第二十一屆旺宏科學獎

成果報告書

參賽編號：SA21-020

作品名稱：

Identifying a robust lncRNA signature for predicting stage of colon adenocarcinoma using an evolutionary learning method

姓名：何彥霖

關 鍵 字 ：lncRNA; signature; colon adenocarcinoma; evolutionary learning

Identifying a robust lncRNA signature for predicting stage of colon adenocarcinoma using an evolutionary learning method

Abstract

Background: Long non-coding RNAs (lncRNAs) are newly identified as potential biological and gene regulators, which are promising biomarkers for cancer diagnosis and prognosis. There are few signatures consisting of a small set of biomarkers for modeling and predicting stage of colon adenocarcinoma (COAD) using lncRNA profiles.

Method: The dataset of 521 COAD patients that contains patient's clinical information and expression profiles of 14,048 lncRNAs from the dataset of The Cancer Genome Atlas was adopted to identify a lncRNA signature using the proposed evolutionary learning method EL-COAD. A dataset gse17536 of Gene Expression Omnibus was additionally used for the signature confirmation. EL-COAD uses an inheritable bi-objective combinatorial genetic algorithm with support vector machine (SVM) for identifying the signature while maximizing the accuracy of predicting the early and advanced stages of COAD. 14 commonly-used prediction models such as Random forest, Sequential minimal optimization (SMO), and Logistic regression were used to evaluate the identified signature. EL-COAD also identified a survival signature with a Cox proportion hazard regression (Cox-PH) model for predicting personal survival proportion.

Results: EL-COAD identified a stage signature of 15 lncRNAs and achieved accuracies and area under receiver operating characteristic curve of 79.4% and 0.792 in terms of 5-fold cross-validation, respectively. The signature with mean test accuracy of 63.05±2.73% was significantly better than the set of 15 top-ranked lncRNAs (59.12±3.08%) using 14 prediction models with a p-value 0.002. The signature with four clinical features (age, aneuploidy score, microsatellite instability status, and microsatellite instability score) using the SMO model can advance the test accuracy from 64.15% to 73.68%. The top-5 ranked lncRNAs were TMEM105, DUXAP8, APCDD1L-DT, PCAT6, and the novel transcript ENSG00000226308. Both KEGG pathway and Disease Ontology (DO) analysis supported the robust signature and that ENSG00000226308 is a promising biomarker. EL-COAD also identified the survival signature of 20 lncRNAs and the Cox-PH model achieved the C-index of 77.03% in terms of 10-fold cross-validation.

Conclusions: This study used an evolutionary learning method to identify the first stage signature of 15 lncRNAs and a survival signature of 20 lncRNA for predicting the stage and survival proportion of patients with COAD.

Introduction

Colorectal cancer (CRC) is the fourth most prevalent cancer in the United States and the second primary cause of cancer-related death [1], and the third highest incidence of all types of cancer and the second highest mortality rate worldwide [2]. CRC incidence and mortality have declined significantly due to improvements in cancer prevention, screening diagnosis, treatment modalities, etc. [3]. Despite this, the prognosis for patients with advanced colon cancer remains poor [4], and 90% of whom have colon adenocarcinoma (COAD) [5]. Stage at diagnosis is highly predictive of cancer mortality, and also effects gene therapy [6]. Early stage detection and diagnosis of cancer remains a challenge for clinicians. Therefore, it is of great practical significance to improve the prognosis of COAD patients through effective prognostic stratification by establishing a stage prediction model.

A review article for mRNA and non-coding RNAs for the diagnosis and prognosis of CRC from the body fluid to tissue level has been reported [7]. Most RNA transcripts are non-coding and only 2% of the genome encodes proteins. mRNAs are single-stranded ribonucleic acid molecules transcribed from a DNA strand as a template, carrying genetic information, and guiding protein synthesis. The related gene expression detected in platelets of CRC patients can also be used for the diagnosis of CRC. A set of five mRNAs establishing a simple formulation could be used for the postoperative treatment of CRC patients [8]. A novel five-gene signature as a prognostic and diagnostic biomarker was proposed for predicting survival in CRC [9]. Non-coding RNAs are important molecules that regulate the expression of genes at different stages such as the epigenetic, transcription, and post-transcription levels. MicroRNAs (miRNAs) have attracted interest as biomarkers due to their critical roles in cancer development and prognosis. miRNA dysregulation is observed in multiple types of cancers [10]. Extensive research has been conducted on miRNAs as clinically relevant biomarkers for CRC [7]. A set of nine key miRNAs related to the survival rate of COAD patients was reported [11]. The authors used a deep learning algorithm with miR-133b and its target genes to predict survival in patients with COAD via multi-omics data integration [12].

Long non-coding RNA (lncRNA) are special non-coding RNA molecules of more than 200 nucleotides in length. Numerous studies have shown that there are some potential relationships between the abnormal expression of lncRNAs and the occurrence of cancer [13-16]. lncRNA deregulation has associated with the development and progression of various cancer types, which makes lncRNA suitable as biomarkers for cancer diagnosis and prognosis

[17]. The detection of cancer-associated lncRNAs has proven to be a particularly valuable method for effective cancer diagnosis [18, 19]. Dysregulation of lncRNAs GAS8-AS [20], H19 [21], NEAT1 [22], and SNHG6 [23] was extensively well-studied and has been demonstrated to contribute to tumorigenesis and poor prognosis of COAD. Furthermore, overexpression of lncRNA SNHG1 [24] has been shown to promote epithelial-mesenchymal transition by binding to miR-497 and miR-195-5p in COAD cells. LINC00312 [25] and BCYRN1 [26] have been shown to play an important regulatory role in CRC cell proliferation and metastasis invasion.

Signature is well recognized as a small set of biomarkers for establishing models and predicting the stage, survival, recurrence, prognosis, etc. [27-30]. To our best knowledge, Table 1 shows existing lncRNA signature identification studies for patients with colon, CRC and COAD. From Table 1, all signature identification methods were published for survival prediction, including statistic method, univariate CoxPH analysis, and LASSO. The statistic and univariate CoxPH analysis methods identified a set of individual lncRNAs. LASSO with Cox regression was a well-recognized effective method for automatically identifying a set of biomarkers and establishing a model to predict survival. Due to the interaction among individual biomarkers, an inheritable bi-objective combinatorial genetic algorithm (IBCGA) was proposed for identifying a miRNA signature instead of a set of individual biomarkers to predict stage of breast cancer [27] and hepatocellular carcinoma [28], and survival of ovarian cancer [29] and bladder urothelial carcinoma [30]. Most studies used The Cancer Genome Atlas (TCGA) to establish prediction models. For validating the prediction model, the Gene Expression Omnibus (GEO) dataset was used as an independent cohort [31]. Few studies have reported lncRNA signatures to predict stage in patients with CRC or COAD due to insufficient samples of lncRNA expression profiles and interaction among numerous lncRNAs in using machine learning approaches.

This study proposed an evolutionary learning method called EL-COAD for identifying the lncRNA signature to predict the early and advanced stages of COAD. The lncRNA expression profiles of 521 and 177 COAD patients were obtained from the TCGA and GEO databases, respectively. EL-COAD is based on the optimal feature selection method IBCGA for coping with the strong interaction among lncRNAs. A dataset gse17536 of GEO was used for the signature confirmation. The biological significance of the identified lncRNA signature was analyzed for supporting the identified signature using Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and Disease Ontology (DO) annotations. For advancing

prediction accuracy, the effective features of clinical information were identified and added into the signature for establishing prediction model of COAD stage. Furthermore, EL-COAD also identified the survival signature using a Cox proportion hazard regression (Cox-PH) model for predicting personal survival proportion. All the top-ranked lncRNA biomarkers and risk factors were investigated and discussed.

Results and Discussion

Identification of a lncRNA signature associated with stage of COAD

The flowchart of the proposed method EL-COAD and the signature analysis is depicted in Fig. 1. First, we attempted to predict the stage of patients with COAD using lncRNA expression profiles. A dataset TCGA-COAD containing 56-lncRNA expression profiles of 506 patients with COAD and clinical information was obtained after preprocessing. EL-COAD with the optimal feature selection algorithm IBCGA identified a signature consisting of m=15 lncRNAs and achieved accuracies and area under receiver operating characteristic curve (AUC) of 79.4% and 0.792, respectively, in terms of 5-fold cross-validation (5-CV). We ranked lncRNAs using the main effect difference (MED) score and their relationship with cancers published, shown in Table 2. The SVM model with the 15-lncRNA signature can score the stage of COAD. The distribution of sample scores in the training set of TCGA-COAD (n=354) obtained using the classification probability of SVM was shown in Fig. 2. The prediction score can be used for quantifying the COAD stage and effectiveness of gene therapy. We computed the correlation among 15 lncRNAs in the signature using the Pearson correlation coefficient. The highest correlation coefficient of lncRNA pairs in the signature was 0.481. The result strongly suggested that the identified signature was effective at predicting the stage of patients with COAD.

Performance evaluation of signatures and prediction models

The effective 15-lncRNA signature obtained relied on the simultaneous optimization of the feature selection of IBCGA and parameter settings of SVM. For evaluating the effectiveness of IBCGA, we compared with the commonly-used statistic method of selecting the most differentially expressed lncRNAs of distinguishing stage in terms of p-value. For evaluating the model dependence of the identified signature, 14 commonly-used models were used including sequential minimal optimization (SMO, a fast SVM), Logistic, naïve Bayes, SVM, and random forest, etc.

Table 3 shows the comparison of the EL-COAD signature and the set of 15 p-value

lncRNAs using 14 models of Waikato Environment for Knowledge Analysis (Weka) and the TCGA-COAD dataset. The signature with mean test accuracy of 63.05±2.73% was significantly better than the set of 15 top-ranked lncRNAs (59.12±3.08%) using the 14 prediction models with a p-value 0.002. The signature of EL-COAD also performed well in the robustness with a small standard deviation of 2.73% revealing that the signature was more effective in model independence. The result was agreed with previous work that the IBCGA-based signature was better than Ranker search and correlation attribute evaluation method of Weka [28].

The identified 15-lncRNA signature was further validated using the second dataset of gse17536 from the GEO database. Because there were only 10 of 15 lncRNAs available, we used the following 10 lncRNAs to establish prediction models: TMEM105, APCDD1L-DT, PCAT6, PINK1-AS, BAIAP2-DT, LEMD1-AS1, H19, RAMP2-AS1, SNHG32, DLG3-AS1. The prediction model of using SMO and the 10-lncRNA signature achieved the 5-CV and test accuracy of 61.29% and 64.15%, respectively. The EL-COAD derived signatures performed equally well in the two independent datasets.

Prioritizing the lncRNA signature

The larger MED score indicates the higher contribution towards the prediction accuracy. The ranking of the MED score considers the interaction among lncRNAs instead of p-value of individual lncRNAs without considering interaction. From Table 2, the lncRNA signature contains 15 lncRNAs in order of decreasing MED scores, TMEM105, DUXAP8, APCDD1L-DT, PCAT6, ENSG00000226308, PINK1-AS, BAIAP2-DT, LINC02474, LEMD1-AS1, H19, RAMP2-AS1, SNHG32, CALML3-AS1, DLG3-AS1, and H1-10-AS1.

There were 13 lncRNAs associated with cancers from the published work, including 3 lncRNAs of colon cancer, 3 lncRNAs of CRC, and 7 lncRNAs of other cancer types. Note that the rank-5 lncRNA ENSG00000226308 is a novel transcript with a p-value 2.19E-6 (Fig. 3) revealing the significance in classifying the early and advanced stage. From the human gene database GeneCards [52], the top-five phenotypes according to the best scores with gene relation were high density lipoprotein cholesterol measurement, apolipoprotein A1 measurement, moderate albuminuria, sex hormone-binding globulin measurement, and colorectal cancer. The five phenotypes highly related to the stage of developing COAD reveals that ENSG00000226308 is a promising biomarker of predicting COAD stage.

Difference of expression profiles between early and advanced stage groups

We measured expression of the 15 lncRNAs in early stage and advanced stage groups that they all have significant difference with p-value <0.001 using the training set of TCGA-COAD. The box-plots representing expression difference in the early stage (low risk) and advanced stage (high risk) groups and p-value of the 15 lncRNAs in the identified signature are showing in Fig. 3. Of the 15 lncRNAs, the mean expression values of TMEM105, DUXAP8, APCDD1L-DT, PCAT6, ENSG00000226308, PINK1-AS, BAIAP2-DT, LINC02474, LEMD1-AS1, H19, RAMP2-AS1, SNHG32, CALML3-AS1, DLG3-AS1, and H1-10-AS1 are 0.45 ± 0.36, 0.25 ± 0.40, 0.04 ± 0.09, 2.00 ± 1.21, 0.06 ± 0.11, 1.59 ± 0.98, 4.15 ± 2.14, 0.56 ± 1.59, 0.05 ± 0.14, 0.19 ± 0.23, 48.45 ± 25.56, 0.04 ± 0.04, 0.68 ± 1.45, and 0.18 ± 0.11, respectively, in the early stage group, and 0.66 ± 0.50, 0.50 ± 0.86, 0.05 ± 0.07, 2.79 ± 1.94, 0.15 ± 0.39, 1.31 ± 0.57, 4.97 ± 2.59, 0.94 ± 2.08, 0.08 ± 0.14, 24.84 ± 73.14, 0.29 ± 0.40, 57.14 ± 28.18, 0.06 ± 0.06, 0.43 ± 0.61, and 0.22 ± 0.17, respectively, in the advanced stage group.

Among the 10-lncRNA signature, 4 lncRNAs including, PCAT6 ($p = 0.018$), PINK1-AS ($p = 0.019$), RAMP2-AS1 ($p = 0.032$), and SNHG32 ($p = 0.002$) were significantly associated with stage of patients in the gse17536 dataset. The sample number of gse17536 (n=177) was far smaller than that of TCGA-COAD (n=506) and the expression profiles were not measured using the same way in the two datasets. The lncRNAs served as biomarkers of distinguishing stage needs further validation.

Pathway analysis of the identified lncRNA signature

We performed the lncRNA–RNA interaction analysis on 15 characteristic lncRNAs in the prognostic signature by the ENCORI database [53] and using LncRNA2Target v3.0 [54] to find other target gene. The lncRNA–RNA interaction network was constructed, consisting of 8 lncRNAs and 330 RNAs, shown in Fig. 4. Note that there were 7 lncRNAs without targeted RNAs reported in the two databases. The three lncRNAs, PINK1-AS, H19 and LINC02474, had a large number of target RNAs. To explore potential functions of these lncRNAs, target RNAs were annotated by the Metascape database [55] and String database [56]. The protein-protein interactions of target genes obtained using the String database were shown in Fig. 5. The top-ranked KEGG pathways significantly enriched by using Metascape were shown in Table 4. The target genes were involved in the pathways in the p-value order: 1) MicroRNAs in cancer, 2) Endocrine resistance, 3) Human papillomavirus infection, 4) Pathways in cancer, 5) Colorectal cancer, and 6) Transcriptional misregulation in cancer. These pathways were highly associated with the stage of COAD and supported the identified signature of predicting the

COAD stage.

Functional annotations of the lncRNA signature

The disease ontology (DO) plays a key role in disease knowledge organization, representation, and standardization, serving as a reference framework for multiscale biomedical data integration and analysis across thousands of clinical, biomedical and computational research projects and genomic resources around the world. The top-10 of 43 diseases using DO analysis of the String database according to the strength were shown in Table 5. From Table 5, The 3rd and 10th diseases were Colon cancer and Colorectal cancer, respectively. Furthermore, most diseases had relationship with cancers, revealing that the 8 lncRNAs were significantly associated with the COAD stage.

Roles of the top ranked lncRNAs

The roles of the top-10 ranked lncRNAs in COAD were analyzed using experimentally validated literature.

1) TMEM105: The lncRNA is a novel transcript associated with COAD. The high expression of TMEM105 predicted poor prognoses of gastric cancer by univariate Cox regression analysis (hazard ratio, HR>1) [37]. Through a series of bioinformatics analyses, TMEM105 could serve as prognostic and diagnostic biomarkers for patients with breast infiltrating duct and lobular carcinoma [38]. The top-1 lncRNA has the potential to be a biomarker to predict the stage of COAD.

2) DUXAP8: The expression level was upregulated in bladder cancer tissues, and it was in a positive correlation with the TNM stage and tumor size, but negatively correlated with the total survival time [57]. DUXAP8 promotes pancreatic carcinoma cell migration and invasion via pathway by miR-448/WTAP/Fak signaling axis [58]. DUXAP8 may serve as a candidate biomarker and represent a novel therapeutic target of pancreatic cancer [59].

3) APCDD1L-DT: APCDD1L-AS1 is the aliases of APCDD1L-DT. APCDD1L-AS1 was able to inhibit the progression of clear cell renal cell carcinoma, and its decreased expression could be caused by DNA hypermethylation and loss of VHL protein expression. Therefore, APCDD1L-AS1 may serve as a new therapeutic target in the treatment of clear cell renal cell carcinoma [60].

4) PCAT6: PCAT6 is a member of the Prostate Cancer Associated Transcripts family of molecules. PCAT6 is highly expressed in gastric cancer, colon cancer, hepatocellular

carcinoma, lung cancer, bladder cancer, ovarian cancer, breast cancer, cervical cancer, osteosarcoma, glioblastoma, and other tumors [61]. PCAT6 functions as an oncogene and may serve as a potential new prognostic biomarker in these tumors [61].

5) ENSG00000226308: The top-five phenotypes according to the best scores with gene relation were high density lipoprotein cholesterol measurement, apolipoprotein A1 measurement, moderate albuminuria, sex hormone-binding globulin measurement, and colorectal cancer [52]. There is few experimentally validated literature of ENSG00000226308. The five phenotypes related to the stage of developing COAD suggests that ENSG00000226308 is a promising biomarker of predicting COAD stage.

6) PINK1-AS: PINK1 Antisense RNA is affiliated with the lncRNA class. Based on the altered expression of PINK1-AS in the peripheral blood of multiple sclerosis patients, PINK1-AS might be a putative culpript in the pathogenesis of multiple sclerosis [62].

7) BAIAP2-DT: BAIAP2 Divergent Transcript (BAIAP2-DT) is an autophagy-related lncRNA. E2F1-induced lncRNA BAIAP2-AS1 overexpression contributes to the malignant progression of hepatocellular carcinoma via miR-361-3p/SOX4 Axis [44].

8) LINC02474: Long Intergenic Non-Protein Coding RNA 2474 (LINC02474) affects metastasis and apoptosis of colorectal cancer by inhibiting the expression of GZMB [45].

9) LEMD1-AS1: LEMD1-AS1 suppresses ovarian cancer progression through sponging miR-183-5p and regulation of TP53, suggesting a novel biomarker and target for ovarian cancer [46].

10) H19: H19 usually controls gene expression by acting as a microRNA sponge, or through mir-675, or by leading various protein complexes to genes at the chromosome level [63].

Stage prediction using the signature with clinical features

In the TCGA-COAD dataset, there were seven clinical features available, namely Diagnosis Age, Aneuploidy Score, MSI MANTIS Score, MSIsensor Score, Mutation Count, Winter Hypoxia Score, and Buffa Hypoxia Score. We used IBCGA with SVM to select a set of clinical features from the seven features by maximizing the prediction accuracy of 5-CV. There were four features selected as follows: Diagnosis Age, Aneuploidy Score, MANTIS Score, and MSIsensor Score. The aneuploidy score for each tumor is calculated as the sum of altered arms, within a range of 0 to 39 [64]. The MSI MANTIS Score is used to predict the

microsatellite instability status. Microsatellite instability-high CRC had a better immunotherapy prognosis than Microsatellite instability-low CRC [65]. The resulting MSIsensor score is a value between 0 and 100 that corresponds to the percentage of mutated microsatellite loci [66]. The signature SigE with the four clinical features using the SMO model can significantly advance the test accuracy from 64.15% to 73.68%. The results suggest that the personalized model using the signature and informative clinical features is helpful in precision medicine.

Identification of survival features

The proposed method EL-COAD can identify nor only stage signature but also survival signature. Using the TCGA-COAD dataset, EL-COAD identify 20 from 54 lncRNAs as a survival signature using the Cox-PH model. The 20-lncRNA signature achieved the C-index of 0.7823 and 0.7703 for the training and 10-CV accuracies, respectively. The relatively high accuracy confirmed that EL-COAD can identify both stage and survival signatures without The 20 lncRNAs in the identified survival signature are listed in Table 6. The lncRNAs with the negative values of MED and the ones with the low rank denotes the less contribution toward the prediction and are not stable due to the underdetermined problem resulting from the insufficient profiles used. Since the lncRNAs are newly identified as potential biological and gene regulators, which are promising biomarkers for cancer diagnosis and prognosis, some top-ranked lncRNAs were uncharacterized in literature. However, we presented the Kaplan–Meier survival curves of the top-five lncRNAs in Fig. 6. The low expression of these lncRNAs has a good survival of patients with COAD.

Conclusions

Numerous studies have shown that there are some potential relationships between the abnormal expression of lncRNAs and the occurrence of cancer. Cancer stage at diagnosis is highly predictive of cancer mortality, survival, treatment, and effects gene therapy. Few studies have reported lncRNA signatures to predict stage in patients with CRC or COAD due to insufficient samples of lncRNA expression profiles and interaction among numerous lncRNAs in using machine learning approaches. Therefore, this study proposed EL-COAD cooperated with the feature selection method IBCGA to identify a lncRNA signature that can distinguish early stage from advanced stage of COAD.

EL-COAD proposed a first robust signature of 15 lncRNAs with derived models for scoring and predicting the COAD stage. KEGG pathway and DO term analyses has revealed the functional mechanisms of lncRNA signature in several cancer pathways and top-rank diseases. The biological analysis using KEGG, protein-protein interaction, DO term, has supported the model-independent signature in predicting stage of COAD. Furthermore, this study discovered a lncRNA ENSG00000226308 which is a novel transcript and is a promising biomarker of predicting COAD stage using the analysis of the GeneCards database. It is worthwhile to experimentally validate the biomarker ENSG00000226308.

The development of technologies for potential identification of lncRNAs and their role in cancer are important for COAD diagnostics and therapeutics. The identified lncRNAs in this study could aid in the development of lncRNA-based targeted cancer therapies in COAD patients. Together, our findings help to explore the role of lncRNAs in COAD and could facilitate early-stage detection and prevention.

Datasets and methods

Datasets

From the TCGA database, we retrieved a dataset (namely TCGA-COAD) containing lncRNA expression profiles of 521 patients with COAD and clinical information. Each profile has 14,048 lncRNAs. The patients with missing data were removed. Consequently, the 506 patients were divided into non-overlapping training and test sets according in a ratio of 7:3. The lncRNAs with p-value <0.001 were retained as candidate biomarkers using Wilcoxon rank sum test. The final dataset of 56-lncRNA expression profiles were used for signature identification. For the classification purpose, the samples of COAD were categorized into two groups: early stage (stages 1 and 2) and advanced stage (stages 3 and 4). There were 293 and 213 patients in the early-stage and advanced-stage groups, respectively. The clinical information plays an important role in predicting stage. Because of much missing data, there were seven informative and available clinical features which can be further investigated, namely Diagnosis Age, Aneuploidy Score, MSI MANTIS Score, MSIsensor Score, Mutation Count, Winter Hypoxia Score, and Buffa Hypoxia Score.

The COAD samples were also classified into two groups according to survival time. The short-term survival group contained patients whose survival was less than 3 years (n=89), and the long-term survival group consisted of patients who lived for more than 3 years (n=127).

The patients who are still alive and whose follow-up time is less than 3 years were classified as the uncertain group (n=302). In the lncRNA filtration process, we applied the Mann–Whitney U test to the training set to select 54 top-ranked lncRNAs by p-values < 0.001 using the short-term and long-term survival groups.

The GEO dataset gse17536 was used for additional validation of the identified signature which consists of 177 lncRNA expression profiles and clinical information of colorectal cancer. There were 81 and 96 patients in the early-stage and advanced-stage groups, respectively. The 177 patients were divided into non-overlapping training and test sets in a ratio of 7:3. The clinical information contained age, sex, stage, overall survival time, disease specific survival time, disease free survival time, overall survival event, disease specific survival event, and disease free survival event.

The proposed EL-COAD method

This study proposed an evolutionary learning method EL-COAD based on an optimal feature selection method IBCGA cooperated with SVM to identify a robust lncRNA signature capable of distinguishing early stage and advanced stage patients and establish models for predicting stage of COAD from lncRNA expression profiles. SVMs are powerful statistical learning algorithms that use non-linear transformation to map data from input space to higher-dimensional space to identify better predictive models. SVMs have become popular in the biomedical sciences, especially in cancer research, due to their potential predictive performance. This study used the LibSVM package, a library of SVM [67]. The lncRNA expression profiles of COAD patients are input data. SVM works implicitly by only computing the corresponding kernels in the feature space between two data points, x_i and x_j. The radial basis function (RBF) kernel function is defined as follows:

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|\right)^2 \quad (1)$$

The SVM parameters C and γ were optimized based on the fitness function of the used intelligent evolutionary algorithm [68] in terms of 5-CV.

The Cox proportional-hazards (Cox-PH) model [69] is the most common method to predict the risk score (i.e., log hazard ratio) and survival function. It represents the hazard function as the following form:

$$\mathrm{H}(\mathrm{t}|x_i) = H_0(t)\exp(\theta_i) \quad (2)$$

$$\theta_i = X_i^T \beta \quad (3)$$

where $H_0(t)$ is the baseline hazard function, θ_i is the log hazard ratio for patient *i*, and β is the model parameters to be estimated. The C-index was used to measure the performance of Cox-PH. The signature to be identified plays a crucial role in the performance of the personalized Cox-PH model.

IBCGA

The inheritable bi-objective combinatorial genetic algorithm (IBCGA) based an intelligent evolutionary algorithm that uses an orthogonal array crossover to solve large parameter optimization problems. In the optimization process, IBCGA selects a minimum number of features, in this study namely lncRNAs or clinical features, while maximizing its prediction performance [70]. In this study, we used the same IBCGA to identify the stage and survival signatures. The fitness function is the only guide of IBCGA to search for an optimal solution. For the identification of the stage signature, the fitness function is the prediction accuracy of the SVM model in terms of five-fold cross-validation (5-CV). For the survival signature, the fitness function is the C-index of the Cox-PH model using 5-CV. We used 56-lncRNA and 54-RNA expression profiles of 354 COAD patients for the identification of the stage and the survival signatures, respectively. IBCGA aims to identify a minimal number *m* from *k* lncRNAs while maximizing the prediction accuracy. IBCGA simultaneously optimizes the feature selection and parameter settings of the SVM/Cox-PH models. IBCGA's parameters were $r_{start} = 10$, $r_{end} = 50$, Npop = 50, and Gmax = 60. Some applications of IBCGA in identifying miRNA signatures can be referred to the work [27-30].

The major steps based on IBCGA are described as follows.

Step 1. (Initialization) Randomly generate the population of Npop individuals encoded by r 1′s and k-r 0′s in the chromosome, where $r = r_{end}$.

Step 2. (Evaluation) Evaluate the fitness value of all individuals using the fitness function.

Step 3. (Selection) Apply a conventional tournament selection method that selects the winner from two randomly selected individuals to generate a mating pool of Npop individuals.

Step 4. (Crossover) Select two parents from the mating pool to perform an orthogonal array crossover operation. The best two individuals among the parents and the generated individuals are selected as the children of the crossover.

Step 5. (Mutation) Apply a conventional mutation operator to the randomly selected individuals. To prevent the highest fitness value from deteriorating, mutation is not applied to the best individual.

Step 6. (Termination test) If the stopping condition of performing Gmax generations is satisfied, then output the best individual as the solution S_r. Otherwise, go to Step 2.

Step 7. (Inheritance) If r < rend, randomly change one bit in the binary genes for each individual from 1 to 0; decrease the number r by one, and go to Step 2.

Step 8. (Output signature) Let S_m be the best solution among the solutions S_r. Obtain a set of selected lncRNAs and parameters C and γ of SVM from the chromosome of S_m.

Weka classifiers

We used the Weka package [71], a powerful data mining tool that uses well-known machine learning algorithms. We compared the prediction performance of 14 machine learning methods such as Random Forest, Sequential minimal optimization (SMO), and Logistic regression. We performed 5-CV to evaluate the performance of the machine learning models. We evaluated the prediction performance of EL-COAD using the prediction accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

where TP is true positive; TN is true negative; FP is false positive; and FN is false negative.

Data Availability

All the data used in this analysis can be found on the TCGA data portal https://portal.gdc.cancer.gov/ and Gene Expression Omnibus https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62564.

Competing Interests

The authors declare that they have no competing interests.

Funding

The author declares that there are no sources of funding to be acknowledged.

References

[1] Benson A.B., Venook A.P., Al-Hawary M.M., Cederquist L., Chen Y.-J., Ciombor K.K. et al. (2018) NCCN guidelines insights: colon cancer, version 2.2018. J. Natl. Comprehensive Cancer Network 16, 359–369 10.6004/jnccn.2018.0021.

[2] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. (2018) Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 68(6):394–424.

[3] Edwards B.K., Ward E., Kohler B.A., Eheman C., Zauber A.G., Anderson R.N. et al. (2010) Annual report to the nation on the status of cancer, 1975-2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates. Cancer: Interdisciplinary Int. J. Am. Cancer Soc. 116, 544–573 10.1002/cncr.24760

[4] Anguraj S., Lyssiotis C.A., Krisztian H., Collisson E.A., Gibb W.J., Stephan W. et al. (2013) A colorectal cancer classification system that associates cellular phenotype and responses to therapy. Nat. Med. 19, 619–625

[5] Fatemeh H., Saeed A., Amir Mohammad K. and Mehdi E. (2014) Clinicopathological features of colon adenocarcinoma in Qazvin, Iran: a 16-year study. Asian Pacific J. Cancer Prevention APJCP 15, 951

[6] Madu CO, Lu Y. (2010) Novel diagnostic biomarkers for prostate cancer. J Cancer. 2010 Oct 6;1:150-77. doi: 10.7150/jca.1.150. PMID: 20975847; PMCID: PMC2962426.

[7] He, J., Wu, F., Han, Z., Hu, M., Lin, W., Li, Y., & Cao, M. (2021). Biomarkers (mRNAs and Non-Coding RNAs) for the Diagnosis and Prognosis of Colorectal Cancer - From the Body Fluid to Tissue Level. Frontiers in oncology, 11, 632834.

[8] Olsson L, Hammarstrom ML, Israelsson A, Lindmark G, Hammarstrom S. (2020) Allocating Colorectal Cancer Patients to Different Risk Categories by Using a Five-Biomarker mRNA Combination in Lymph Node Analysis. PloS One (2020) 15(2):e0229007. doi: 10.1371/journal.pone.0229007.

[9] Ghatak S, Mehrabi SF, Mehdawi LM, Satapathy SR, Sjölander A. (2022). Identification of a Novel Five-Gene Signature as a Prognostic and Diagnostic Biomarker in Colorectal Cancers. Int J Mol Sci. 2022 Jan 12;23(2):793. doi: 10.3390/ijms23020793. PMID: 35054980; PMCID: PMC8776147.

[10] Volinia, S., Calin, G. A., Liu, C. G., Ambs, S., Cimmino, A., Petrocca, F., Visone, R., Iorio, M., Roldo, C., Ferracin, M., Prueitt, R. L., Yanaihara, N., Lanza, G., Scarpa, A., Vecchione, A., Negrini, M., Harris, C. C., & Croce, C. M. (2006). A microRNA expression signature of human solid tumors defines cancer gene targets. Proceedings of the National Academy of Sciences of the United States of America, 103(7), 2257–2261.

[11] Zhu, J., Xu, Y., Liu, S., Qiao, L., Sun, J., & Zhao, Q. (2020). MicroRNAs Associated With Colon Cancer: New Potential Prognostic Markers and Targets for Therapy. Frontiers in bioengineering and biotechnology, 8, 176.

[12] Lv, J., Wang, J., Shang, X., Liu, F., & Guo, S. (2020). Survival prediction in patients with colon adenocarcinoma via multi-omics data integration using a deep learning algorithm. Bioscience reports, 40(12), BSR20201482. Advance online publication.

[13] Chen X, Yan GY. (2013). Novel human lncRNA-disease association inference based on lncRNA expression profiles. Bioinformatics. 2013 Oct 15;29(20):2617-24. doi: 10.1093/bioinformatics/btt426. Epub 2013 Sep 2. PMID: 24002109.

[14] Zhong, Y., Gao, D., He, S., Shuai, C., & Peng, S. (2016). Dysregulated Expression of Long Noncoding RNAs in Ovarian Cancer. International journal of gynecological cancer : official journal of the International Gynecological Cancer Society, 26(9), 1564–1570.

[15] Jiang, Y., Zhou, J., Zou, D., Hou, D., Zhang, H., Zhao, J., Li, L., Hu, J., Zhang, Y., & Jing, Z. (2019). Overexpression of Limb-Bud and Heart (LBH) promotes angiogenesis in human glioma via VEGFA-mediated ERK signalling under hypoxia. EBioMedicine, 48, 36 – 48.
[16] Esteller M. (2011). Non-coding RNAs in human disease. Nat Rev Genet. 2011 Nov 18;12(12):861-74. doi: 10.1038/nrg3074. PMID: 22094949.
[17] Bolha L, Ravnik-Glavač M, Glavač D. (2017). Long Noncoding RNAs as Biomarkers in Cancer. Dis Markers. 2017;2017:7243968. doi: 10.1155/2017/7243968. Epub 2017 May 29. PMID: 28634418; PMCID: PMC5467329.
[18] Chi, Y., Wang, D., Wang, J., Yu, W., & Yang, J. (2019). Long Non-Coding RNA in the Pathogenesis of Cancers. Cells, 8(9), 1015.
[19] Wang, P., Ning, S., Zhang, Y., Li, R., Ye, J., Zhao, Z., Zhi, H., Wang, T., Guo, Z., & Li, X. (2015). Identification of lncRNA-associated competing triplets reveals global patterns and prognostic markers for cancer. Nucleic acids research, 43(7), 3478–3489.
[20] Zhao Y, Chu Y, Sun J, Song R, Li Y, Xu F. (2019). LncRNA GAS8-AS inhibits colorectal cancer (CRC) cell proliferation by downregulating lncRNA AFAP1-AS1. Gene. 2019 Aug 20;710:140-144. doi: 10.1016/j.gene.2019.05.040. Epub 2019 May 24. PMID: 31132513.
[21] Li CF, Li YC, Wang Y, Sun LB. (2018). The Effect of LncRNA H19/miR-194-5p Axis on the Epithelial-Mesenchymal Transition of Colorectal Adenocarcinoma. Cell Physiol Biochem. 2018;50(1):196-213. doi: 10.1159/000493968. Epub 2018 Oct 2. PMID: 30278464.
[22] Zhang M, Weng W, Zhang Q, Wu Y, Ni S, Tan C, Xu M, Sun H, Liu C, Wei P. (2018). The lncRNA NEAT1 activates Wnt/β-catenin signaling and promotes colorectal cancer progression via interacting with DDX5. J Hematol Oncol. 2018;11(1):113.
[23] Xu M, Chen X, Lin K, Zeng K, Liu X, Xu X, Pan B, Xu T, Sun L, He B, Pan Y, Sun H, Wang S. (2019). lncRNA SNHG6 regulates EZH2 expression by sponging miR-26a/b and miR-214 in colorectal cancer. J Hematol Oncol. 2019 Jan 9;12(1):3. doi: 10.1186/s13045-018-0690-5. PMID: 30626446; PMCID: PMC6327409.
[24] Bai J, Xu J, Zhao J, Zhang R. (2019). lncRNA SNHG1 cooperated with miR-497/miR-195-5p to modify epithelial-mesenchymal transition underlying colorectal cancer exacerbation. J Cell Physiol. 2020 Feb;235(2):1453-1468. doi: 10.1002/jcp.29065. Epub 2019 Jul 5. PMID: 31276207.
[25] Li, G., Wang, C., Wang, Y., Xu, B., & Zhang, W. (2018). LINC00312 represses proliferation and metastasis of colorectal cancer cells by regulation of miR-21. Journal of cellular and molecular medicine, 22(11), 5565–5572.
[26] Gu L, Lu L, Zhou D, Liu Z. (2018). Long Noncoding RNA BCYRN1 Promotes the Proliferation of Colorectal Cancer Cells via Up-Regulating NPR3 Expression. Cell Physiol Biochem. 2018;48(6):2337-2349. doi: 10.1159/000492649. Epub 2018 Aug 16. PMID: 30114690.
[27] Yerukala Sathipati, S., Ho, SY. (2018). Identifying a miRNA signature for predicting the stage of breast cancer. Sci Rep 8, 16138 (2018).
[28] Yerukala Sathipati, S., Ho, SY. (2020). Novel miRNA signature for predicting the stage of hepatocellular carcinoma. Sci Rep 10, 14452 (2020).
[29] Sathipati SY, Ho, SY. (2021). Identification of the miRNA signature associated with survival in patients with ovarian cancer. Aging (Albany NY). 2021 Apr 27;13(9):12660-12690. doi: 10.18632/aging.202940. Epub 2021 Apr 27. PMID: 33910165; PMCID: PMC8148489.

[30] Yerukala Sathipati, S., Tsai, MJ., Shukla, S.K. Ho, SY. (2022). MicroRNA signature for estimating the survival time in patients with bladder urothelial carcinoma. Sci Rep 12, 4141
[31] Zhang, H., Wang, Z., Wu, J., Ma, R., & Feng, J. (2019). Long noncoding RNAs predict the survival of patients with colorectal cancer as revealed by constructing an endogenous RNA network using bioinformation analysis. Cancer medicine, 8(3), 863–873.
[32] Lin, Y., Pan, X., Chen, Z., Lin, S., & Chen, S. (2020). Identification of an Immune-Related Nine-lncRNA Signature Predictive of Overall Survival in Colon Cancer. Frontiers in genetics, 11, 318.
[33] Li, S., Chen, S., Wang, B., Zhang, L., Su, Y., & Zhang, X. (2020). A Robust 6-lncRNA Prognostic Signature for Predicting the Prognosis of Patients with Colorectal Cancer Metastasis. Frontiers in medicine, 7, 56.
[34] Huang, X., Cai, W., Yuan, W., & Peng, S. (2020). Identification of key lncRNAs as prognostic prediction models for colorectal cancer based on LASSO. International journal of clinical and experimental pathology, 13(4), 675–684.
[35] Gao, M., Guo, Y., Xiao, Y. et al. (2021). Comprehensive analyses of correlation and survival reveal informative lncRNA prognostic signatures in colon cancer. World J Surg Onc 19, 104 (2021).
[36] Tang, X., Lin, Y., He, J. et al. (2022). Establishment and validation of a prognostic model based on HRR-related lncRNAs in colon adenocarcinoma. World J Surg Onc 20, 74 (2022).
[37] Chen, X., Zhu, Z., Li, X., Yao, X., & Luo, L. (2021). The Ferroptosis-Related Noncoding RNA Signature as a Novel Prognostic Biomarker in the Tumor Microenvironment, Immunotherapy, and Drug Screening of Gastric Adenocarcinoma. Frontiers in oncology, 11, 778557.
[38] Wei, T., Zhu, N., Jiang, W., & Xing, X. L. (2022). Development and Validation of Ferroptosis- and Immune-Related lncRNAs Signatures for Breast Infiltrating Duct and Lobular Carcinoma. Frontiers in oncology, 12, 844642.
[39] Shengxun Mao, Zhaohong Mo, Runxin Wu, Bin Lai, Zhiyong Zhou, Yi Song, Xi Ouyang & Xingen Zhu (2022) The double homeobox a pseudogene 8 accelerates cell proliferation, migration, and invasion in colon cancer, Bioengineered, 13:4, 8164-8173, DOI: 10.1080/21655979.2022.2053802
[40] Wu, J., Zheng, C., Wang, Y., Yang, Z., Li, C., Fang, W., Jin, Y., Hou, K., Cheng, Y., Qi, J., Qu, X., Liu, Y., Che, X., & Hu, X. (2021). LncRNA APCDD1L-AS1 induces icotinib resistance by inhibition of EGFR autophagic degradation via the miR-1322/miR-1972/miR-324-3p-SIRT5 axis in lung adenocarcinoma. Biomarker research, 9(1), 9.
[41] Huang, W., Su, G., Huang, X., Zou, A., Wu, J., Yang, Y., Zhu, Y., Liang, S., Li, D., Ma, F., & Guo, L. (2019). Long noncoding RNA PCAT6 inhibits colon cancer cell apoptosis by regulating anti-apoptotic protein ARC expression via EZH2. Cell cycle (Georgetown, Tex.), 18(1), 69–83.
[42] Yin K, Lee J, Liu Z, Kim H, Martin DR, Wu D, Liu M, Xue X. (2021). Mitophagy protein PINK1 suppresses colon tumor growth by metabolic reprogramming via p53 activation and reducing Acetyl-CoA production. Cell Death Differ. 2021 Aug; 28(8):2421-2435. doi: 10.1038/s41418-021-00760-9. Epub 2021 Mar 15. PMID: 33723373; PMCID: PMC8329176.
[43] D'Onofrio, N., Martino, E., Mele, L., Colloca, A., Maione, M., Cautela, D., Castaldo, D., & Balestrieri, M. L. (2021). Colorectal Cancer Apoptosis Induced by Dietary δ-Valerobetaine Involves PINK1/Parkin Dependent-Mitophagy and SIRT3. International journal of molecular sciences, 22(15), 8117.

[44] Yang, Y., Ge, H., Li, D. Q., & Xu, A. X. (2021). E2F1-Induced lncRNA BAIAP2-AS1 Overexpression Contributes to the Malignant Progression of Hepatocellular Carcinoma via miR-361-3p/SOX4 Axis. Disease markers, 2021, 6256369.
[45] Du, T., Gao, Q., Zhao, Y., Gao, J., Li, J., Wang, L., Li, P., Wang, Y., Du, L., & Wang, C. (2021). Long Non-coding RNA LINC02474 Affects Metastasis and Apoptosis of Colorectal Cancer by Inhibiting the Expression of GZMB. Frontiers in oncology, 11, 651796.
[46] Guo, R., & Qin, Y. (2020). LEMD1-AS1 Suppresses Ovarian Cancer Progression Through Regulating miR-183-5p/TP53 Axis. OncoTargets and therapy, 13, 7387–7398.
[47] Chen, S. W., Zhu, J., Ma, J., Zhang, J. L., Zuo, S., Chen, G. W., Wang, X., Pan, Y. S., Liu, Y. C., & Wang, P. Y. (2017). Overexpression of long non-coding RNA H19 is associated with unfavorable prognosis in patients with colorectal cancer and increased proliferation and migration in colon cancer cells. Oncology letters, 14(2), 2446–2452.
[48] Liu, S., Mitra, R., Zhao, M. M., Fan, W., Eischen, C. M., Yin, F., & Zhao, Z. (2016). The Potential Roles of Long Noncoding RNAs (lncRNA) in Glioblastoma Development. Molecular cancer therapeutics, 15(12), 2977–2986.
[49] Chodary Khameneh, S., Razi, S., Shamdani, S. et al. Weighted correlation network analysis revealed novel long non-coding RNAs for colorectal cancer. Sci Rep 12, 2990 (2022).
[50] Wang F, Zu Y, Huang W, Chen H, Xie H, Yang Y. LncRNA CALML3-AS1 promotes tumorigenesis of bladder cancer via regulating ZBTB2 by suppression of microRNA-4316. Biochem Biophys Res Commun. 2018 Sep 26;504(1):171-176. doi: 10.1016/j.bbrc.2018.08.150. Epub 2018 Sep 1. PMID: 30177388.
[51] Chen, F. B., Wu, P., Zhou, R., Yang, Q. X., Zhang, X., Wang, R. R., Qi, S. C., & Yang, X. (2020). LINC01315 Impairs microRNA-211-Dependent DLG3 Downregulation to Inhibit the Development of Oral Squamous Cell Carcinoma. Frontiers in oncology, 10, 556084.
[52] Safran M, Rosen N, Twik M, BarShir R, Iny Stein T, Dahary D, Fishilevich S, and Lancet D. The GeneCards Suite Chapter, Practical Guide to Life Science Databases (2022) pp 27-56
[53] Li JH, et al. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data , Nucleic Acids Res. 2014 Jan;42:D92-7.
[54] Pingping Wang, Hongxin Liu, Liang Cheng, Wenyang Zhou, Xiyun Jin, Zhaochun Xu, Meng Luo, Liran Juan, and Qinghua Jiang, 'Lncrna2target V3.0: A Comprehensive Database for Target Genes of lncRNAs in Human and Mouse'. http://bio-annotation.cn/lncrna2target/.
[55] Zhou et al., Metascape, Nature Communication (2019), 10(1):1523
[56] Szklarczyk D*, Gable AL*, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021 Jan 8;49(D1):D605-12
[57] M.-G. Lin, Y.-K. Hong, Y. Zhang, B.-B. Lin, X.-J. He, Mechanism of lncRNA DUXAP8 in promoting proliferation of bladder cancer cells by regulating PTEN
[58] Li, Jia-Rong MD, MM; Liu, Ling MD, PhD; Luo, Hui MD; Chen, Ze-Guo MD; Wang, Jian-Hua MD; Li, Nian-Feng MD, PhD. Long Noncoding RNA DUXAP8 Promotes Pancreatic Carcinoma Cell Migration and Invasion Via Pathway by miR-448/WTAP/Fak Signaling Axis. Pancreas: March 2021 - Volume 50 - Issue 3 - p 317-326 doi: 10.1097/MPA.0000000000001751
[59] Lian, Y., Yang, J., Lian, Y. et al. DUXAP8, a pseudogene derived lncRNA, promotes growth of pancreatic carcinoma cells by epigenetically silencing CDKN1A and KLF2. Cancer Commun 38,

64 (2018). https://doi.org/10.1186/s40880-018-0333-9

[60] Yang W, Zhou J, Zhang Z, Zhang K, Xu Y, Li L, Cai L, Gong Y, Gong K. Downregulation of lncRNA APCDD1L-AS1 due to DNA hypermethylation and loss of VHL protein expression promotes the progression of clear cell renal cell carcinoma. Int J Biol Sci 2022; 18(6):2583-2596. doi:10.7150/ijbs.71519.

[61] Wang S, Chen Z, Gu J, Chen X, Wang Z. The Role of lncRNA PCAT6 in Cancers. Front Oncol. 2021 Jul 13;11:701495. doi: 10.3389/fonc.2021.701495. PMID: 34327141; PMCID: PMC8315724.

[62] Patoughi M, Ghafouri-Fard S, Arsang-Jang S, Taheri M. Expression analysis of PINK1 and PINK1-AS in multiple sclerosis patients versus healthy subjects. Nucleosides Nucleotides Nucleic Acids. 2021;40(2):157-165. doi: 10.1080/15257770.2020.1844229. Epub 2020 Nov 9. PMID: 33161812.

[63] Wang B, Suen CW, Ma H, Wang Y, Kong L, Qin D, Lee YWW and Li G (2020) The Roles of H19 in Regulating Inflammation and Aging. Front. Immunol. 11:579687. doi: 10.3389/fimmu.2020.579687.

[64] Auslander, N., Wolf, Y.I. & Koonin, E. Interplay between DNA damage repair and apoptosis shapes cancer evolution through aneuploidy and microsatellite instability. Nat Commun 11, 1234 (2020). https://doi.org/10.1038/s41467-020-15094-2.

[65] Lin A, Zhang J and Luo P (2020) Crosstalk Between the MSI Status and Tumor Microenvironment in Colorectal Cancer. Front. Immunol. 11:2039. doi: 10.3389/fimmu.2020.02039

[66] Johansen, A.F.B., Kassentoft, C.G., Knudsen, M. et al. Validation of computational determination of microsatellite status using whole exome sequencing data from colorectal cancer patients. BMC Cancer 19, 971 (2019). https://doi.org/10.1186/s12885-019-6227-7

[67] Chang, C-C and Lin, C-J, LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at

[68] Ho, SY, H., Li-Sun, S. & Jian-Hung, C. (2004) Intelligent evolutionary algorithms for large parameter optimization problems. IEEE Transactions on Evolutionary Computation 8, 522-541.

[69] Bender, R., T. Augustin, and M. Blettner, Generating survival times to simulate Cox proportional hazards models. Statistics in medicine, 2005. 24(11): p. 1713-1723.

[70] Ho SY, Chen JH, Huang MH. (2004) Inheritable genetic algorithm for biobjective 0/1 combinatorial optimization problems and its applications. IEEE Trans Syst Man Cybern B Cybern. 2004;34:609–620.

[71] Smith TC, Frank E. (2016) Introducing Machine Learning Concepts with WEKA. Methods Mol Biol. 2016;1418:353-78.

Tables

Table 1. Existing lncRNA signature identification studies of predicting patients with colon, CRC and COAD.

	Cancer	Feature selection	Prediction model	Prediction	database	Reference
1	CRC	Statistic	Cox proportional hazard regression model	Survival	TCGA+ GEO	[31] (2019)
2	colon	LASSO	Univariate, lasso and multivariate Cox regression analyses	Survival	TCGA	[32] (2020)
3	CRC	Statistic	univariate Cox regression analysis, followed by a stepwise multivariate Cox regression model	Survival	TCGA	[33] (2020)
4	CRC	LASSO	LASSO regression	Survival	TCGA	[34] (2020)
5	colon	LASSO	Cox analysis, random survival forest analysis, and lasso regression analysis	Survival	TCGA	[35] (2021)
6	COAD	Statistic	LASSO Cox regression	Survival	TCGA	[36] (2022)
7	COAD	Genetic algorithm	Support vector machine	Stage	TCGA+ GEO	This study

Table 2. Ranking of lncRNAs using the main effect difference (MED) score and their relationship with cancers published.

Rank	Ensemble Gene ID	lncRNA	MED	Colon/ CRC	Cancer	Reference
1	ENSG00000185332	TMEM105	0.465		Gastric, Breast	[37][38]
2	ENSG00000206195	DUXAP8	0.435	Colon		[39]
3	ENSG00000231290	APCDD1L-DT	0.375		Lung	[40]
4	ENSG00000228288	PCAT6	0.315	Colon		[41]
5	ENSG00000226308	Uncharacterized	0.285			
6	ENSG00000117242	PINK1-AS	0.225	Colon		[42][43]
7	ENSG00000226137	BAIAP2-DT	0.165		HCC	[44]
8	ENSG00000228437	LINC02474	0.165	CRC		[45]
9	ENSG00000226235	LEMD1-AS1	0.135		Ovarian	[46]
10	ENSG00000130600	H19	0.105	CRC		[47]
11	ENSG00000197291	RAMP2-AS1	0.105		Glioblastoma	[48]

12	ENSG00000204387	SNHG32	0.105	CRC		[49]
13	ENSG00000205488	CALML3-AS1	0.045		Bladder	[50]
14	ENSG00000231651	DLG3-AS1	0.045		Oral	[51]
15	ENSG00000206417	H1-10-AS1	0.015			

Table 3. Comparison of the EL-COAD signature and the set of 15 p-value lncRNAs using 14 models of Weka and the TCGA-COAD dataset.

	Prediction model	5-CV (%)	Test (%)	5-CV (%)	Test (%)
		15-lncRNA signature		15 p-value lncRNAs	
1	SVM	79.66	62.50	70.62	59.87
2	Naïve Bayes	68.36	64.47	64.41	60.53
3	Logistic	72.60	65.13	67.51	60.53
4	Multilayer Perceptron	67.80	59.21	65.54	57.24
5	Random Forest	70.34	61.04	70.34	60.53
6	REP tree	61.30	59.87	64.12	50.55
7	SMO	72.03	64.47	67.23	61.84
8	J48	66.95	68.42	65.82	61.18
9	SGD	72.03	64.47	66.67	60.53
10	LMT	71.75	65.79	64.97	59.06
11	IBK	50.47	59.21	60.17	55.26
12	LWL	63.84	62.50	66.10	62.50
13	Decision Table	65.82	64.47	64.12	59.06
14	JRip	65.25	61.10	65.02	59.06
	mean	67.73	63.05	65.90	59.12
	standard deviation	6.74	2.73	2.63	3.08

Table 4. The top-ranked KEGG pathways by using the Metascape database.

Rank	KEGG pathway (ID)	Log10 (p-value)	Target genes
1	MicroRNAs in cancer (hsa05206)	-32.827	CDKN1A,DNMT1,DNMT3B,E2F1,EGFR,EZH2,MTOR,MMP9,ABCC1,MYC,NOTCH1,NOTCH2,NOTCH3,ABCB1,PIK3CD,MAPK1,MAPK3,TP53,VEGFA,VIM,HMGA2,DICER1,SIRT1,MIRLET7A1,MIRLET7B,MIR107,MIR141,MIR152,MIR18A,MIR1941,MIR19A,MIR19B1,MIR200A,MIR200B,MIR200C,MIR29A,MIR29B1,MIR326,MIR342,MIR615
2	Endocrine resistance (hsa01522)	-12.368	BAX,CDKN1A,E2F1,EGFR,MTOR,IGF1R,MMP9,NOTCH1,NOTCH2,NOTCH3,PIK3CD,MAPK1,MAPK3,TP53

3	Human papillomavirus infection (hsa05165)	-12.084	BAX,CDKN1A,COL2A1,CTNNB1,E2F1,EGFR,FOXO1,MTOR,HLA-E,ITGA1,LAMA2,LAMA3,NOTCH1,NOTCH2,NOTCH3,PIK3CD,MAPK1,MAPK3,SPP1,TP53,TUBG1,VEGFA
4	Pathways in cancer (has05200)	-8.88514	BAX,CAMK2D,CDH1,CDKN1A,CTNNB1,E2F1,EGFR,FOXO1,MTOR,IGF1R,LAMA2,LAMA3,MMP9,MYC,NOTCH1,NOTCH2,NOTCH3,PIK3CD,MAPK1,MAPK3,TP53,VEGFA,TRAF4
5	Colorectal cancer (has05210)	-8.10336	BAX,CDKN1A,CTNNB1,EGFR,MTOR,MYC,PIK3CD,MAPK1,MAPK3,TP53
6	Transcriptional misregulation in cancer (hsa05202)	-7.436	BAX,RUNX2,CDKN1A,FOXO1,FUS,GZMB,TLX1,IGF1R,MMP9,MYC,TP53,HMGA2,PROM1

Table 5. The top-10 of 43 diseases using Disease Ontology (DO) analysis according to the strength.

Rank	disease	description	count in network	strength	false discovery rate
1	DOID:3308	Embryonal carcinoma	3 of 5	1.72	0.0105
2	DOID:3307	Teratoma	4 of 7	1.70	0.0013
3	DOID:219	Colon cancer	5 of 14	1.50	0.00066
4	DOID:686	Liver carcinoma	5 of 15	1.47	0.00072
5	DOID:4007	Bladder carcinoma	3 of 9	1.47	0.0276
6	DOID:0060108	Brain glioma	3 of 9	1.47	0.0276
7	DOID:5520	Head and neck squamous cell carcinoma	3 of 10	1.42	0.032
8	DOID:684	Hepatocellular carcinoma	4 of 14	1.40	0.0065
9	DOID:5603	T-cell acute lymphoblastic leukemia	3 of 11	1.38	0.0374
10	DOID:9256	Colorectal cancer	6 of 25	1.33	0.00066

Table 6. The 20 lncRNAs in the identified survival signature.

Rank	Ensemble Gene ID	lncRNA	MED
1	ENSG00000232386	Uncharacterized	0.026
2	ENSG00000250509	Uncharacterized	0.022
3	ENSG00000259840	Uncharacterized	0.018
4	ENSG00000254605	Uncharacterized	0.014
5	ENSG00000272512	Uncharacterized	0.012
6	ENSG00000237945	LINC00649	0.012
7	ENSG00000267317	Uncharacterized	0.010
8	ENSG00000281120	Uncharacterized	0.010
9	ENSG00000203993	ARRDC1-AS1	0.010
10	ENSG00000266903	CEACAM16-AS1	0.008
11	ENSG00000273199	Uncharacterized	0.008
12	ENSG00000124835	LOC93463	0.008
13	ENSG00000265478	Uncharacterized	0.006
14	ENSG00000260597	Uncharacterized	0.005
15	ENSG00000197251	LINC00336	0.004
16	ENSG00000273230	Uncharacterized	0.004
17	ENSG00000273314	Uncharacterized	-0.001
18	ENSG00000230975	Uncharacterized	-0.002
19	ENSG00000271871	Uncharacterized	-0.004
20	ENSG00000223768	LINC00205	-0.006

Figures

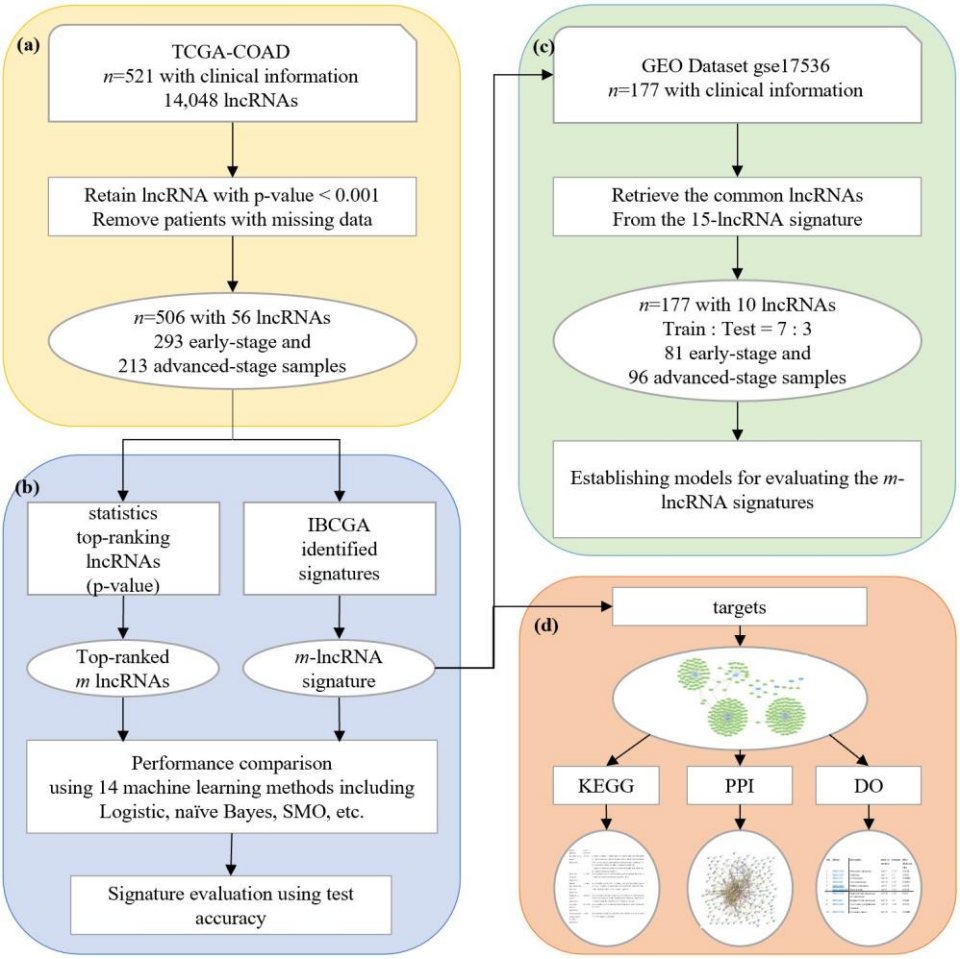


Figure 1. The flowchart of the proposed method EL-COAD and the signature analysis. (a) preprocessing of training dataset, (b) signature identification and comparison, (c) signature confirmation using a GEO dataset, (d) KEGG and DO analysis.

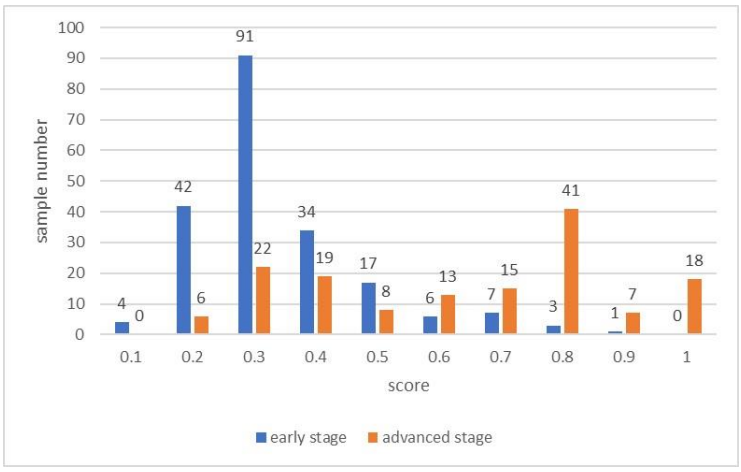
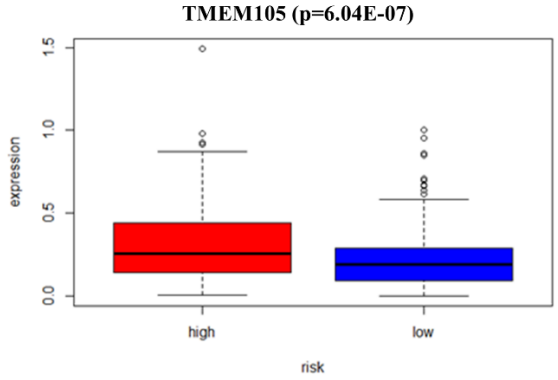
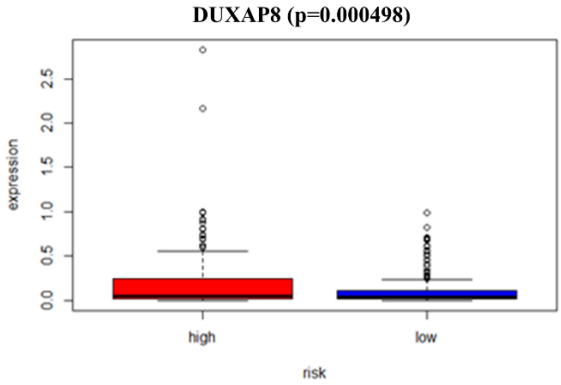
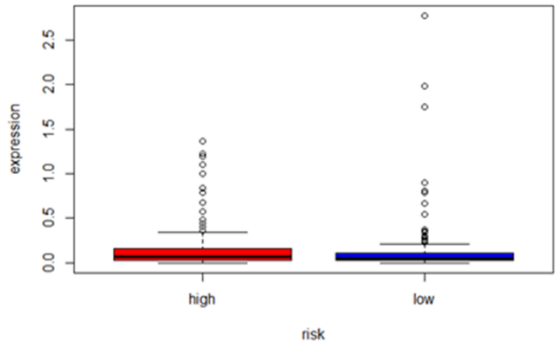
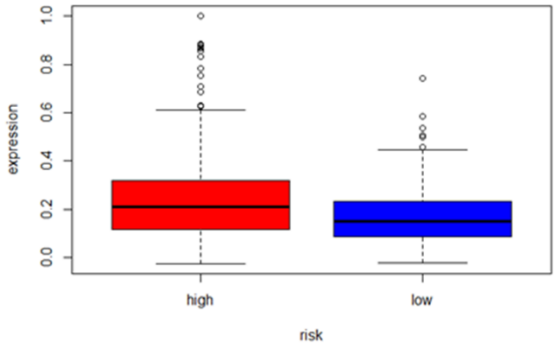


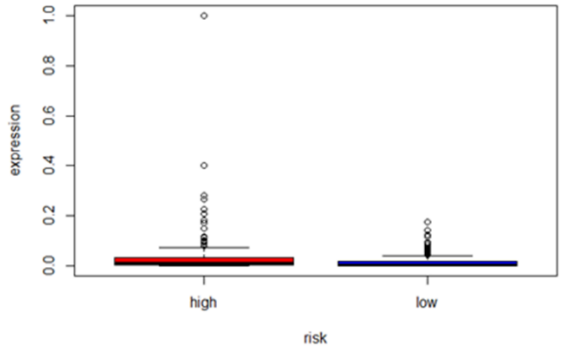
Figure 2. The distribution of sample scores in the training set of TCGA-COAD using the SVM model.

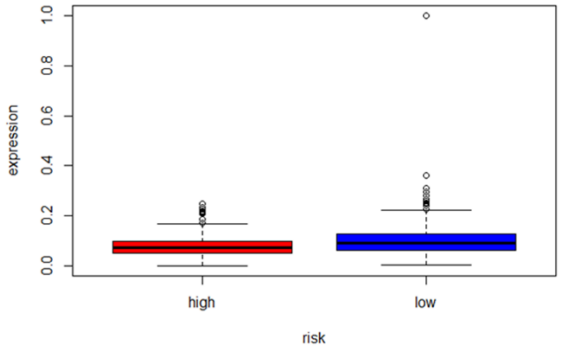

TMEM105 (p=6.04E-07)
expression
0.0
0.5
1.0
1.5
high
low
risk

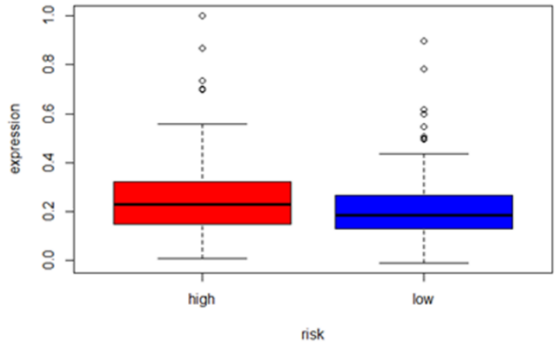

DUXAP8 (p=0.000498)
expression
0.0
0.5
1.0
1.5
2.0
2.5
high
low
risk

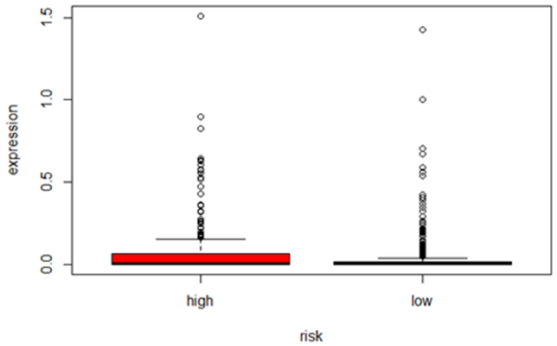

APCDD1L-DT (p=0.000938)
expression
0.0
0.5
1.0
1.5
2.0
2.5
high
low
risk


PCAT6 (p=2.12E-06)
expression
0.0
0.2
0.4
0.6
0.8
1.0
high
low
risk


ENSG00000226308 (p=2.19E-06)
expression
0.0
0.2
0.4
0.6
0.8
1.0
high
low
risk


PINK1-AS (p=5.63E-05)
expression
0.0
0.2
0.4
0.6
0.8
1.0
high
low
risk


BAIAP2-DT (p=0.000179)
expression
0.0
0.2
0.4
0.6
0.8
1.0
high
low
risk


LINC02474 (p=0.000296)
expression
0.0
0.5
1.0
1.5
high
low
risk

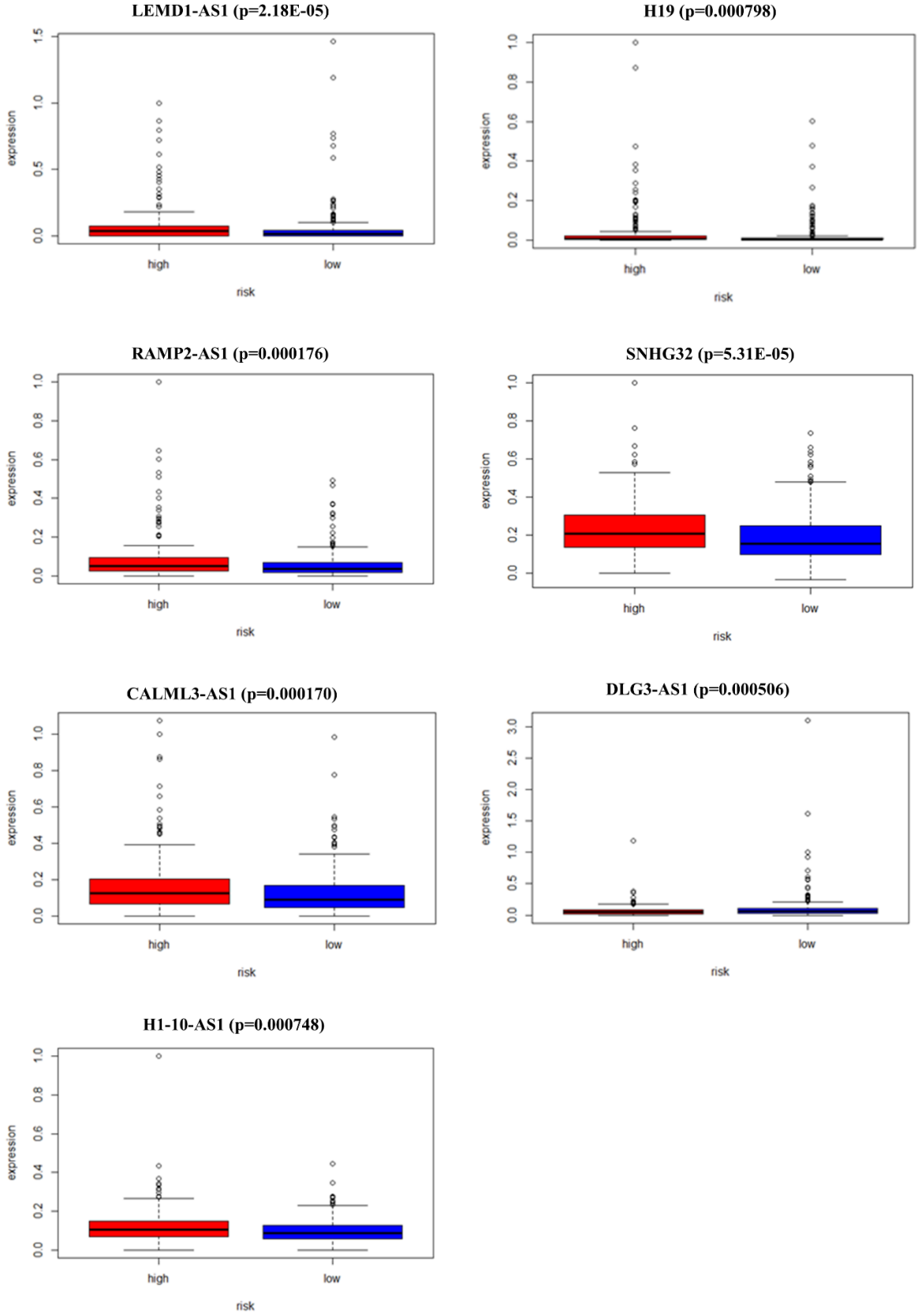


Figure 3. The box plots and p-value of the 15 lncRNAs in the identified signature.

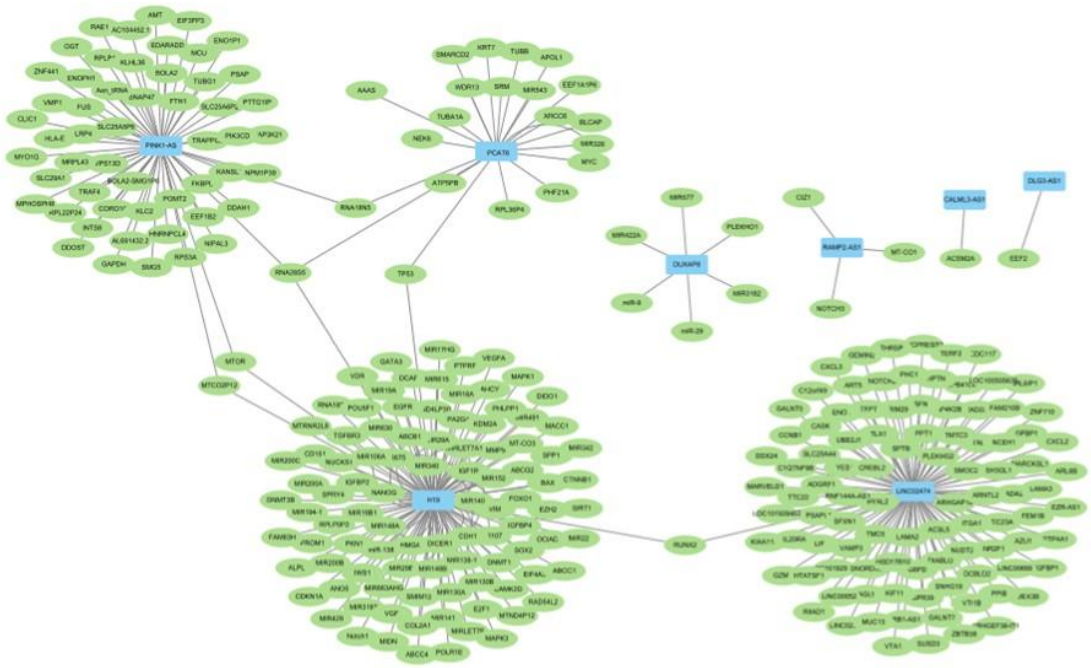


Figure 4. The lncRNA–RNA interaction network consisting of 8 lncRNAs and 330 RNAs.

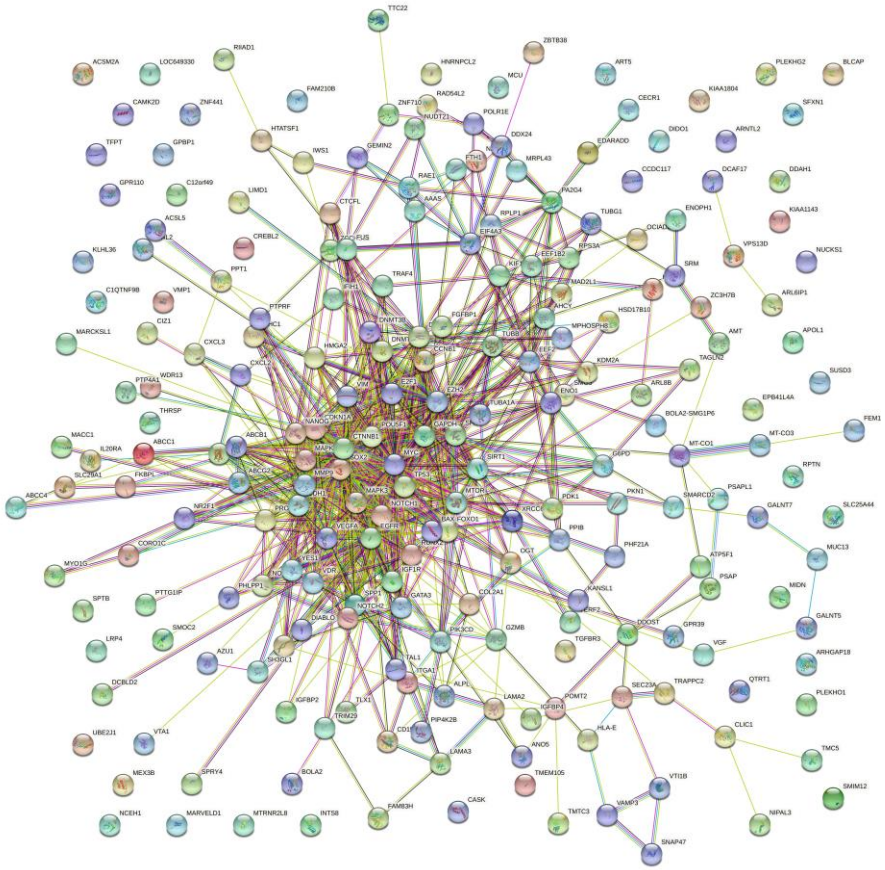


Figure 5. The protein-protein interactions of target genes obtained using the String database.

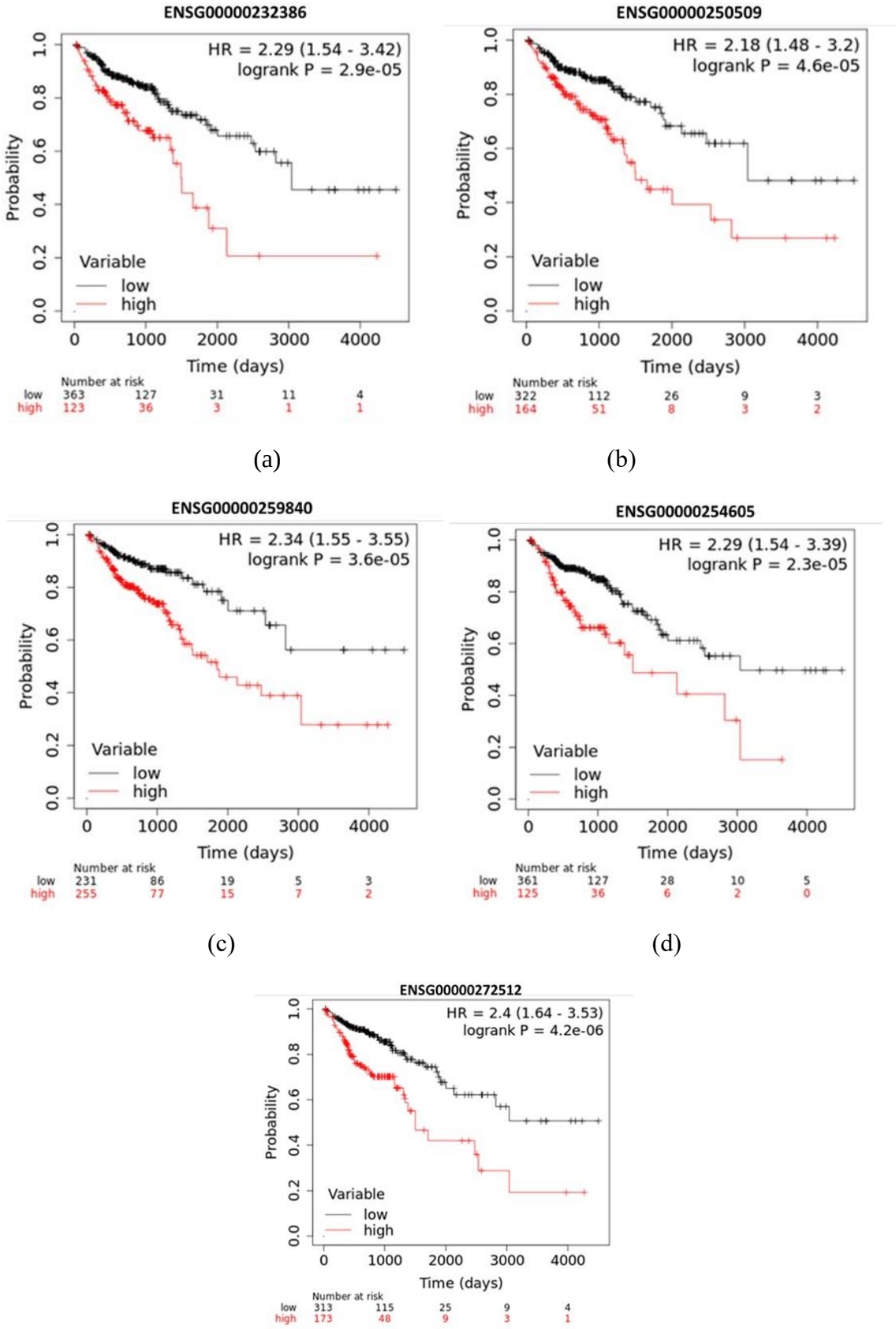


Figure 6. The Kaplan–Meier survival curves of the top-five lncRNAs. The low expression of these lncRNAs has a good survival of patients with COAD.