Establishment of a 12-gene expression signature to predict colon cancer prognosis

Dalong Sun [1] , **Jing Chen** [2] , **Longzi Liu** [3] , **Guangxi Zhao** [4] , **Pingping Dong** [1] , **Bingrui Wu** [5] , **Jun Wang** [Corresp., 6] , **Ling Dong** [Corresp. 1]

1 Department of Gastroenterology and Hepatology, Zhongshan Hospital, Fudan University, Shanghai, China

2 Department of Neurology, Shanghai Fifth People's Hospital, Fudan University, Shanghai, China

3 Department of Hepatic Surgery, Liver Cancer Institute, and Key Laboratory of Carcinogenesis and Cancer Invasion (Ministry of Education), Zhongshan Hospital, Fudan University, Shanghai, China

4 Department of Gastroenterology, Shanghai East Hospital, Tongji University School of Medicine, Shanghai, China

5 Key Laboratory of Glycoconjugate Research Ministry of Public Health, Department of Biochemistry and Molecular Biology, Shanghai Medical College, Fudan University, Shanghai, China

6 Guangzhou Institute of Pediatrics, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, Guangdong Province, China

Corresponding Authors: Jun Wang, Ling Dong
Email address: jwang03@sibs.ac.cn, dong.ling@zs-hospital.sh.cn

Robust and accurate gene expression signature is essential to assist oncologists to determine which subset of patients at similar Tumor-Lymph Node-Metastasis (TNM) stage has high recurrence risk and could benefit from adjuvant therapies. Here we applied a two-step supervised machine-learning method and established a 12-gene expression signature to precisely predict colon adenocarcinoma (COAD) prognosis by using COAD RNA-seq transcriptome data from The Cancer Genome Atlas (TCGA). The predictive performance of the 12-gene signature was validated with two independent gene expression microarray datasets: GSE39582 includes 566 COAD cases for the development of six molecular subtypes with distinct clinical, molecular and survival characteristics; GSE17538 is a dataset containing 232 colon cancer patients for the generation of a metastasis gene expression profile to predict recurrence and death in COAD patients. The signature could effectively separate the poor prognosis patients from good prognosis group [disease specific survival (DSS): Kaplan Meier (KM) Log Rank $p=0.0034$; overall survival (OS): KM Log Rank $p=0.0336$] in GSE17538. For patients with proficient mismatch repair system (pMMR) in GSE39582, the signature could also effectively distinguish high risk group from low risk group [OS: KM Log Rank $p=0.005$; Relapse free survival (RFS): KM Log Rank $p=0.022$]. Interestingly, advanced stage patients were significantly enriched in high 12-gene score group (Fisher's exact test $p=0.0003$). After stage stratification, the signature could still distinguish poor prognosis patients in GSE17538 from good prognosis within stage II (Log Rank $p=0.01$) and stage II&III (Log Rank $p=0.017$) in the outcome of DFS. Within stage III or II / III pMMR patients treated with Adjuvant ChemoTherapies (ACT),

patients with higher 12-gene score showed poorer prognosis (III, OS: KM Log Rank p=0.046; III&II, OS: KM Log Rank p=0.041). Among stage II/III pMMR patients with lower 12-gene scores in GSE39582, subgroup receiving ACT showed significantly longer OS time compared with those who received no ACT (Log Rank p=0.021), while there is no obvious difference between counterparts among patients with higher 12-gene scores (Log Rank p=0.12). Besides COAD, our 12-gene signature is multifunctional in several other cancer types including kidney cancer, lung cancer, uveal and skin melanoma, brain cancer, and pancreatic cancer. Functional classification showed that the seven of the twelve genes are involved in immune system function and regulation. So our 12-gene signature could potentially be used to guide decisions about adjuvant therapy for patients with stage II/III and pMMR COAD.

Establishment of a 12-gene expression signature to predict colon cancer

prognosis

Dalong Sun[1#], Jing Chen[2#], Longzi Liu[3#], Guangxi Zhao[4], Pingping Dong[1], Bingrui Wu[5], Jun Wang[6 *], Ling Dong[1*]

1 Department of Gastroenterology and Hepatology, Zhongshan Hospital, Fudan University, Shanghai 200032, China;

2 Department of Neurology, Shanghai Fifth People's Hospital, Fudan University, Shanghai 200240, China;

3 Department of Hepatic Surgery, Liver Cancer Institute, and Key Laboratory of Carcinogenesis and Cancer Invasion (Ministry of Education), Zhongshan Hospital, Fudan University, Shanghai 200032, China;

4 Department of Gastroenterology, Shanghai East Hospital, Tongji University School of Medicine, Shanghai 200120, China;

5 Key Laboratory of Glycoconjugate Research Ministry of Public Health, Department of Biochemistry and Molecular Biology, Shanghai Medical College, Fudan University, Shanghai 200032, China;

6 Guangzhou Institute of Pediatrics, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, Guangdong, 510623, China;

[#] These authors contributed equally.

Corresponding authors:

Ling Dong,

dong.ling@zs-hospital.sh.cn

Jun Wang,

jwang03@sibs.ac.cn

Abstract

Robust and accurate gene expression signature is essential to assist oncologists to determine which subset of patients at similar Tumor-Lymph Node-Metastasis (TNM) stage has high recurrence risk and could benefit from adjuvant therapies. Here we applied a two-step supervised machine-learning method and established a 12-gene expression signature to precisely predict colon adenocarcinoma (COAD) prognosis by using COAD RNA-seq transcriptome data from The Cancer Genome Atlas (TCGA). The predictive performance of the 12-gene signature was validated with two independent gene expression microarray datasets: GSE39582 includes 566 COAD cases for the development of six molecular subtypes with distinct clinical, molecular and survival characteristics; GSE17538 is a dataset containing 232 colon cancer patients for the generation of a metastasis gene expression profile to predict recurrence and death in COAD patients. The signature could effectively separate the poor prognosis patients from good prognosis group [disease specific survival (DSS): Kaplan Meier (KM) Log Rank p=0.0034; overall survival (OS): KM Log Rank p=0.0336] in GSE17538. For patients with proficient mismatch repair system (pMMR) in GSE39582, the signature could also effectively distinguish high risk group from low risk group [OS: KM Log Rank p=0.005; Relapse free survival (RFS): KM Log Rank p=0.022]. Interestingly, advanced stage patients were significantly enriched in high 12-gene score group (Fisher's exact test p=0.0003). After stage stratification, the signature could still distinguish poor prognosis patients in GSE17538 from good prognosis within stage II (Log Rank p=0.01) and stage II&III (Log Rank p=0.017) in the outcome of DFS. Within stage III or II/III pMMR patients treated with Adjuvant ChemoTherapies (ACT), patients with higher 12-gene score showed poorer prognosis (III, OS: KM Log Rank p=0.046; III&II, OS: KM Log Rank p=0.041). Among stage II/III pMMR patients with lower 12-gene scores in GSE39582, subgroup receiving ACT showed significantly longer OS time compared with those who received no ACT (Log Rank p=0.021), while there is no obvious difference between counterparts among patients with higher 12-gene scores (Log Rank p=0.12). Besides COAD, our 12-gene signature is multifunctional in several other cancer types including kidney cancer, lung cancer, uveal and skin melanoma, brain cancer, and pancreatic cancer. Functional classification showed that the seven of the twelve genes are involved in immune system function and regulation. So our 12-gene signature could potentially be used to guide decisions about adjuvant therapy for patients with stage II/III and pMMR COAD.

Introduction

Colorectal cancer (CRC) is one of the most common cancers in men and women, representing almost 10% of the global cancer incidents and the third leading cause of cancer death worldwide(*McGuire, 2016*). CRC comprises three different subtypes according to distinct pathway operate: chromosomal-instable, microsatellite-instable, and CpG island methylator phenotype, all of which differ in morphology, genetic background, molecular profile, clinical behavior, and response to therapy(*De Sousa et al., 2013*). Current prognostic model based on the classic tumor-node-metastasis (TNM) staging is the standard prognosis factor for CRC in clinical practice. However, due to the high heterogeneity of disease, the patients at similar stage behave differently in terms of recurrence and response to chemotherapy often differs. Better parameters to guide patients' prognostic stratification and personalized medicine are urgently needed. Currently, some prognostic and predictive molecular markers have been developed. Microsatellite instability (MSI) is the molecular hallmark of DNA mismatch repair (MMR) deficiency. In stage II disease, MSI status helps select patients with high risk of developing recurrence(*Brychtova et al., 2017*). MSI status can also be a predictor of the benefit of adjuvant chemotherapy with fluorouracil in stage II and stage III colon cancer(*Ribic et al., 2003*). KRAS mutation status has been validated as a molecular marker for prediction of non-response to EGFR targeted drugs in metastatic CRC *(Cunningham et al., 2010; Karapetis et al., 2008; Siena et al., 2009)*. However, due to complex pathways contributing to cancer progression, single molecular marker might not be efficient enough to predict prognosis and individualize in selecting adjuvant therapy.

The development of gene expression profiling technologies such as microarray and Next Generation Sequencing (NGS) provide further opportunities to comprehensively characterize the molecular features of cancer. Gene-expression profiling has been used to develop genomic tests that may provide better predictions of clinical outcomes in combination with traditional clinicopathologic factors *(Gray et al., 2007; Venook et al., 2011; Meropol et al., 2011; Ebata et al., 2016; Moloney & Cotter, 2017; Guinney et al., 2015; Marisa et al., 2013; Smith et al., 2010; Gentles et al., 2015)*. Some commercially genomic assays are available for the prediction of clinical outcome in CRC patients. The most well-known one is the Oncotype DX Colon Cancer Assay, which is a 12-gene (7 cancer related genes and 5 reference genes) genomic test that has been used to help identify individuals with high recurrence risk from stage II colon cancer patients with T3 and MMR proficient tumors *(Gray et al., 2007; Venook et al., 2011; Meropol et al., 2011)*. However, the five reference genes in Oncotype DX Assay contain PGK1 and GPX1, which are important players in the process of energy metabolism and cellular oxidative stress, both of which are actively involved in cancer development and metastasis(*Ebata et al., 2016; Moloney & Cotter, 2017*). Normalization with PGK1 and GPX1 might have diluted the tumorous heterogeneities among cancer patients. In this work, we applied two steps of supervised machine-learning method and established a 12-gene expression signature to precisely predict colon adenocarcinoma (COAD) prognosis by exhaustively using expression of all genes of TCGA COAD patients.

99

Materials and Methods

TCGA and GEO datasets

RNA-seq data and clinic information for all cancer types were obtained from the Cancer Genome Altas (TCGA) RNA-seq database (https://cancergenome.nih.gov/). Microarray expression data and clinic information for COAD patients were retrieved from Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/).

Development of the gene expression signature

The development process has training and validation phase.

Training stage has two phases:

Phase I

Grouping

The TCGA colon adenocarcinoma (COAD) patients were used for the development of prototype of the 118-gene signature that could predict COAD prognosis. We applied a similar supervised machine learning method that was used for MammaPrint(*van T et al., 2002*). Forty-two patients experienced relapse within 3 years were designated as poor prognosis. Forty-nine patients who were relapse free for at least three years were categorized as good prognosis. The gene expression values were centered and scaled before grouping. For training dataset, 32 and 39 patients were randomly chosen from poor and good prognosis category, respectively. The rest patients were grouped as test dataset. Detailed clinic information was listed in **Suppl. Table 1**.

Selection of genes with high correlation to real prognosis status

Overall, there are 20530 genes in the raw RNA-seq data. The Pearson correlation coefficients with real prognosis status were calculated for all genes. Genes with absolute correlation coefficient greater than 0.3 were selected. To test whether such correlation coefficient distribution could be found by chance, a permutation method was used to generate 10,000 Monte-Carlo simulations randomizing the correlation between gene expression data of the 71 training patients and corresponding prognostic categories.

Supervised machine-learning method

Gene number incorporated in the signature needs to be optimized. 1510 genes were reordered by absolute coefficients from maximum to minimum. Starting from the top 2 genes on the list, 755 signatures were generated by adding two more genes from the top list each time until all the 1510 genes were exhaustively used as reporters. A Leave-One-Out Cross-Validation (LOOCV) method was employed to check the performances of these signatures:

Step 1: leave one tumor out;
Step 2: calculate the good- and poor-prognosis expression template by averaging the expression values for each gene incorporated in good-prognosis group and poor-prognosis group,

135 respectively. Then we defined a parameter called risk coefficient (risk-coef.). For a tumor, risk
136 coefficient was calculated with its gene expression profile and good- and poor-prognosis
137 expression template:

139 Risk-coef = cor-coef. to good template – cor-coef. to poor template;

141 Step 3, calculate the risk-coefs for all the remaining 70 training samples and the left out sample.
142 Reorder the 71 samples by ranking their risk-coefs from small to large. Determine the genomic
143 risk by taking first 32 tumors as high genomic risk and the rest 39 as low genomic risk. Check
144 the consistency between genomic risk and real risk for the left out sample;
145 Step 4, repeat step 1-3 iteratively until all the 71 samples have been left out once. Collect the
146 error counts when there is a disagreement between genomic risk and real risk for the left out
147 sample.

149   Better signatures with least error counts were selected.

151 *Cross-validation without information leak*
152   The 1510 genes were obtained using all training samples including the one left out for cross
153 validation. So there might have over-fitting issue due to information leak. A modified LOOCV
154 with no information leak was performed as below:
155 Step 1, leave one patient out;
156 Step 2, calculate the Pearson correlation coefficients with real prognosis status for all genes with
157 the reminding 70 training samples. Filter the genes with |coefficient|>=0.3.
158 Step 3, generate the signature with the genes selected and predict the genomic risk for the left out
159 sample.
160 Step 4, repeat step 1-3 iteratively until all the 71 samples have been left out once.

162 *Phase II*

163   Further machine learning process was applied to generate a concise scoring system. Before
164 machine learning, the RPKM (Reads Per Kilobase per Million mapped reads) values need
165 normalization, which was done through dividing them by geometric mean of RPKM values of
166 TFRC, GUSB, and RPLP0. Firstly, the TCGA COAD patients (**Suppl. Table 2**) were split into
167 training and test dataset. There is no significant difference between the clinicopathologic factors
168 of these two groups (**Table 1**). For each of the 118 genes, we calculated the coefficient and *p*-
169 value in univariate Cox Proportional Hazard regression model (CPH) with training dataset. Then
170 we reordered the gene list by sorting the univaraite Cox-regression *p*-value from minimum to
171 maximum. So the top genes have stronger correlations with cancer prognosis. Starting from the
172 top one gene in the list, we added one more gene iteratively from the top for multivariate CPH
173 analysis. In every iteration step, the fitness of established signature on test dataset was checked
174 by calculating Kaplan Meier Log Rank *p*-value (KM-*p*). At the end of iteration, signature

175   incorporating the top 12 genes has the minimum test dataset KM-*p* and was deemed as the
176   optimal one. The multivariate Cox coefficient of each gene in the final signature was extracted to
177   generate the scoring system:

178

179   $$Riskscore = \sum_{i=1}^{n} Ei * \beta i$$

180   *Ei: expression level of gene i; βi: multivariate Cox-regression coefficient of gene i.*
181   **Validation Stage**

182   The GEO microarray datasets were used to validate the final gene expression signature. For
183   genes with more than one probe, the probe showing minimum univariate CPH *p*-value was
184   selected. Relative expression level was obtained via dividing the probe signal by geometric mean
185   of signals of TFRC, GUSB, and RPLP0. For each tumor, a risk score was obtained by calculating
186   the weighted summation of relative expressions of the 12-gene. For a certain dataset, patients
187   with risk scores below the median value of the population were designated as low risk group,
188   while the rest of the patients were categorized as high risk group. Survival comparisons between
189   high and low risk group were conducted by Kaplan-Meier plotting. Log Rank *p* value <0.05 was
190   considered as significantly different. Other cancer types in TCGA library were also retrieved to
191   validate the 12-gene signature.

192

193   **Results**

194   **Development of signature prototype**

195   The development process was shown as the flow chart in **Figure 1**. With the TCGA COAD
196   data, an unbiased screening method was used to obtain 1510 genes showing absolute correlations
197   greater than 0.3 with disease outcomes. The frequency distribution of number of genes with
198   absolute coefficient no less than 0.3 in the 10,000 Monte-Carlo trials was displayed in **Figure S1**.
199   The probability of obtaining 1510 genes or more with an absolute correlation coefficients of at
200   least 0.3 with prognosis categories purely by chance was 0.0019 (p<0.05), which was fair for us
201   to reject the null hypothesis.
202   During the (Leave-One-Out Cross Validation) LOOCV process, 755 signatures were
203   generated. Least violation times were observed when signature employed the top 16, 36, 40, 42,
204   44, 46, 48, 50, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, or 118 genes. We
205   further found that the predictive accuracy rates were high towards the 71 training samples with
206   the signature containing the top 118 genes (**Figure 2**). We had the luxury to further validate the
207   established signatures using the remaining 20 independent samples in test dataset. For each
208   signature, receiver operating characteristic curve (ROC) was plotted with the information of risk-
209   coefs and real risk of the 91 TCGA patients to compare the performances of the 25 signatures.
210   There was no significant difference among the performances of these signatures (**Figure 3 and**
211   **Table 2**).
212   Because the above 1510 genes were obtained using all the training samples including the

213   one left out for cross validation, a modified LOOCV without information leak was performed.
214   Seventy-one additional signatures were created. The vast majority of the original 1510 genes
215   were shared by most of the 71 signatures (**Figure S2**). So there was very limited information
216   leak introduced during the previous training process.
217

Development of 12-gene signature

219   For the purpose of concise and simplicity, we further established a 12-gene expression signature
220   based on the 118 genes obtained in phase I training stage. Expressional coefficients were
221   assigned to respective genes. Each patient has a risk score by calculating the weighted
222   summation of expression values of the 12 genes. The Kaplan-Meier (KM) survival analysis
223   showed that among TCGA COAD patients, the high risk group displayed significantly poorer
224   prognosis than low risk group regarding to disease free survival (DFS) (training dataset: KM Log
225   Rank p=0.0001; test dataset: KM Log Rank p=0.0005) (**Figure 4**).
226

Prognostic values of the 12-gene signature in other COAD datasets

228   GSE17538 (GSE17536 and GSE17537) was used to validate the 12-gene expression signature.
229   With both clinic information and microarray gene expression of 232 colon cancer patients, Smith
230   and Freeman (15,18) established a metastasis gene expression profile to predict recurrence and
231   death in COAD patients. The 12-gene signature could effectively separate the poor prognosis
232   patients from good prognosis group [**Figure 5 (a)-(c)**, Disease specific survival (DSS): KM Log
233   Rank p=0.0034; Overall survival (OS): KM Log Rank p=0.0336; Disease free survival (DFS):
234   KM Log Rank p=0.0004]. After stage stratification, the signature could still distinguish poor
235   prognosis patients from good within stage II [**Figure 5(d)**, Log Rank p=0.01] and stage II&III
236   [**Figure 5(e)**: Log Rank p=0.017] in terms of DFS.

237   GSE39582 is a dataset including 566 COAD cases and 19 non-tumoral colorectal mucosas. With
238   this dataset, Marisa *et al.* developed gene expression classification of colon cancer defining six
239   molecular subtypes with distinct clinical, molecular and survival characteristics (**19**). In patients
240   with proficient mismatch repair system (pMMR), our 12-gene signature could effectively
241   distinguish high risk group from low risk group [**Figure 6 (a)** and **(b)**, Relapse free survival
242   (RFS): KM Log Rank p=0.022; OS: KM Log Rank p=0.005]. No significant difference was
243   found in KM analysis performed among dMMR patients. Further survival analysis was
244   performed within stage III or II&III and pMMR patients treated with Adjuvant Chemotherapies
245   (ACT): patients with higher 12-gene score showed poorer prognosis [**Figure 6 (c)** and **(d):** III,
246   OS: KM Log Rank p=0.046; III&II, OS: KM Log Rank p=0.041]. Interestingly, among stage
247   II&III pMMR patients with lower 12-gene scores, subgroup receiving adjuvant chemotherapies
248   showed significantly longer OS time compared with those who received no adjuvant
249   chemotherapy [**Figure 6 (e)**: Log Rank p=0.021], while there is no obvious difference between
250   counterparts among patients with higher 12-gene scores[**Figure 6 (f)**: Log Rank p=0.12**]**.

251   Interestingly, advanced stage patients were significantly enriched in high 12-gene score group
252   (**Table 3**).

253
254 **Predictive performances of the 12-gene signature in other cancer types**
255 We also tested the performance of the signature in other cancer types. TCGA RNA-seq data and
256 corresponding clinic information for 24 cancer types were retrieved for validation. Surprisingly,
257 KM results showed that our signature successfully separated good prognosis patients from poor
258 prognosis patients in several other cancer types including pan-kidney cohort (KIPAN) (**Figure**
259 **7a**,OS: KM Log Rank p=6.815e-6), kidney renal clear cell carcinoma (KIRC) (**Figure 7b**, DFS:
260 KM Log Rank p=0.0480), kidney renal papillary cell carcinoma (KIRP) (**Figure 7c**, DFS:KM
261 Log Rank p=0.0027; **Figure 7d** OS: Log Rank p=0.0129), lung squamous cell carcinoma
262 (LUSC) (**Figure 7e**, DFS: Log Rank p=0.0071), and skin cutaneous melanoma (SKCM) (**Figure**
263 **7f**, DFS: Log Rank p=0.01117), brain lower grade glioma (LGG) (**Figure 8a**, OS: Log Rank
264 p=0.0031), uveal melanoma (UVM) (**Figure 8b**, OS: Log Rank p=0.0054), glioblastoma (GBM)
265 (**Figure 8c**, OS: Log Rank p=0.0074), cervical and endocervical cancers (CESC) (**Figure 8d**,
266 OS: Log Rank p=0.0090), pancreatic adenocarcinoma (PAAD) (**Figure 8e**, OS: Log Rank
267 p=0.0127), stomach adenocarcinoma (STAD) (**Figure 8f**, OS: Log Rank p=0.0456).
268

269 **Discussion**

270   Numerous attempts have been made to establish gene expression signatures for the purpose
271 of precise prediction of colorectal cancer prognosis(*Gray et al., 2007; Venook et al., 2011;*
272 *Meropol et al., 2011; Ebata et al., 2016; Moloney & Cotter, 2017; Guinney et al., 2015; Marisa*
273 *et al., 2013; Smith et al., 2010; Gentles et al., 2015*). A meta-analysis was done to assess the
274 clinical value of several published prognosis gene expression signatures in colorectal cancer
275 (*Sanz-Pamplona et al., 2012*). Although most of the published signatures showed significant
276 statistical association with prognosis, their accuracy to classify independent tumor samples into
277 high-risk and low-risk group is limited. So we need more robust and accurate gene expression
278 signature that can predict prognosis cross independent COAD datasets. Here we established a
279 gene expression signature by applying two steps of supervised machine-learning method. The
280 predicative accuracy of our gene expression signature was proven by validation in two large
281 independent gene expression microarray datasets (GSE39582, N=459; GSE17538, N=232).
282 Decision making regarding adjuvant therapy has been a debate among professional clinical
283 organizations over the past 20 years *(Dotan & Cohen, 2011; Meropol, 2011; Vachani C, 2013*).
284 Currently speaking, uncertainty presents in adjuvant chemotherapeutic effect among stage II
285 COAD patients who are mismatch repair system proficient. The Scottish Intercollegiate
286 Guidelines Network (SIGN), ASCO, and NCCN are following different guidelines regarding this
287 issue (*Gao et al., 2016*). Resectable COAD patients with pMMR routinely receive 5-FU based
288 postoperative adjuvant chemotherapy (POCT) which has been shown to provide a relatively
289 small absolute benefit *(Andre et al., 2009; Gill et al., 2004; Sargent et al., 2009; Gray et al.,*
290 *2011; Alex et al., 2017*), indicating many COAD patients might have been over-treated due to the
291 lacking of an effective test to stratify the patients further. Our gene signature showed important
292 prognostic value for stage II or/and III pMMR COAD patients. There validation results in

293 GSE39582 indicate that lower 12-gene score patients have gained survival benefit from adjuvant
294 chemotherapies, while high score patients treated with adjuvant chemotherapies didn't receive
295 survival benefit. So our 12-gene signature could potentially be used to guide decisions about
296 adjuvant therapy for patients with stage II&III and pMMR colon cancer.

297   Seven of the proteins encoded by the 12 genes were related to immune system, they are
298 TREML2, PADI4, NCKIPSD, PTPRN, PGLYRP1, C5orf53, and TREML3, indicating the
299 essential roles of deregulated immune response in COAD progression and metastasis (**Suppl.**
300 **Table 3**). TREML2, acting as the counter-receptor for B7-H3, promotes T cell
301 responses(*Hashiguchi et al., 2008*). PADI4 protein catalyzes the conversion of arginine to
302 citrulline residue. With specific high expression in blood lymphocytes*(Asaga et al., 2001;*
303 *Anzilotti et al., 2010*), PADI4 is believed to be an active autoimmune player in synovial tissue of
304 rheumatoid arthritis (*Chang et al., 2005*). It is reported that cell free circulation PADI4 mRNA
305 level (together with cfDNA, PPBP, and haptoglobin) in peripheral blood of non-small cell lung
306 cancer patients was significantly higher than that in healthy donors. So PADI4 may serve as a
307 potential marker for NSCLC diagnosis (*Ulivi et al., 2013*). As a member of protein tyrosine
308 phosphatase (PTP), PTPRN is an autoantigen in the sera of insulin-dependent diabetes mellitus
309 (IDDM) patients, making it a promising therapeutic target of autoimmunity in IDDM (*Rabin et*
310 *al., 1994; Solimena et al., 1996*). Hypermethylation in PTPRN was associated with longer
311 progression-free survival in ovarian cancer (Bauerschlag et al., 2011). If that is the case,
312 hypomethylation (upregulated mRNA expression level) in PTPRN may be associated with poor
313 prognosis, which is consistent with our results. NCKIPSD is a protein containing SH3 and
314 proline-rich domains. Reports have shown that NCKIPSD is involved in the maintenance of
315 sarcomeres and assembly of myofibrils into sarcomeres*(Lim et al., 2001)*. A very recent study
316 reported that NCKIPSDdownregulation and increased α-tublin acetylation promotes stiffness of
317 tumor stroma, which in turn, may inhibit chemotherapeutic drug uptake and regulate tumor
318 sensitivity to chemotherapy, resulting in high risk of breast cancer recurrence within 5 years (*You*
319 *et al., 2017*). Consistently, our findings also showed decreased NCKIPSD expression is
320 associated with high risk of colon cancer recurrence. PGLYRP1 is a member of peptidoglycan
321 recognition proteins which are conserved innate immunity proteins, recognize bacterial
322 peptidoglycan, and play a role in antibacterial immunity and inflammation(*Dziarski & Gupta,*
323 *2010*). PGLYRP1 interacts with Hsp70 to induces cytotoxic activity in tumor cells via TNFR1
324 receptor (*Yashin et al., 2015*). C5orf53 is also called IgA inducing protein, which enhances IgA
325 secretion from B-cells stimulated via CD40 (*Endsley et al., 2009*). TREML3 is a inhibitory
326 receptor involved in antigen processing(*Cella et al., 1997*). Numerous studies have shown that
327 cancer patients' prognosis and sensitivity to therapy are closely associated with infiltration and
328 density of immunologic cells within primary tumors (*Wels et al., 2008; McConnell & Yang,*
329 *2009; McLean et al., 2011; Sethi & Kang, 2011; Smith & Kang, 2013*). Of particular note, by
330 applying a novel machine-learning method, called Cell-type Identification By Estimating
331 Relative Subsets of known RNA Transcripts (CIBERSORT), Gentles et al. developed several
332 gene expression signatures to inferring distinct leukocyte subsets representation in bulk tumor

333   transcriptomes (*Gentles et al., 2015*). In several solid tumors including colon cancer, the
334   signatures relating to plasma cells and polymorphonuclear cells were the most significant
335   favorable and adverse module to cancer outcomes, respectively. The broad spectrum
336   involvement of lymphocyte infiltration and intra-tumor immune-suppression implies that this
337   could be the main reason why our 12-gene signature could also predict patient prognosis in
338   several other cancer types including kidney cancer, lung cancer, uveal and skin melanoma, brain
339   cancer, and pancreatic cancer.

340   Other 5 genes (NOG, VIP, RIMKLB, NKAIN4, and FAM171B) in the 12-gene signature
341   are functionally sporadic. NOG is related to mesodermal commitment and differentiation
342   pathway(*Costamagna et al., 2016*). High expressing of gene signature including NOG showed a
343   strong trend for a worse prognosis of patients with lung adenocarcinomas *(Rajski et al., 2015)*.
344   VIP, a member of glucagon/secretin superfamily, is the ligand of class II G protein-coupled
345   receptor (*Umetsu et al., 2011*). It causes vasodilation and lowers arterial blood pressure. VIP
346   signaling is enhanced in human prostate cancer (Fernandez-Martinez et al., 2012). Elevated VIP
347   secretion is associated with advanced tumor stage in colorectal carcinoma (*Hirayasu et al.,*
348   *2002*). RIMKLB is involved in alanine, aspartate and glutamate metabolism. RIMKLB up-
349   regulation is associated with radio-resistance in nasopharyngeal carcinomas *(Li et al.,*
350   *2016)*.NKAIN4 may interact with the beta subunit of Na, K-ATPase*(Gorokhova et al., 2007)*.
351   FAM171B which is a single-pass type I membrane protein, belongs to the FAM171 family. It is
352   up-regulated in gemcitabine-resistant pancreatic cancer cell line (*Zhou et al., 2015*).The
353   associations of these genes with cancer and cancer outcomes are very relevant to our findings in
354   this study.

355   Our signature generated a novel scoring system with relative gene expression values by
356   dividing the raw expression with geometric mean of RPKM values of three house-keeping genes
357   (TFRC, GUSB, and RPLP0). In order to preserve the heterogeneities among tumors to the most
358   extent, ACTB and GAPDH were avoided using as reference genes due to the fact that
359   cytoskeleton and energy metabolism might be greatly deregulated among cancer individuals
360   (*Xiang et al., 2017; Stine & Dang, 2013*). A recent study overcomes hypoxia-induced tumor cell
361   resistance by synergistic GAPDH-siRNA and chemotherapy (*Guan et al., 2017*), indicating the
362   important roles of GAPDH in tumor cell resistance. Our normalization process also makes the
363   gene expression scoring system very friendly to different gene expression detection systems
364   including qPCR, RNA-seq, and QuantiGene Plex.

365

Conclusion

366

367   Robust and accurate gene expression signature is essential to assist oncologists to determine
368   which subset of patients at similar TNM stage has high recurrence risk and could benefit from
369   adjuvant therapies. Here we report a new 12-gene expression signature that could separate
370   resectable COAD patients into poor- and good-prognosis group in several independent TCGA
371   and microarray datasets. Functional classification showed that the seven of the twelve genes are

372 involved in immune system function and regulation. Our gene expression signature could
373 potentially serve as an effective genomic test in helping identify resectable COAD patients with
374 high risk of poor prognosis. The accuracy and robustness of the signature as a prognostic
375 classification requires further confirmation with large prospective patient cohorts.

376 References:

377 **Alex, A.K., Siqueira, S., Coudry, R., Santos, J., Alves, M., Hoff, P.M., and Riechelmann, R.P. 2017**. Response
378 to Chemotherapy and Prognosis in Metastatic Colorectal Cancer With DNA Deficient Mismatch Repair. *Clin*
379 *Colorectal Cancer* **16**:228-239. DOI: 10.1016/j.clcc.2016.11.001

380 **Andre, T., Boni, C., Navarro, M., Tabernero, J., Hickish, T., Topham, C., Bonetti, A., Clingan, P.,**
381 **Bridgewater, J., Rivera, F., and de Gramont, A. 2009**. Improved overall survival with oxaliplatin,
382 fluorouracil, and leucovorin as adjuvant treatment in stage II or III colon cancer in the MOSAIC trial.
383 *JOURNAL OF CLINICAL ONCOLOGY* **27**:3109-3116. DOI: 10.1200/JCO.2008.20.6771

384 **Anzilotti, C., Pratesi, F., Tommasi, C., and Migliorini, P. 2010.** Peptidylarginine deiminase 4 and citrullination in
385 health and disease. *AUTOIMMUNITY REVIEWS* 9:158-160. DOI 10.1016/j.autrev.2009.06.002

386 **Asaga, H., Nakashima, K., Senshu, T., Ishigami, A., and Yamada, M. 2001.** Immunocytochemical localization
387 of peptidylarginine deiminase in human eosinophils and neutrophils. *J Leukoc Biol* **70**:46-51.

388 **Bauerschlag, D.O., Ammerpohl, O., Brautigam, K., Schem, C., Lin, Q., Weigel, M.T., Hilpert, F., Arnold, N.,**
389 **Maass, N., Meinhold-Heerlein, I., and Wagner, W. 2011.** Progression-free survival in ovarian cancer is
390 reflected in epigenetic DNA methylation profiles. *ONCOLOGY* **80**:12-20. DOI: 10.1159/000327746

391 **Brychtova, V., Sefr, R., Hrstka, R., Videnska, P., Bencsikova, B., Hanakova, B., Zdrazilova, D.L., Nenutil, R.,**
392 **and Budinska, E. 2017.** Molecular Pathology of Colorectal Cancer, Microsatellite Instability - the Detection,
393 the Relationship to the Pathophysiology and Prognosis. *Klin Onkol* **30**:153-155.

394 **Cella, M., Dohring, C., Samaridis, J., Dessing, M., Brockhaus, M., Lanzavecchia, A., and Colonna, M. 1997.** A
395 novel inhibitory receptor (ILT3) expressed on monocytes, macrophages, and dendritic cells involved in antigen
396 processing. *JOURNAL OF EXPERIMENTAL MEDICINE* **185**:1743-1751.

397 **Chang, X., Yamada, R., Suzuki, A., Sawada, T., Yoshino, S., Tokuhiro, S., and Yamamoto, K. 2005.**
398 Localization of peptidylarginine deiminase 4 (PADI4) and citrullinated protein in synovial tissue of rheumatoid
399 arthritis. *Rheumatology (Oxford)* **44**:40-50. DOI 10.1093/rheumatology/keh414**Costamagna, D., Mommaerts,**
400 **H., Sampaolesi, M., and Tylzanowski, P. 2016.** Noggin inactivation affects the number and differentiation
401 potential of muscle progenitor cells in vivo. *Sci Rep* **6**:31949. DOI 10.1038/srep31949

402 **Cunningham, D., Atkin, W., Lenz, H.J., Lynch, H.T., Minsky, B., Nordlinger, B., and Starling, N. 2010.**
403 Colorectal cancer. *LANCET* **375**:1030-1047. DOI: 10.1016/S0140-6736(10)60353-4

404 **De Sousa, E.M.F., Wang, X., Jansen, M., Fessler, E., Trinh, A., de Rooij, L.P., de Jong, J.H., de Boer, O.J.,**
405 **van Leersum, R., Bijlsma, M.F., Rodermond, H., van der Heijden, M., van Noesel, C.J., Tuynman, J.B.,**
406 **Dekker, E., Markowetz, F., Medema, J.P., and Vermeulen, L. 2013.** Poor-prognosis colon cancer is defined
407 by a molecularly distinct subtype and develops from serrated precursor lesions. *NATURE MEDICINE* **19**:614-
408 618. DOI: 10.1038/nm.3174

409 **Dotan, E., and Cohen, S.J. 2011.** Challenges in the management of stage II colon cancer. *SEMINARS IN*
410 *ONCOLOGY* **38**:511-520. DOI: 10.1053/j.seminoncol.2011.05.005

411   **Dziarski, R., and Gupta, D. 2010.** Review: Mammalian peptidoglycan recognition proteins (PGRPs) in innate
412   immunity. *Innate Immun* **16**:168-174. DOI 10.1177/1753425910366059

413   **Ebata, T., Hirata, H., and Kawauchi, K. 2016.** Functions of the Tumor Suppressors p53 and Rb in Actin
414   Cytoskeleton Remodeling. *Biomed Research International* **2016**: 9231057. DOI: 10.1155/2016/9231057

415   **Endsley, M.A., Njongmeta, L.M., Shell, E., Ryan, M.W., Indrikovs, A.J., Ulualp, S., Goldblum, R.M., Mwangi,**
416   **W., and Estes, D.M. 2009.** Human IgA-inducing protein from dendritic cells induces IgA production by naive
417   IgD+ B cells. *JOURNAL OF IMMUNOLOGY* **182**:1854-1859. DOI 10.4049/jimmunol.0801973

418   **Fernandez-Martinez, A.B., Carmena, M.J., Arenas, M.I., Bajo, A.M., Prieto, J.C., and Sanchez-Chapado, M.**
419   **2012.** Overexpression of vasoactive intestinal peptide receptors and cyclooxygenase-2 in human prostate cancer.
420   Analysis of potential prognostic relevance. *HISTOLOGY AND HISTOPATHOLOGY* 27:1093-1101. DOI:
421   10.14670/HH-27.1093

422   **Gao, S., Tibiche, C., Zou, J., Zaman, N., Trifiro, M., O'Connor-McCourt, M., and Wang, E. 2016.**
423   Identification and Construction of Combinatory Cancer Hallmark-Based Gene Signature Sets to Predict
424   Recurrence and Chemotherapy Benefit in Stage II Colorectal Cancer. *JAMA Oncology* **2**:37-45. DOI:
425   10.1001/jamaoncol.2015.3413

426   **Gentles, A.J., Newman, A.M., Liu, C.L., Bratman, S.V., Feng, W., Kim, D., Nair, V.S., Xu, Y., Khuong, A.,**
427   **Hoang, C.D., Diehn, M., West, R.B., Plevritis, S.K., and Alizadeh, A.A. 2015.** The prognostic landscape of
428   genes and infiltrating immune cells across human cancers. *NATURE MEDICINE* **21**:938-945. DOI:
429   10.1038/nm.3909

430   **Gill, S., Loprinzi, C.L., Sargent, D.J., Thome, S.D., Alberts, S.R., Haller, D.G., Benedetti, J., Francini, G.,**
431   **Shepherd, L.E., Francois, S.J., Labianca, R., Chen, W., Cha, S.S., Heldebrant, M.P., and Goldberg, R.M.**
432   **2004.** Pooled analysis of fluorouracil-based adjuvant therapy for stage II and III colon cancer: who benefits and
433   by how much? *JOURNAL OF CLINICAL ONCOLOGY* **22**:1797-1806. DOI: 10.1200/JCO.2004.09.059

434   **Gorokhova, S., Bibert, S., Geering, K., and Heintz, N. 2007.** A novel family of transmembrane proteins
435   interacting with beta subunits of the Na,K-ATPase. *HUMAN MOLECULAR GENETICS* **16**:2394-2410. DOI
436   10.1093/hmg/ddm167

437   **Gray, R., Barnwell, J., McConkey, C., Hills, R.K., Williams, N.S., and Kerr, D.J. 2007.** Adjuvant chemotherapy
438   versus observation in patients with colorectal cancer: a randomised study. *LANCET* **370**:2020-2029. DOI:
439   10.1016/S0140-6736(07)61866-2

440   **Gray, R.G., Quirke, P., Handley, K., Lopatin, M., Magill, L., Baehner, F.L., Beaumont, C., Clark-Langone,**
441   **K.M., Yoshizawa, C.N., Lee, M., Watson, D., Shak, S., and Kerr, D.J. 2011.** Validation study of a
442   quantitative multigene reverse transcriptase-polymerase chain reaction assay for assessment of recurrence risk
443   in patients with stage II colon cancer. *JOURNAL OF CLINICAL ONCOLOGY* **29**:4611-4619. DOI:
444   10.1200/JCO.2010.32.8732

445   **Guan, J., Sun, J., Sun, F., Lou, B., Zhang, D., Mashayekhi, V., Sadeghi, N., Storm, G., Mastrobattista, E., and**
446   **He, Z. 2017.** Hypoxia-induced tumor cell resistance is overcome by synergistic GAPDH-siRNA and
447   chemotherapy co-delivered by long-circulating and cationic-interior liposomes. *Nanoscale* **9**:9190-9201. DOI:
448   10.1039/c7nr02663c

449   **Guinney, J., Dienstmann, R., Wang, X., de Reynies, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P.,**
450   **Nyamundanda, G., Angelino, P., Bot, B.M., Morris, J.S., Simon, I.M., Gerster, S., Fessler, E., De Sousa,**
451   **E.M.F., Missiaglia, E., Ramay, H., Barras, D., Homicsko, K., Maru, D., Manyam, G.C., Broom, B., Boige,**

V., Perez-Villamil, B., Laderas, T., Salazar, R., Gray, J.W., Hanahan, D., Tabernero, J., Bernards, R., Friend, S.H., Laurent-Puig, P., Medema, J.P., Sadanandam, A., Wessels, L., Delorenzi, M., Kopetz, S., Vermeulen, L., and Tejpar, S. 2015. The consensus molecular subtypes of colorectal cancer. *NATURE MEDICINE* **21**:1350-1356. DOI: 10.1038/nm.3967

Hashiguchi, M., Kobori, H., Ritprajak, P., Kamimura, Y., Kozono, H., and Azuma, M. 2008. Triggering receptor expressed on myeloid cell-like transcript 2 (TLT-2) is a counter-receptor for B7-H3 and enhances T cell responses. *Proc Natl Acad Sci U S A* **105**:10495-10500. DOI 10.1073/pnas.0802423105

Hirayasu, Y., Oya, M., Okuyama, T., Kiumi, F., and Ueda, Y. 2002. Vasoactive intestinal peptide and its relationship to tumor stage in colorectal carcinoma: an immunohistochemical study. *JOURNAL OF GASTROENTEROLOGY* **37**:336-344. DOI: 10.1007/s005350200047

Karapetis, C.S., Khambata-Ford, S., Jonker, D.J., O'Callaghan, C.J., Tu, D., Tebbutt, N.C., Simes, R.J., Chalchal, H., Shapiro, J.D., Robitaille, S., Price, T.J., Shepherd, L., Au, H.J., Langer, C., Moore, M.J., and Zalcberg, J.R. 2008. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *N Engl J Med* **359**:1757-1765. DOI: 10.1056/NEJMoa0804385

Li, G., Liu, Y., Liu, C., Su, Z., Ren, S., Wang, Y., Deng, T., Huang, D., Tian, Y., and Qiu, Y. 2016. Genome-wide analyses of long noncoding RNA expression profiles correlated with radioresistance in nasopharyngeal carcinoma via next-generation deep sequencing. *BMC CANCER* **16**:719. DOI: 10.1186/s12885-016-2755-6

Lim, C.S., Park, E.S., Kim, D.J., Song, Y.H., Eom, S.H., Chun, J.S., Kim, J.H., Kim, J.K., Park, D., and Song, W.K. 2001. SPIN90 (SH3 protein interacting with Nck, 90 kDa), an adaptor protein that is developmentally regulated during cardiac myocyte differentiation. *JOURNAL OF BIOLOGICAL CHEMISTRY* **276**:12871-12878. DOI 10.1074/jbc.M009411200

Marisa, L., de Reynies, A., Duval, A., Selves, J., Gaub, M.P., Vescovo, L., Etienne-Grimaldi, M.C., Schiappa, R., Guenot, D., Ayadi, M., Kirzin, S., Chazal, M., Flejou, J.F., Benchimol, D., Berger, A., Lagarde, A., Pencreach, E., Piard, F., Elias, D., Parc, Y., Olschwang, S., Milano, G., Laurent-Puig, P., and Boige, V. 2013. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLOS MEDICINE* **10**:e1001453. DOI: 10.1371/journal.pmed.1001453

McConnell, B.B., and Yang, V.W. 2009. The Role of Inflammation in the Pathogenesis of Colorectal Cancer. *Curr Colorectal Cancer Rep* **5**:69-74. DOI: 10.1007/s11888-009-0011-z

McGuire, S. 2016. World Cancer Report 2014. Geneva, Switzerland: World Health Organization, International Agency for Research on Cancer, WHO Press, 2015. *Advances in Nutrition* **7**:418-419. DOI: 10.3945/an.116.012211

McLean, M.H., Murray, G.I., Stewart, K.N., Norrie, G., Mayer, C., Hold, G.L., Thomson, J., Fyfe, N., Hope, M., Mowat, N.A., Drew, J.E., and El-Omar, E.M. 2011. The inflammatory microenvironment in colorectal neoplasia. *PLoS One* **6**:e15366. DOI: 10.1371/journal.pone.0015366

Meropol, N.J. 2011. Ongoing challenge of stage II colon cancer. *JOURNAL OF CLINICAL ONCOLOGY* **29**:3346-3348. DOI: 10.1200/JCO.2011.35.4571

Meropol, N.J., Lyman, G.H., Chien, R., and Hornberger, J.C. 2011. Use of a multigene prognostic assay for selection of adjuvant chemotherapy in patients with stage II colon cancer: Impact on quality-adjusted life expectancy and costs. *JOURNAL OF CLINICAL ONCOLOGY* **29S**. DOI: 10.1200/jco.2011.29.4_suppl.491

Moloney, J.N., and Cotter, T.G. 2017. ROS signalling in the biology of cancer. *SEMINARS IN CELL & DEVELOPMENTAL BIOLOGY*. DOI: 10.1016/j.semcdb.2017.05.023

493 **Rabin, D.U., Pleasic, S.M., Shapiro, J.A., Yoo-Warren, H., Oles, J., Hicks, J.M., Goldstein, D.E., and Rae,**
494 **P.M. 1994.** Islet cell antigen 512 is a diabetes-specific islet autoantigen related to protein tyrosine phosphatases.
495 *JOURNAL OF IMMUNOLOGY* **152**:3183-3188.

496 **Rajski, M., Saaf, A., and Buess, M. 2015.** BMP2 response pattern in human lung fibroblasts predicts outcome in
497 lung adenocarcinomas. *BMC Medical Genomics* **8**:16. DOI: 10.1186/s12920-015-0090-4

498 **Ribic, C.M., Sargent, D.J., Moore, M.J., Thibodeau, S.N., French, A.J., Goldberg, R.M., Hamilton, S.R.,**
499 **Laurent-Puig, P., Gryfe, R., Shepherd, L.E., Tu, D., Redston, M., and Gallinger, S. 2003.** Tumor
500 microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for
501 colon cancer. *N Engl J Med* **349**:247-257. DOI: 10.1056/NEJMoa022289

502 **Sanz-Pamplona, R., Berenguer, A., Cordero, D., Riccadonna, S., Sole, X., Crous-Bou, M., Guino, E., Sanjuan,**
503 **X., Biondo, S., Soriano, A., Jurman, G., Capella, G., Furlanello, C., and Moreno, V. 2012.** Clinical value
504 of prognosis gene expression signatures in colorectal cancer: a systematic review. *PLoS One* **7**:e48877. DOI:
505 10.1371/journal.pone.0048877

506 **Sargent, D., Sobrero, A., Grothey, A., O'Connell, M.J., Buyse, M., Andre, T., Zheng, Y., Green, E., Labianca,**
507 **R., O'Callaghan, C., Seitz, J.F., Francini, G., Haller, D., Yothers, G., Goldberg, R., and de Gramont, A.**
508 **2009.** Evidence for cure by adjuvant therapy in colon cancer: observations based on individual patient data
509 from 20,898 patients on 18 randomized trials. *JOURNAL OF CLINICAL ONCOLOGY* **27**:872-877. DOI:
510 10.1200/JCO.2008.19.5362

511 **Sethi, N., and Kang, Y. 2011.** Unravelling the complexity of metastasis - molecular understanding and targeted
512 therapies. *NATURE REVIEWS CANCER* **11**:735-748. DOI: 10.1038/nrc3125

513 **Siena, S., Sartore-Bianchi, A., Di Nicolantonio, F., Balfour, J., and Bardelli, A. 2009.** Biomarkers predicting
514 clinical outcome of epidermal growth factor receptor-targeted therapy in metastatic colorectal cancer. *J Natl*
515 *Cancer Inst* **101**:1308-1324. DOI: 10.1093/jnci/djp280

516 **Smith, H.A., and Kang, Y. 2013.** The metastasis-promoting roles of tumor-associated immune cells. *J Mol Med*
517 *(Berl)* **91**:411-429. DOI: 10.1007/s00109-013-1021-5

518 **Smith, J.J., Deane, N.G., Wu, F., Merchant, N.B., Zhang, B., Jiang, A., Lu, P., Johnson, J.C., Schmidt, C.,**
519 **Bailey, C.E., Eschrich, S., Kis, C., Levy, S., Washington, M.K., Heslin, M.J., Coffey, R.J., Yeatman, T.J.,**
520 **Shyr, Y., and Beauchamp, R.D. 2010**. Experimentally derived metastasis gene expression profile predicts
521 recurrence and death in patients with colon cancer. *GASTROENTEROLOGY* **138**:958-968. DOI:
522 10.1053/j.gastro.2009.11.005

523 **Solimena, M., Dirkx, R.J., Hermel, J.M., Pleasic-Williams, S., Shapiro, J.A., Caron, L., and Rabin, D.U. 1996.**
524 ICA 512, an autoantigen of type I diabetes, is an intrinsic membrane protein of neurosecretory granules. *EMBO*
525 *JOURNAL* **15**:2102-2114.

526 **Stine, Z.E., and Dang, C.V. 2013.** Stress eating and tuning out: Cancer cells re-wire metabolism to counter stress.
527 *CRITICAL REVIEWS IN BIOCHEMISTRY AND MOLECULAR BIOLOGY* **48**:609-619. DOI:
528 10.3109/10409238.2013.844093

529 **Ulivi, P., Mercatali, L., Casoni, G.L., Scarpi, E., Bucchi, L., Silvestrini, R., Sanna, S., Monteverde, M.,**
530 **Amadori, D., Poletti, V., and Zoli, W. 2013**. Multiple marker detection in peripheral blood for NSCLC
531 diagnosis. *PLoS One* **8**:e57401. DOI: 10.1371/journal.pone.0057401

532 **Umetsu, Y., Tenno, T., Goda, N., Shirakawa, M., Ikegami, T., and Hiroaki, H. 2011.** Structural difference of
533 vasoactive intestinal peptide in two distinct membrane-mimicking environments. *Biochim Biophys Acta*

534   **1814**:724-730. DOI 10.1016/j.bbapap.2011.03.009

535   **Vachani C, G.B. 2013.** Stage II Colon Cancer: To Treat or Not to Treat? *Available at*
536   *https://www.oncolink.org/cancers/gastrointestinal/colon-cancer/treatments/stage-ii-colon-cancer-to-treat-or-*
537   *not-to-treat* (accessed April 18 2018).

538   **van T, V.L., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K.,**
539   **Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards,**
540   **R., and Friend, S.H. 2002.** Gene expression profiling predicts clinical outcome of breast cancer. *NATURE*
541   **415**:530-536. DOI: 10.1038/415530a

542   **Venook, A.P., Niedzwiecki, D., Lopatin, M., Lee, M., Friedman, P.N., Frankel, W., Clark-Langone, K.,**
543   **Yoshizawa, C., Millward, C., Shak, S., Goldberg, R.M., Mahmoud, N.N., Schilsky, R.L., and Bertagnolli,**
544   **M.M. 2011.** Validation of a 12-gene colon cancer recurrence score (RS) in patients (pts) with stage II colon
545   cancer (CC) from CALGB 9581. *JOURNAL OF CLINICAL ONCOLOGY* **29S**:3518-3518.

546   **Wels, J., Kaplan, R.N., Rafii, S., and Lyden, D. 2008.** Migratory neighbors and distant invaders: tumor-associated
547   niche cells. *Genes Dev* **22**:559-574. DOI: 10.1101/gad.1636908

548   **Xiang, C., Chen, J., and Fu, P. 2017.** HGF/Met Signaling in Cancer Invasion: The Impact on Cytoskeleton
549   Remodeling. *Cancers* **9**:44. DOI: 10.3390/cancers9050044

550   Yashin, D.V., Ivanova, O.K., Soshnikova, N.V., Sheludchenkov, A.A., Romanova, E.A., Dukhanina, E.A.,
551   Tonevitsky, A.G., Gnuchev, N.V., Gabibov, A.G., Georgiev, G.P., and Sashchenko, L.P. 2015. Tag7
552   (PGLYRP1) in Complex with Hsp70 Induces Alternative Cytotoxic Processes in Tumor Cells via TNFR1
553   Receptor. *JOURNAL OF BIOLOGICAL CHEMISTRY* **290**:21724-21731. DOI 10.1074/jbc.M115.639732

554   **You, E., Huh, Y.H., Kwon, A., Kim, S.H., Chae, I.H., Lee, O.J., Ryu, J.H., Park, M.H., Kim, G.E., Lee, J.S.,**
555   **Lee, K.H., Lee, Y.S., Kim, J.W., Rhee, S., and Song, W.K. 2017.** SPIN90 Depletion and Microtubule
556   Acetylation Mediate Stromal Fibroblast Activation in Breast Cancer Progression. *CANCER RESEARCH*
557   **77**:4710-4722. DOI: 10.1158/0008-5472.CAN-17-0657

558   **Zhou, M., Ye, Z., Gu, Y., Tian, B., Wu, B., and Li, J. 2015.** Genomic analysis of drug resistant pancreatic cancer
559   cell line by combining long   non-coding RNA and mRNA expression profiling. *Int J Clin Exp Pathol* **8**:38-52.

560

Figure 1

The flow chart of the development process of the COAD gene expression signature.

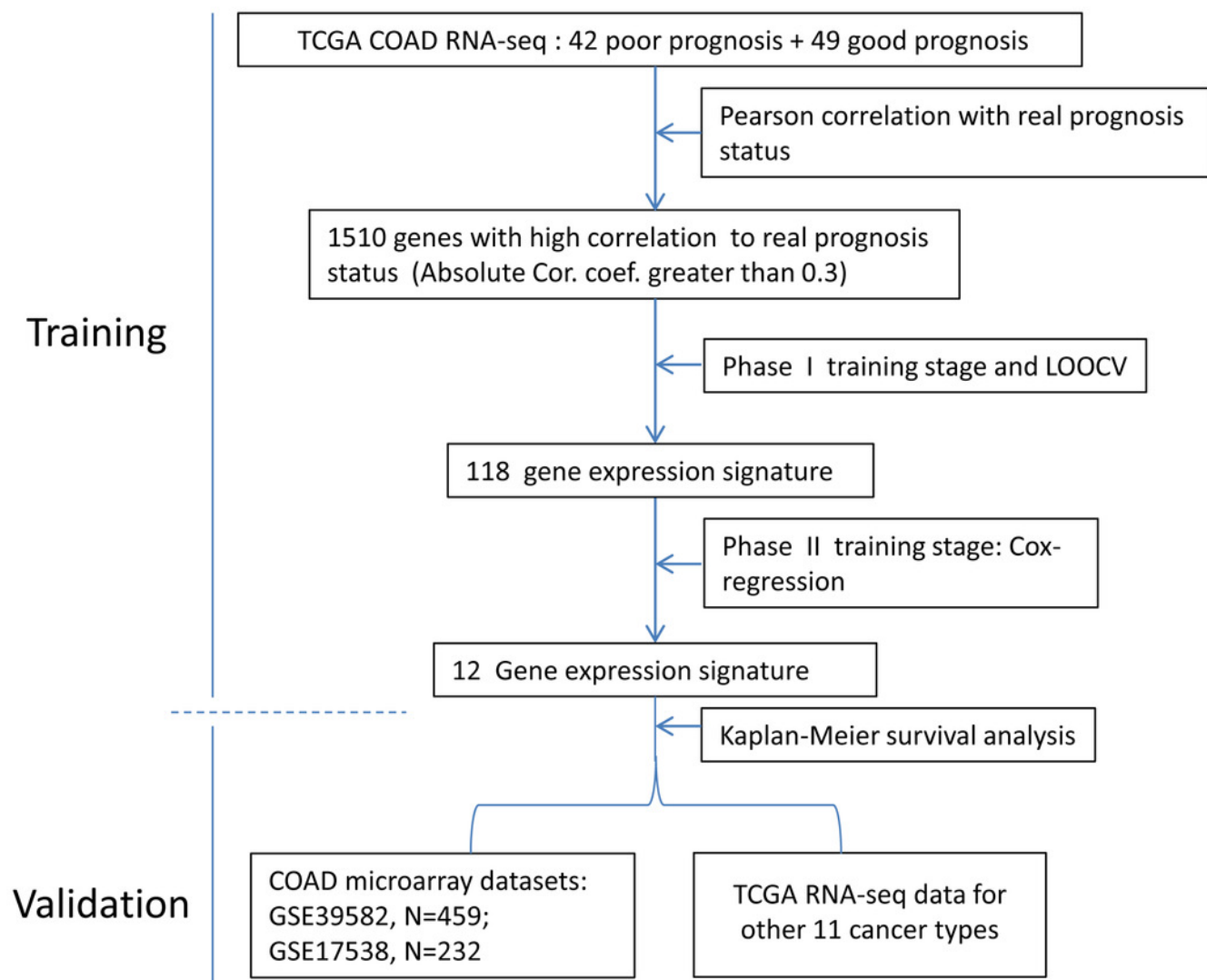



Figure 2

Prototype of the gene expression signature.

Expression heatmap plotting of 118 prognostic marker genes in training dataset (Up panel) and 20 patients in test dataset (Lower panel). Each row represents an observation (patient) and each column is a gene, whose name is labeled at the bottom. Tumors are ordered by the correlation to the average expression pattern of the good and poor prognosis group (Left panel). Genes are ordered by their correlation coefficients with the two prognosis categories. The real prognosis status for each tumor is displayed in the middle panel.

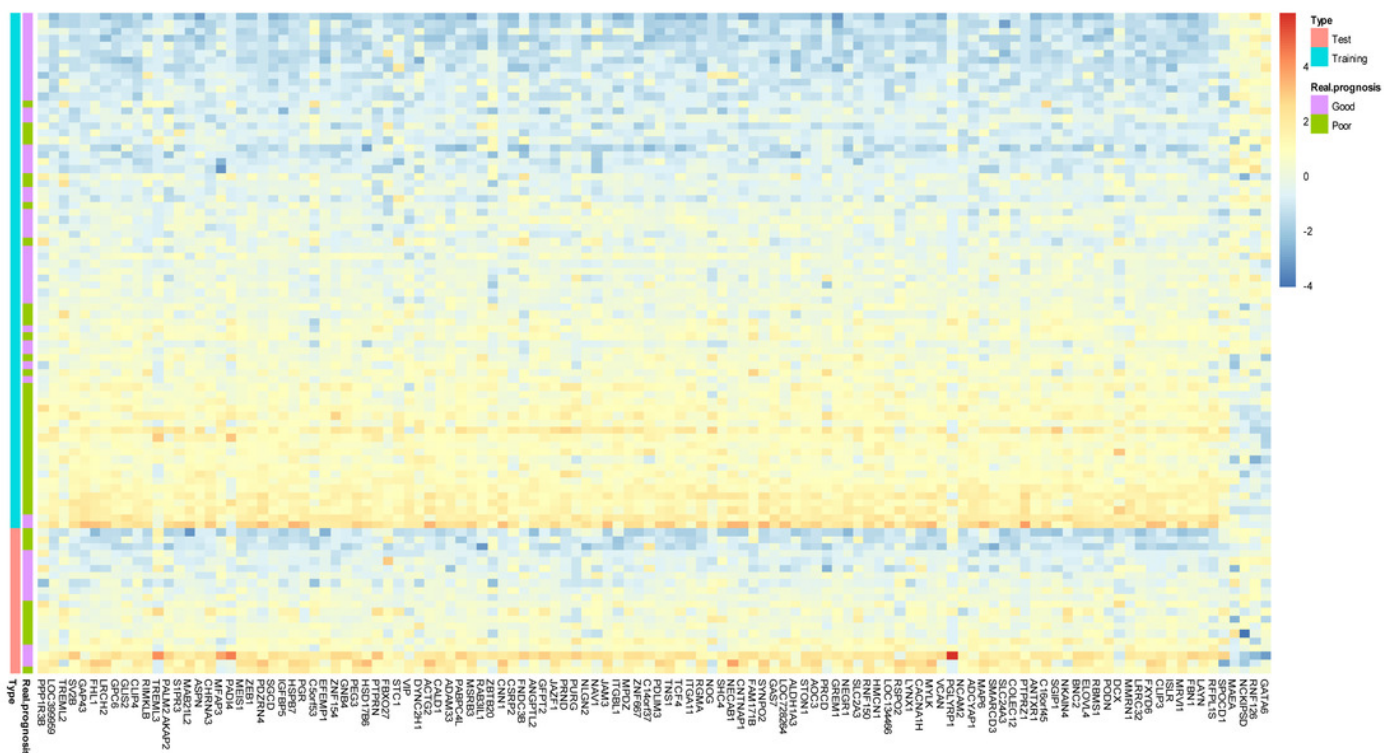



Figure 3

ROC plotting for the 25 signatures generated during phase I training process.

ROC with the information of risk-coefs and real risk of the 91 TCGA patients. ROC, receiver operating characteristic curve. TPR, true positive rate. FPR, false positive rate.

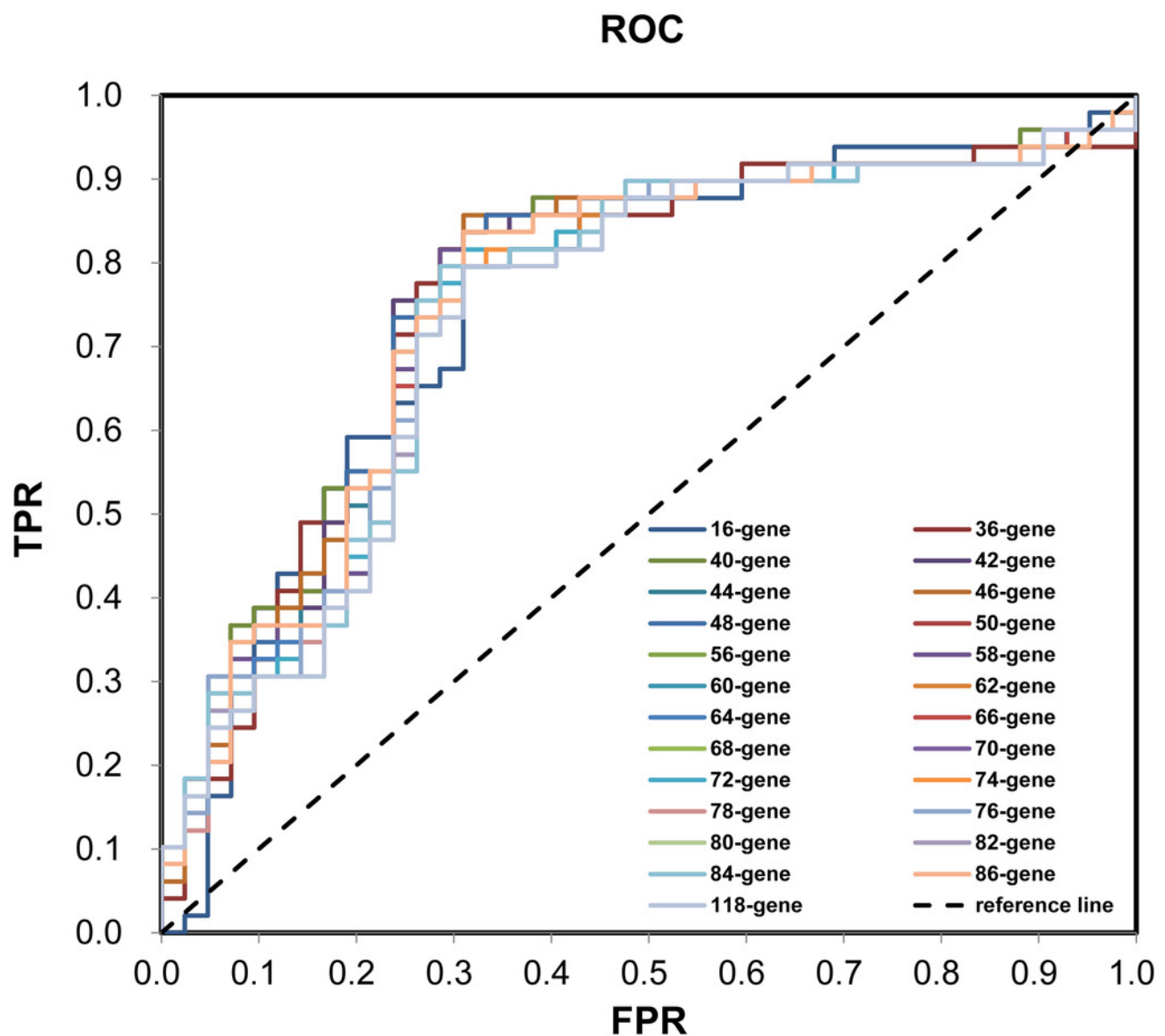



Figure 4

Prognostic values of the 12-gene signature.

Kaplan-Meier analysis of the high and low 12-gene risk score patients among TCGA COAD patients in training (a) and test dataset (b) in phase II training stage.

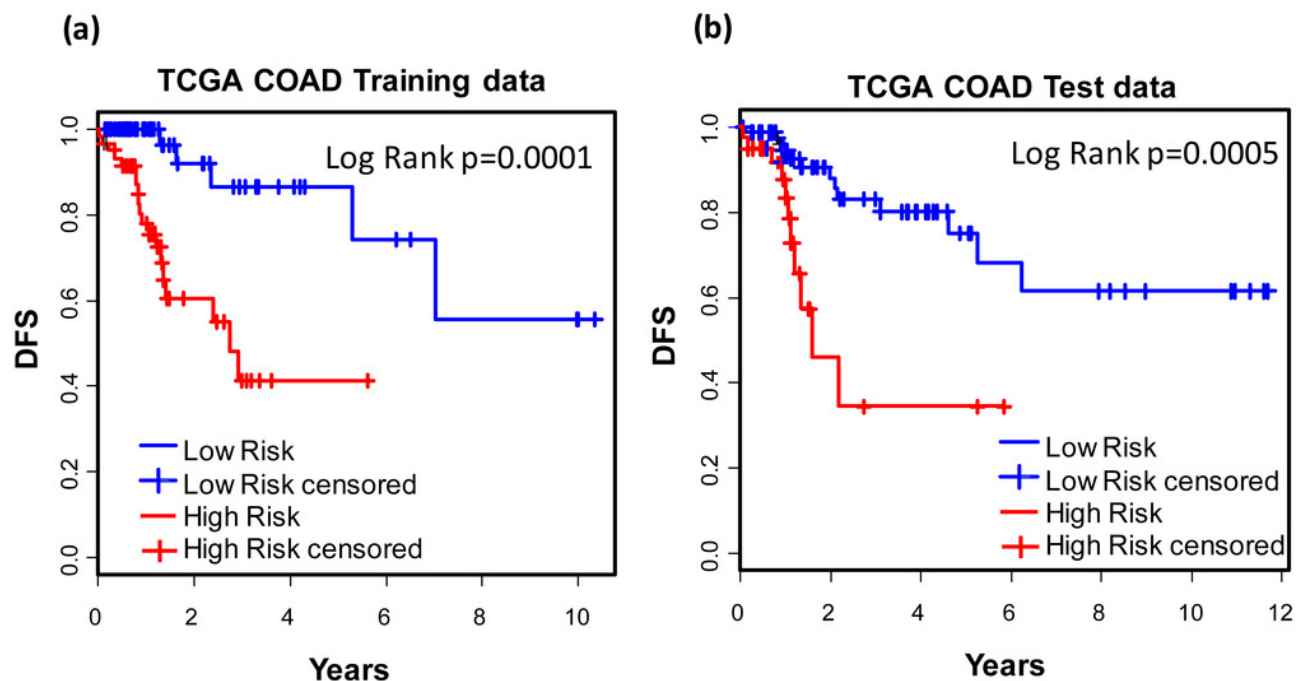



Figure 5

Prognostic values of the 12-gene signature in other COAD datasets.

(a)-(c): Kaplan-Meier curves showing patients (stageI-IV) with high and low12-gene risk score in endpoints of DSS, OS, and DFS, respectively; Kaplan-Meier curves showing patients at stage II (d) or II&III (e) with high and low 12-gene risk score in terms of DFS. DFS, disease free survival; DSS, disease specific survival; OS, overall survival.

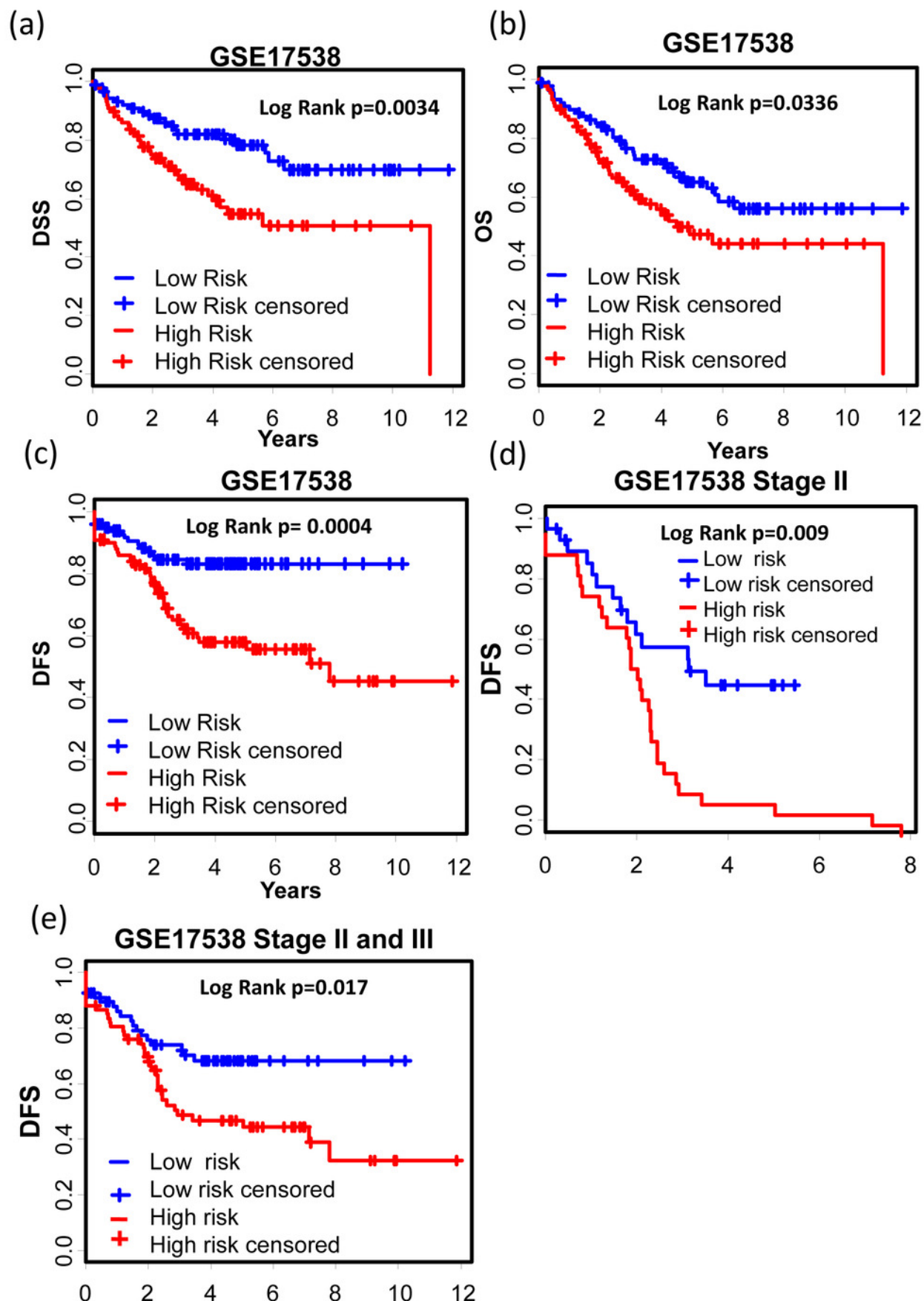



Figure 6

Prognostic values of the 12-gene signature in GSE39582.

(a) and (b): Kaplan-Meier curves showing patients (stageI-IV) with high and low 12-gene risk score in endpoints of RFS and OS, respectively; Kaplan-Meier curves showing stage III (c) or II&III (d) pMMR patients (treated with ACT) with high and low 12-gene risk score in respect to the endpoint of OS; (e) In stage II&III pMMR patients with low 12-gene scores, ACT subgroup displayed better OS outcome than control; (f) In stage II&III pMMR patients with high 12-gene scores, ACT and control group displayed no significant difference in the outcome of OS. RFS: relapse-free survival; OS: overall survival; pMMR: proficient mismatch repair system.

(a) GSE39582 pMMR

Log Rank p=0.0224

RFS

- Low Risk
- Low Risk censored
- High Risk
- High Risk censored

Years

(b) GSE39582 pMMR

Log Rank p=0.0052

OS

- Low Risk
- Low Risk censored
- High Risk
- High Risk censored

Years

(c) GSE39582 Stage III pMMR patients with ACT

Log Rank p=0.046

OS

- Low Risk
- Low Risk censored
- High Risk
- High Risk censored

Years

(d) GSE39582 Stage II&III pMMR patients with ACT

Log Rank p=0.041

OS

- Low Risk
- Low Risk censored
- High Risk
- High Risk censored

Years

(e) GSE39582 Stage II&III pMMR patients with low 12-gene score

Log Rank p=0.021*

OS

- With ACT
- With ACT censored
- Without ACT
- Without ACT censored

Years

ACT	Deaths	N	Death %
No	35	96	36.5
Yes	17	90	18.9

(f) GSE39582 Stage II &III pMMR patients with high 12-gene score

Log Rank p=0.117

OS

- With ACT
- With ACT censored
- Without ACT
- Without ACT censored

Years

ACT	Deaths	N	Death %
No	34	92	37.0
Yes	30	93	32.3

Figure 7

KM analysis of the high and low 12-gene risk score patients for the major outcomes in other cancer types.

(a) OS in pan-kidney cohort (KIPAN), (b) DFS in kidney renal clear cell carcinoma (KIRC). (c) & (d) DFS and OS in kidney renal papillary cell carcinoma (KIRP), respectively. (e) DFS in lung squamous cell carcinoma (LUSC). (f) DFS in skin cutaneous melanoma (SKCM). OS, overall survival. DFS, disease free survival.

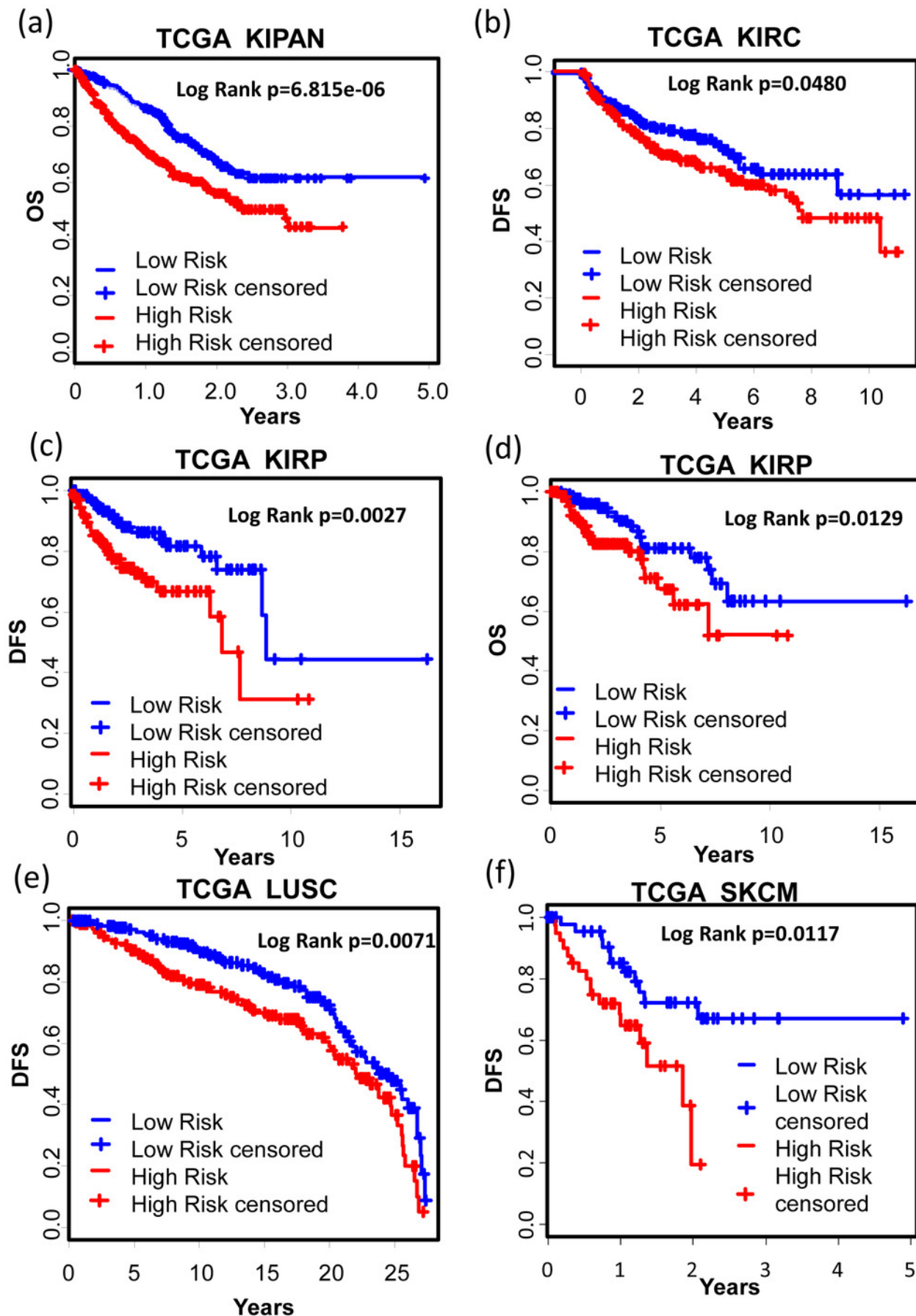
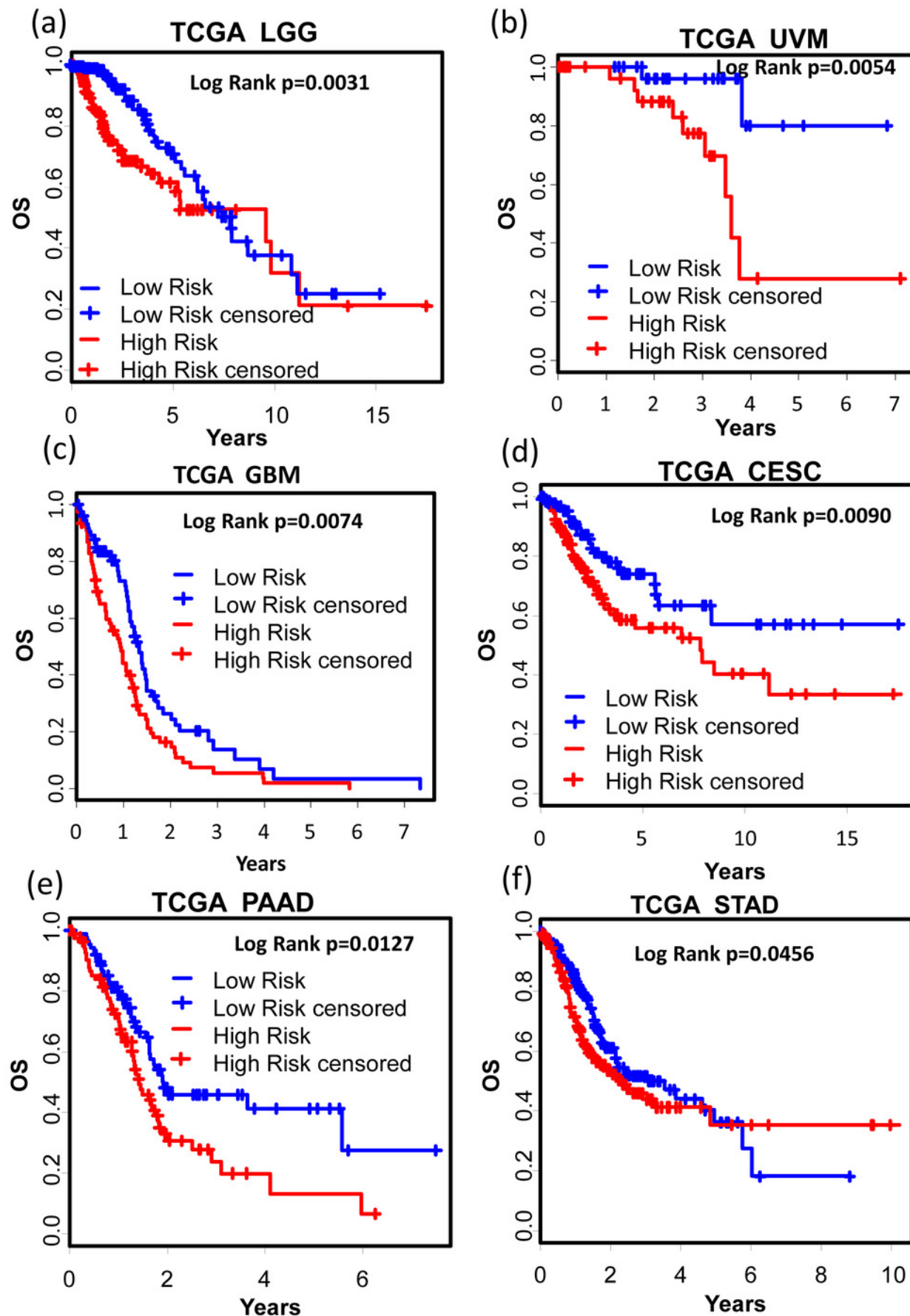



Figure 8

Kaplan-Meier analysis of the high and low 12-gene risk score patients for the major outcomes in other cancer types.

(a) OS in brain lower grade glioma (LGG). (b) OS in uveal melanoma (UVM). (c) OS in glioblastoma (GBM). (d) OS in cervical and endocervical cancers (CESC). (e) OS in pancreatic adenocarcinoma (PAAD). (f) OS in stomach adenocarcinoma (STAD).OS, overall survival.

Clinicopathologic features of 240 TCGA COAD patients.

Characteristic	Training set (N=119) No. of patients (%)	Testing set (N=121) No. of patients (%)	*p* value
Age (mean±SD)	66.4±13.0	63.2±13.8	0.069[†]
Gender			
Male	60 (50.4%)	54 (44.6%)	0.438[‡]
Female	59 (49.6%)	67 (55.4%)	
Stage			
I	20 (16.8%)	20 (16.5%)	
II	47 (39.5%)	48 (39.7%)	0.998[§]
III and IV	52 (43.7%)	53 (43.8%)	
Primary tumor			
T1 and T2	20 (16.8%)	23 (19.0%)	0.737[‡]
T3 and T4	99 (83.2%)	98 (81.0%)	
Microsatellite status			
MSI-L	23 (19.3%)	23 (19.0%)	
MSI-H	20 (16.8%)	20 (16.5%)	0.995[§]
MSS	76 (63.9%)	78 (64.4%)	
Lymphatic_invasion			
No	77 (64.7%)	77 (63.6%)	0.999[‡]
Yes	33 (27.7%)	34 (28.1%)	
Unknown	*9 (7.6%)*	*10 (8.3%)*	Excluded

1   [†]t test.

2   [‡]Fisher's exact test.

3   [§]Chi-squared test.

Statistics of the ROC analysis.

Signature	AUC	SE	Progressive *p*	Progressive 95% CIs	
				Lower bound	Upper bound
16-gene	0.7517	0.0531	0.0000	0.6476	0.8558
36-gene	0.7600	0.0529	0.0000	0.6562	0.8637
40-gene	0.7653	0.0520	0.0000	0.6634	0.8672
42-gene	0.7609	0.0525	0.0000	0.6581	0.8638
44-gene	0.7604	0.0525	0.0000	0.6576	0.8633
46-gene	0.7614	0.0524	0.0000	0.6588	0.8641
48-gene	0.7575	0.0528	0.0000	0.6540	0.8610
50-gene	0.7541	0.0530	0.0000	0.6503	0.8580
56-gene	0.7493	0.0532	0.0000	0.6450	0.8536
58-gene	0.7488	0.0533	0.0000	0.6444	0.8532
60-gene	0.7483	0.0532	0.0000	0.6439	0.8527
62-gene	0.7478	0.0533	0.0000	0.6433	0.8524
64-gene	0.7468	0.0534	0.0001	0.6421	0.8515
66-gene	0.7459	0.0535	0.0001	0.6409	0.8508
68-gene	0.7449	0.0534	0.0001	0.6402	0.8496
70-gene	0.7459	0.0534	0.0001	0.6412	0.8505
72-gene	0.7468	0.0534	0.0001	0.6422	0.8514
74-gene	0.7444	0.0535	0.0001	0.6395	0.8493
76-gene	0.7444	0.0535	0.0001	0.6396	0.8492
78-gene	0.7430	0.0537	0.0001	0.6377	0.8482
80-gene	0.7410	0.0539	0.0001	0.6354	0.8467
82-gene	0.7420	0.0538	0.0001	0.6365	0.8474
84-gene	0.7410	0.0539	0.0001	0.6353	0.8467
86-gene	0.7410	0.0539	0.0001	0.6354	0.8466
118-gene	0.7347	0.0542	0.0001	0.6284	0.8410

Note. AUC: Area-Under-Curve;

SE: Standard Error.

95% CIs: 95% Confidence Intervals.

Distribution of advanced stage patients between high- and low-score group. Fisher's exact test was used for statistical analysis.

Dataset	Group	Stage I&II	Stage III&IV	*p* value
GSE17538	High score group	19 (20%)	78 (80%)	0.0003
	Low score group	43 (44%)	54 (56%)	
TCGA	High score group	53 (49%)	55 (51%)	0.0277
	Low score group	69 (64%)	38 (36%)	