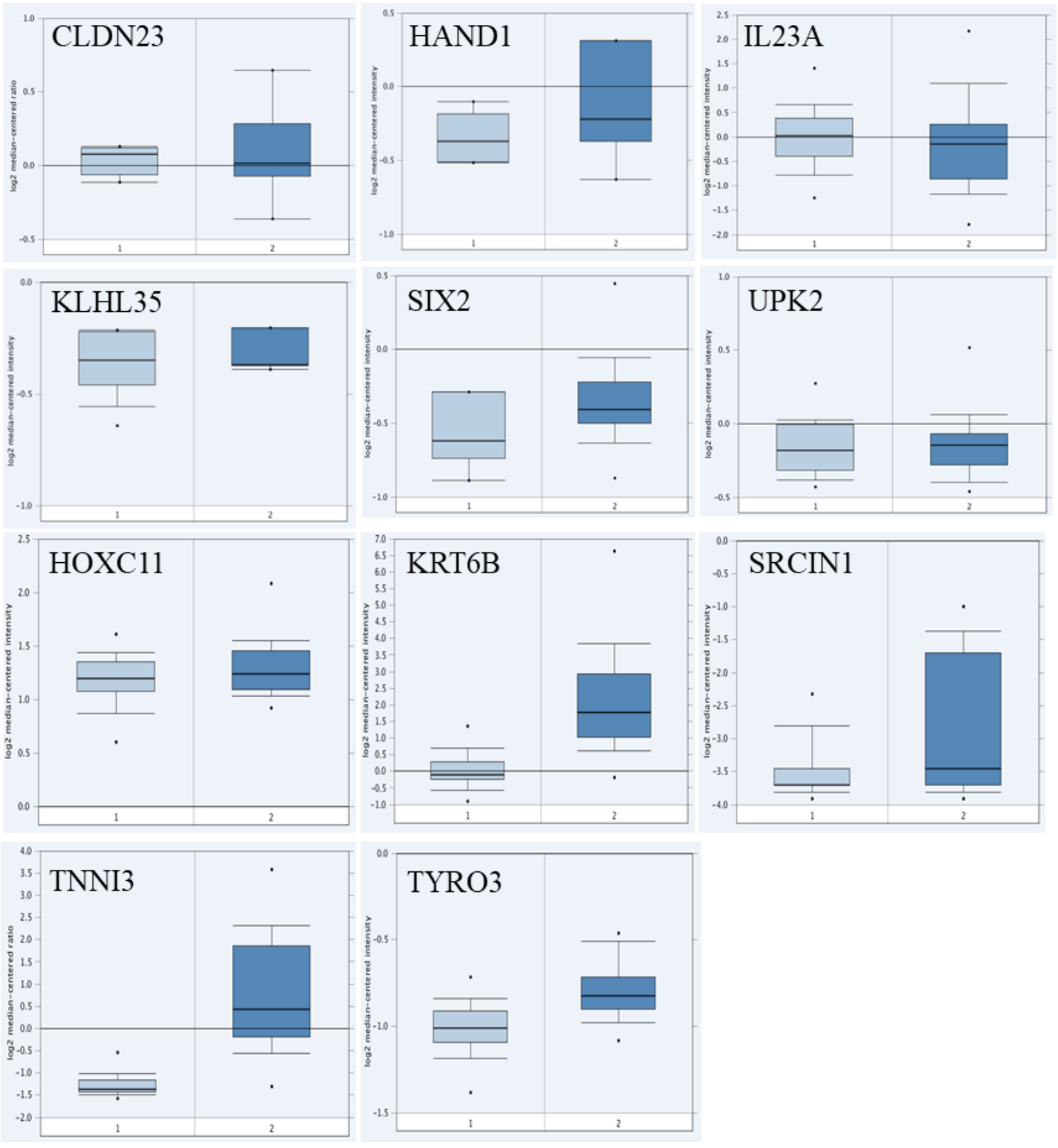

Figure 4

Verification of 11 candidate genes expression in COAD and normal colon tissue using the Oncomine database.

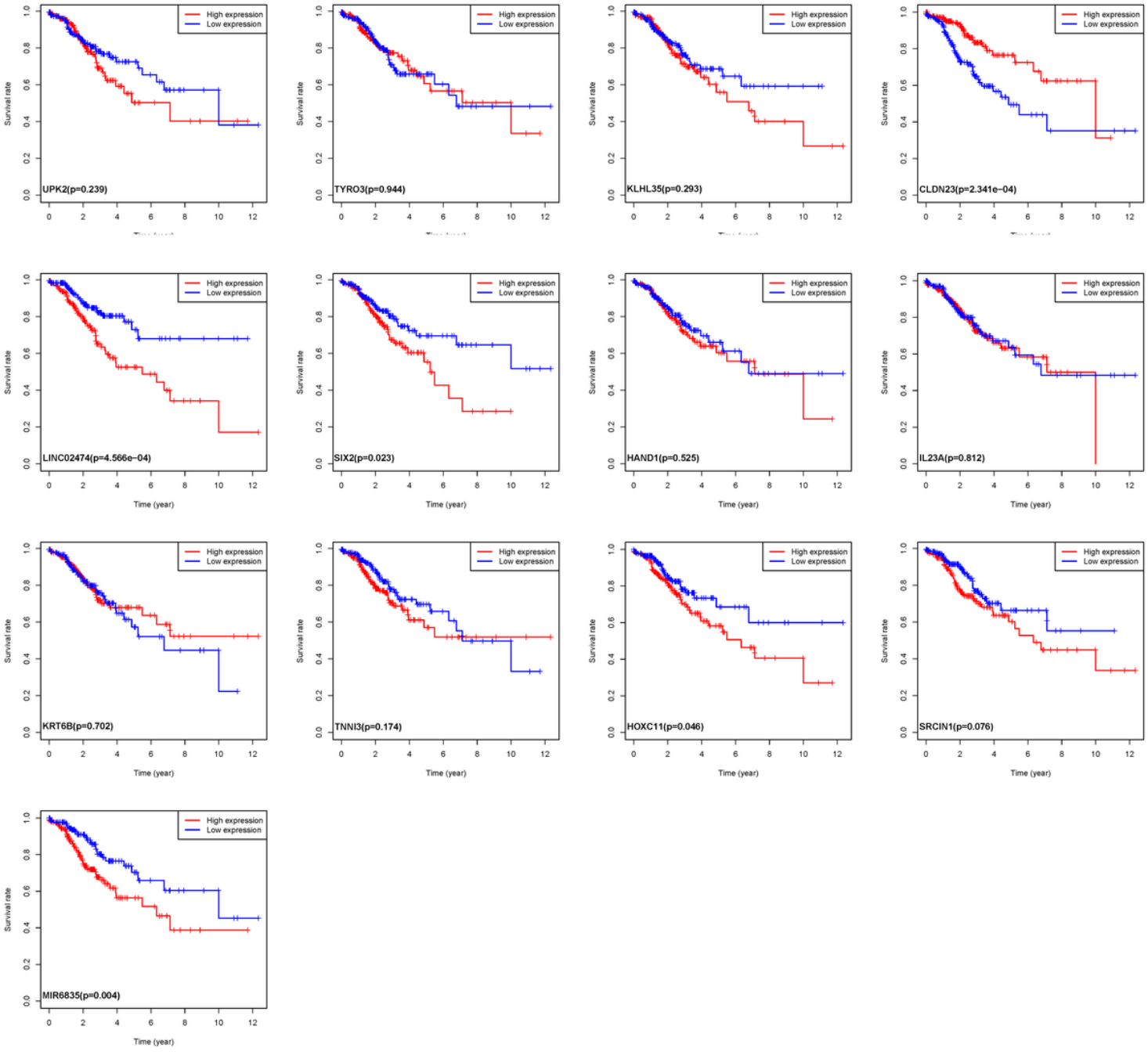

Figure 5

The comprehensive survival with clinical characteristics regarding the OS of the 13 individual genes of the colon cancer patients were analyzed by the Kaplan-Meier method.

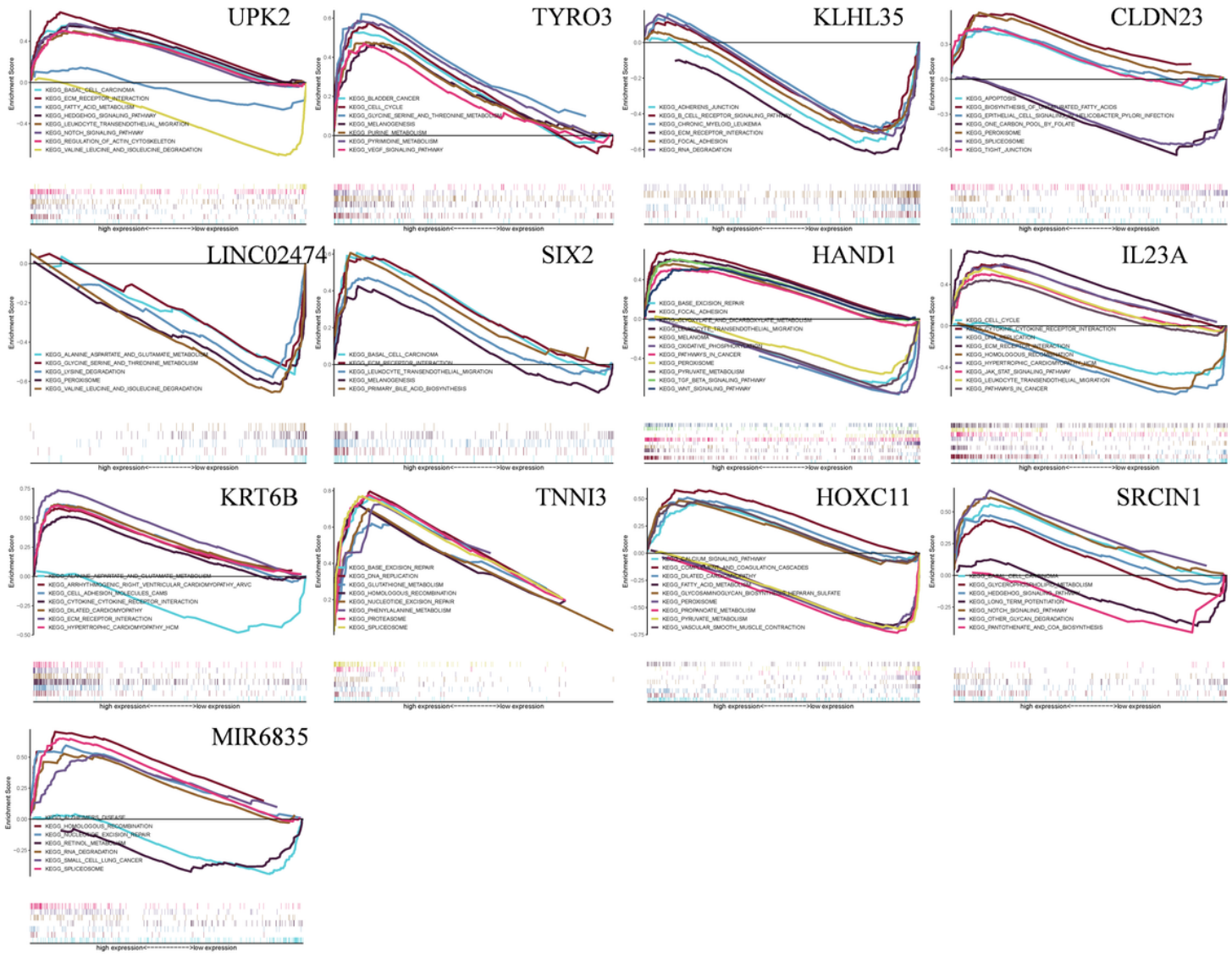

Figure 6

The KEGG pathway enrichment analysis of the TCGA colon cancer patients to each signature gene obtained above by GSEA signaling.

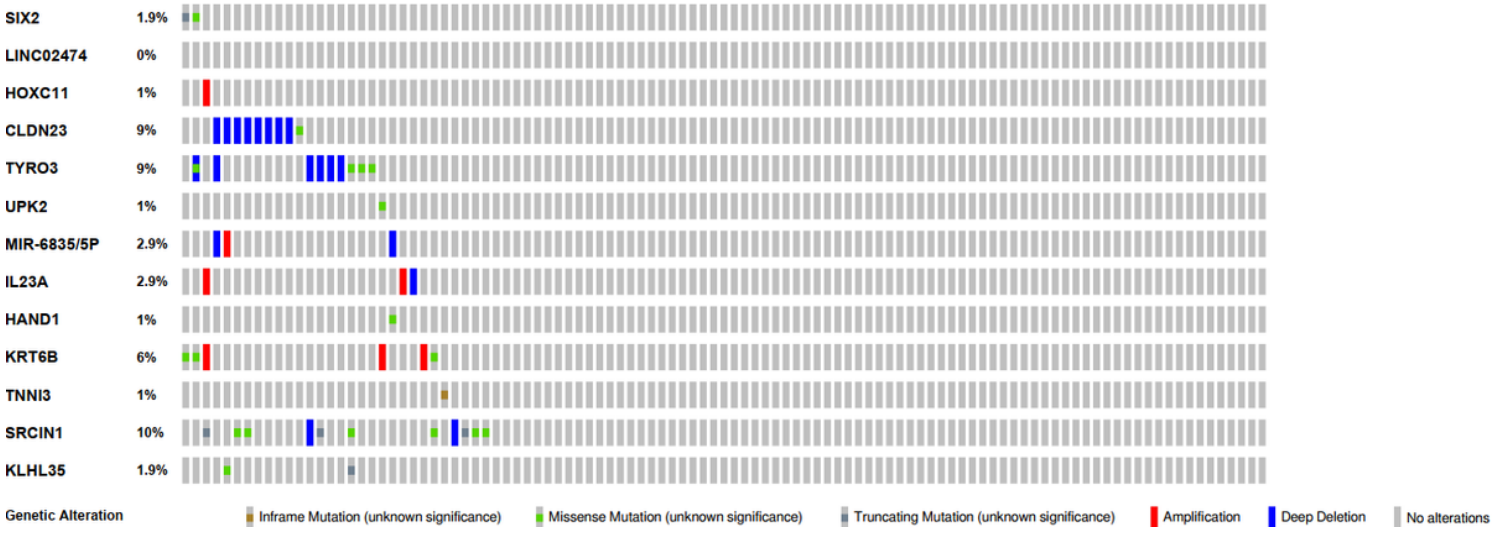

Figure 7

The genetic alterations of 13 signature genes in colon in GEPIA (cbioportal CPTAC-2 Prospective, Cell 2019 cohort) database.