Iran Red Crescent Med J. 2019 May; 21(5):e87526. doi: 10.5812/ircmj.87526.

Published online 2019 June 11. Research Article

Identification of a Novel Genetic Signature in Staging of Colorectal Cancer: A Bayesian Approach

Fatemeh Mohammadzadeh[1], Ebrahim Hajizadeh[1, *], Aliakbar Rasekhi[1] and Sadegh Azimzadeh Jamalkandi[2]

[1]Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
[2]Chemical Injuries Research Center, Systems Biology and Poisonings Institute, Baqiyatallah University of Medical Sciences, Tehran, Iran

[*]*Corresponding author*: Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran. Email: hajizadeh@modares.ac.ir

Received 2018 December 14; **Revised** 2019 March 26; **Accepted** 2019 April 03.

Abstract

Background: Tumor stage is one of the most reliable prognostic factors in the clinical characterization of colorectal cancer. The identification of genes associated with tumor staging may facilitate the personalized molecular diagnosis and treatment along with better risk stratification in colorectal cancer.

Objectives: The study aimed to identify genetic signatures associated with tumor staging and patients' survival in colorectal cancer and recognize the patients' risk category for clinical outcomes based on transcriptomic data.

Methods: In this retrospective cohort study, two available transcriptomic datasets, including 232 patients with colorectal cancer under accession number GSE17537 and GSE17536 were used as discovery and validation sets, respectively. A Bayesian sparse group selection method in the discovery set was applied to identify the associated genes with the tumor staging. Then further screening was performed using survival analysis, and significant genes were used to develop a gene signature model. Finally, the robust performance of the signature model was assessed in the validation set.

Results: A total of 56 genes were significantly associated with the tumor staging in colorectal cancer. Survival analysis resulted in a shortlist of 19 genes, including ADH1B (P = 0.012), AHI (P = 0.006), AKAP12 (P = 0.018), BNIP3 (P = 0.015), CLDN11 (P = 0.015), CST9L (P = 0.028), DPP10 (P = 0.029), FBXO33 (P = 0.036), HEBP (P = 0.025), INTS4 (P = 0.003), LIPJ (P = 0.001), MMP21 (P = 0.006), NGRN (P = 0.014), PAFAH1B2 (P = 0.035), PCOLCE2 (P = 0.009), PIM1 (P = 0.007), TBKBP1 (P = 0.003), TCEB3B (P = 0.001), and TIPARP (P = 0.018), developing the signature model and validation. In both discovery and validation sets, the discrimination ability of the signature model to categorize patients with colorectal cancer into low- and high-risk subgroups for mortality and recurrence at 3- and 5-years showed good discrimination performances, with the area under the receiver operating characteristic curve (ROC) ranging from 0.64 to 0.88. It also had good sensitivity (discovery set 63.1%, validation set 61.7%) and specificity (discovery set 75.0%, validation set 59.3%) to discriminate between early- and late-stage groups.

Conclusions: We identified a 19-gene signature associated with tumor staging and survival of colorectal cancer, which may represent potential diagnosis and prognosis markers, and help to classify patients with colorectal cancer into low- or high-risk subgroups.

Keywords: Bayesian Approach, Colorectal Cancer, Gene Expression Signatures, Microarray Analysis, Prognosis, Recurrence, Overall Survival, Tumor Staging, Classification, Gene Ontology, Risk, Transcriptome

1. Background

Colorectal cancer (CRC) is the third most frequent cancer and the fourth leading cause of cancer-related death worldwide (1). As a complex disease, CRC is affected by several genetic and environmental factors (2). In recent years, intensive studies have been conducted to provide more insight into molecular alterations in the CRC. However, the molecular mechanisms underlying the CRC progression and tumor metastasis are still unclear.

Identification of the tumor stage is currently the most reliable prognostic factor in the CRC (3). It is strongly associated with the survival of patients with CRC and influences the decision-making about the treatment plans (4). The tumor node metastasis (TNM) staging system is a traditional approach to divide the cancer into four categories, namely I, II, III, and IV, based on the clinicopathologic features, including tumor size, nodal spread, and metastasis. A higher stage is corresponding to a further progression of cancer and poorer clinical outcomes. Therefore, identification of genes associated with the TNM staging of the CRC can help to discover novel diagnostic and prognostic biomarkers, and develop an improved prognostic tool to optimize the personalized treatment and risk stratification of patients with CRC.

Currently, the systems biology approach in solving biological problems is a new approach that simultaneously considers all changes at different levels of the gene expression. Meanwhile, there is a huge amount of transcriptomics data that can be explored using a variety of statistical analysis methods or data mining techniques.

2. Objectives

In this study, we attempted to identify a robust gene signature based on genes associated with the tumor stage for risk stratification of patients with CRC by taking into consideration the high correlation between genes in microarray data through a Bayesian approach. Finally, the robust performance of the signature model was validated in an independent cohort dataset.

3. Methods

3.1. Microarray Data and Data Preprocessing

In this retrospective cohort study, two previously published transcriptomics datasets, including 232 patients with colorectal cancer (5) under accession number GSE17537 and GSE17536 were used as discovery and validation sets, respectively in the Gene Expression Omnibus (GEO) (www.ncbi.nlm.nih.gov/geo) for which met the following inclusion criteria: (a) being human gene expression data, (b) profiled on the Affymetrix HG-U133_Plus_2 platform, (c) information about age, sex, pathological tumor staging (stage I-IV), death status, and survival time. Details of protocols and procedures of data collection, instruments, and variables measurement are available (5). Raw data of both datasets were downloaded and normalized using the Robust Multi-array Average (RMA) algorithm (6). The Z-score transformed was used to standardize the expression value of each gene.

3.2. Statistical Analysis

All analyses were performed using R programming software version 3.5.0. Figure 1 shows the workflow of the study. At first, primary screening was performed using the significance analysis of microarrays (SAM) algorithm (7) available in "siggenes" package (8) to detect differentially expressed genes (DEG) between the different stages of CRC in the discovery set (GSE17537). A false discovery rate (FDR) $<$ 0.1 was set for identification DEGs.

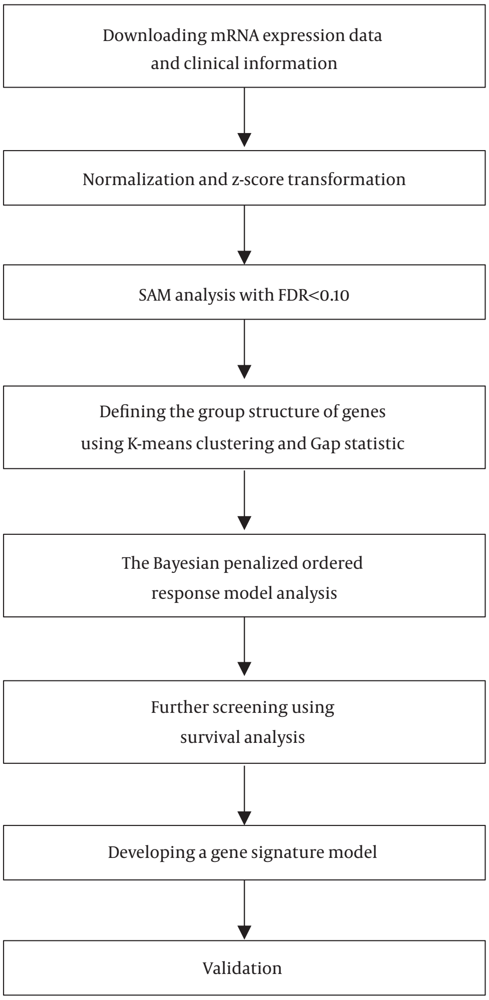
Downloading mRNA expression data
and clinical information
Normalization and z-score transformation
SAM analysis with FDR<0.10
Defining the group structure of genes
using K-means clustering and Gap statistic
The Bayesian penalized ordered
response model analysis
Further screening using
survival analysis
Developing a gene signature model
Validation

Figure 1. Flow chart of the analysis procedure, including data collection, preprocessing, analysis, and validation

3.3. Detection of Associated Genes with the TNM-Staging of CRC in the Discovery Set

After differential gene expression analysis, the parallel line assumption was tested using a score test in a univariate ordered probit model for each gene. Then a multiple penalized ordinal probit model was used to discover the associated genes with the TNM-staging of patients with CRC. In this model, we used a Bayesian sparse group selection (BSGS) method for variable selection, which took into account the group structures between genes in microarray data. The BSGS method was proposed by Xu and Ghosh (9) for a linear model. They showed the superior accuracy and excellent performance of this method for variable selection and prediction by simulation. In the BSGS method, it is assumed that predictors to be partitioned into G groups and spike and slab priors are used to select variables at both group and within-group levels. The formulation of the prior for the coefficient vector of the gth group can be written as (9):

$$\beta_g = V_g^{\frac{1}{2}} b_g,$$

$$where\ V_g^{\frac{1}{2}} = diag\left\{\tau_{g1}, \ldots, \tau_{gm_g}\right\}$$

$\tau_{gj} \geq 0$,
g = 1, …, G,
j = 1, …, m_g

$$b_g \overset{ind}{\sim} (1-\pi_0)\, N_{m_g}\left(0, I_{m_g}\right) + \pi_0 \delta_0\left(b_g\right),$$

$$\tau_{gj} \overset{ind}{\sim} (1-\pi_1)\, N^+\left(0, s^2\right) + \pi_1 \delta_0\left(\tau_{gj}\right),$$

$$\pi_0 \sim Beta\left(1,1\right),\ \pi_1 \sim Beta\left(1,1\right),\ s^2 \sim IG\left(1,t\right).$$

where β_g represents the group level coefficients of length m_g, b_g is the regression coefficients within groups, $N_{m_g}\left(0, I_{m_g}\right)$ denotes a multivariate standard normal distribution, $N^+\left(0, s^2\right)$ denotes a normal distribution truncated below at 0, δ_0 (.) is a point mass density function at zero, and t is the scale parameter of inverse gamma distribution for s^2 and is estimated with the Monte Carlo EM algorithm. In this study, we defined the group structure of genes using the K-means approach owing to its efficiency in clustering large datasets, low computational cost, and relatively robustness (10). We determined the optimal number of groups by using the Gap statistic (11). Then we considered the stage of CRC as an ordinal response variable and used an extension of BSGS method for an ordinal probit model using the latent variable approach, as described by Albert and Chip (12).

3.4. Further Screening and Validation of Candidate Genes

The prognostic value of candidate genes was investigated by establishing a gene signature (risk score) model. The risk score model developed based on the expression level of survival-associated genes with a P value < 0.05 as the significant difference, weighted by the corresponding regression coefficients in the univariate Cox models (13, 14). Then the patients were divided into low- or high-risk groups by the peak value in the frequency distribution histogram of the risk scores as the cut-off point (15, 16) and the following statistical analyses were used to assess the prognostic properties of the gene signature model in the discovery set. The Kaplan-Meier curve and the log-rank test were used to assess overall survival (OS) and recurrence-free survival (RFS) differences between the two groups, respectively. The time-dependent receiver operating characteristic (ROC) curve analysis was used to evaluate the accuracy of the gene signature for predicting OS and RFS. We also investigated the association between the gene signature model and tumor stages of patients with CRC. The ROC curve analysis was applied to quantify how accurately the gene signature model can discriminate between two early- and late-stage groups (I and II stages vs. III and IV stages). The age- and gender-independent prognostic value of the signature were also evaluated using multivariate Cox regression models. We validated the gene signature model using the independent GSE17536 dataset. To this end, the same gene signature model and the cut-off value were used to categorize the patients in the validation set. Next, the mentioned statistical analysis was applied to assess the signature model performance in the validation set.

3.5. Gene Set Enrichment

All 19 identified genes were used for enrichment analysis in Enrichr database (http://amp.pharm.mssm.edu/Enrichr). The Pathway and Gene Ontology (GO) results were used for annotation of the results as the biological process and involved pathways.

4. Results

4.1. Demographic and Clinical Characteristics

In this study, we analyzed the data on a total of 232 CRC samples, including 55 samples in the GSE17537 discovery set and 177 samples in the GSE17536 validation set. The median OS was more than 111 months for the discovery set and 135 months for the validation set. The median RFS was more than 76 months for the discovery set and more than 142 months for the validation set. Other demographic and clinical characteristics of the patients are shown in Table 1.

Table 1. Demographic and Clinical Characteristics of the Patients in Discovery Set (GSE17537), and Validation Set (GSE17536)[a]

Variable	GSE17537	GSE17536
Age, y	65.4 ± 13.1	62.3 ± 14.4
Gender		
Male	29 (52.7)	85 (48.0)
Female	26 (47.3)	92 (52.0)
Stage		
I	4 (7.3)	24 (13.6)
II	15 (27.3)	57 (32.2)
III	19 (34.5)	57 (32.2)
IV	17 (30.9)	39 (22.0)
No. of death	20 (36.3)	73 (41.2)
No. of recurrence	6 (10.9)	36 (20.3)

[a] Values are expressed as mean ± SD or No. (%).

4.2. The Discovery of Genes Associated with the TNM-Stage of CRC

A total of 23,520 genes were obtained from each dataset after preprocessing. The SAM analysis resulted in a list of 1,850 genes, which were clustered in 9 groups using the K-means approach and the Gap statistic. The Bayesian penalized ordinal probit model based on the predefined clusters resulted in the identification of 56 genes as potential biomarkers associated with the tumor stages of CRC. The prediction accuracy of the Bayesian model was 83.6% based on leave-one-out cross-validation method.

4.3. Further Screening and Validation of Candidate Genes

The results of the univariate Cox model showed that 19 genes of 56 candidate genes were associated with the OS of patients with CRC. We summarized the overall information of these genes in Table 2.

Table 2. Overall Information of the 19 Genes Associated with TNM Staging and OS of Patients with CRC in the Discovery Set

Gene Symbol	Bayesian Penalized Ordinal Response Model	Univariate Cox model	
	β (SD)	HR (95% CI)	P Value
ADH1B	0.11 (0.05)	1.55 (1.10, 2.19)	0.012
AHI1	-0.08 (0.04)	0.45 (0.25, 0.80)	0.006
AKAP12	0.13 (0.05)	1.48 (1.07, 2.05)	0.018
BNIP3	0.13 (0.05)	1.65 (1.10, 2.47)	0.015
CLDN11	0.10 (0.06)	1.59 (1.09, 2.30)	0.015
CST9L	-0.10 (0.05)	0.58 (0.36, 0.94)	0.028
DPP10	0.31 (0.08)	1.42 (1.04, 1.95)	0.029
FBXO33	0.14 (0.06)	1.72 (1.04, 2.84)	0.036
HEBP2	0.10 (0.05)	1.73 (1.07, 2.79)	0.025
INTS4	-0.09 (0.05)	0.50 (0.34, 0.73)	0.003
LIPJ	-0.10 (0.05)	0.53 (0.35, 0.78)	0.001
MMP21	-0.12 (0.05)	0.55 (0.35, 0.84)	0.006
NGRN	0.10 (0.05)	1.72 (1.12, 2.64)	0.014
PAFAH1B2	-0.19 (0.05)	0.54 (0.31, 0.96)	0.035
PCOLCE2	0.13 (0.06)	1.60 (1.12, 2.27)	0.009
PIM1	0.15 (0.05)	1.94 (1.20, 3.12)	0.007
TBKBP1	-0.16 (0.06)	0.54 (0.36, 0.81)	0.003
TCEB3B	-0.11 (0.05)	0.41 (0.25, 0.68)	0.001
TIPARP	0.14 (0.05)	1.62 (1.09, 2.41)	0.018

Abbreviations: SD, standard deviation; HR, hazard ratio; CI, confidence interval.

The coefficients of these genes were used to develop the gene signature model. The risk score for each patient was calculated based on the gene signature model and divided into low- or high-risk groups based on the cut-off value of -1.21 (the peak value of the frequency distribution histogram of the risk scores). The expression heat map of the 19 candidate genes of all patients is shown in Figure 2.

The log-rank test showed a significant difference between OS and also RFS curves for the two groups in both discovery set (OS: P = 0.0002, RFS: P = 0.031, Figures 3A and 2B), and validation set (OS: P = 0.009, RFS: P = 0.026, Figures 3C and 2D).

Time-dependent ROC curve analysis showed that in the discovery set, the signature model could predict the 3-, and 5-year OS of patients with CRC, as the AUC values were 0.88, and 0.81, respectively (Figure 4A). While in the validation set, the AUC values at 3- and 5-year OS of patients with CRC were 0.65, and 0.64, respectively (Figure 4B). In the discovery set, the AUC values at 3- and 5-year RFS of patients with CRC were 0.83, and 0.80, respectively (Figure 4C). Whereas in the validation set, the AUC values at 3- and 5-year RFS of patients with CRC were 0.64, and 0.68, respectively (Figure 4D). All of the time-dependent AUC values exceed 0.6. These results suggested that our 19-mRNA signature model has good performance for the prediction of disease course in patients with CRC.

The patients with early-stage had a significantly lower risk score than patients with late-stage in both discovery set (t = -4.3503, df = 42.915, P = 0.008), and validation set (t = -2.681, df = 170.09, P < 0.001). The sensitivity, specificity, and AUC for discriminating between early- and late-stage patients in the discovery set by the 19-gene signature model were 63.1%, 75.0%, and 70.9%, respectively, as shown in Figure 5A. These values were 61.7%, 59.3%, and 60.0% respectively, for the validation set (Figure 5B). Therefore, the 19-mRNA signature model had relatively good sensitivity and specificity to discriminate between two early- and late-stage groups.

The age-independent and gender-independent prognostic values of the signature were further evaluated using multivariate Cox regression models. The results from multivariate Cox regression suggested that our signature model is independent of both age and gender in the prognosis of patients with CRC in the discovery set (OS: HR, 1.17; 95% CI, 1.09 - 1.26, and RFS: HR, 1.41; 95% CI, 1.08 - 1.85) and in the validation set (OS: HR, 1.92; 95% CI, 1.11 - 3.32, and RFS: HR, 1.13; 95% CI, 1.04 - 1.23).

4.4. Gene Set Enrichment

The results of gene set enrichment indicated that most of the identified genes are involved and formerly reported in the CRC. Meanwhile, most of the genes showed basal expression in the CRC. Also, most of the genes are mitochondrial genes and are located within the outer membrane of the mitochondria involved in energy metabolism in CRC cells. According to biological processes of the

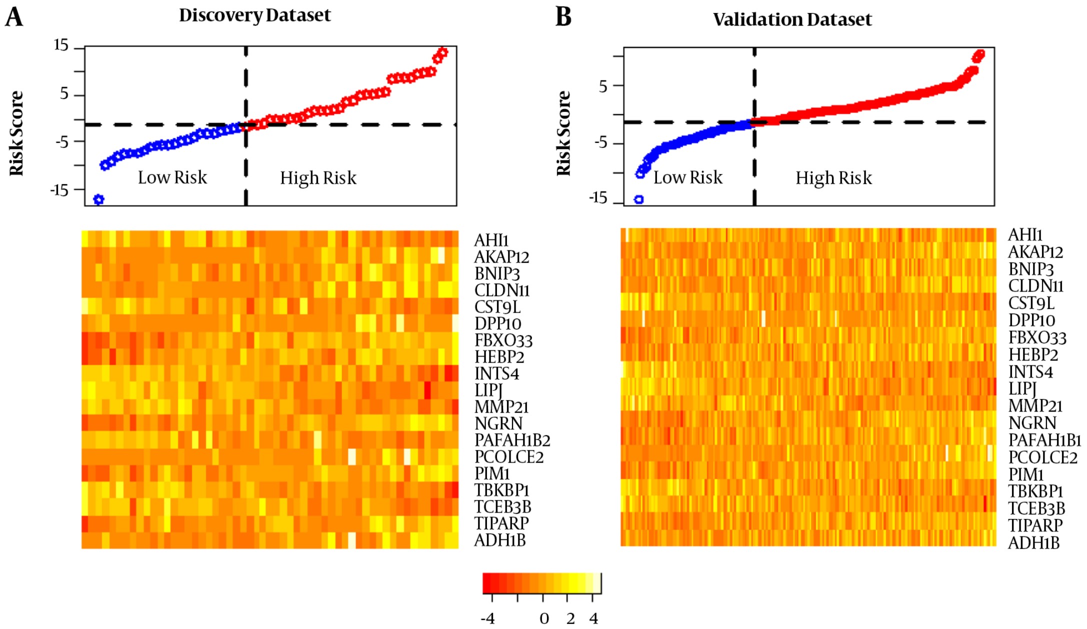

Figure 2. Expression heat map of 19-gene signature (sorted by risk score) (A) in the discovery dataset and (B) validation dataset

GO database, different types of autophagy, including mitophagy, xenophagy, aggrophagy, and lipophagy are the most important enriched signaling pathways. Perturbations in fatty acid and lipid metabolism are also frequent in the KEGG pathway database (Figure 6).

5. Discussion

In the present study, we identified a novel 19-mRNA gene signature model associated with the TNM staging and OS of patients with CRC using transcriptomic data analysis. Our signature model was able to classify the patients with CRC into low- and high-risk subgroups for mortality, and recurrence and also had rather good sensitivity and specificity to discriminate between two early- and late-stage groups. Therefore, the gene signature model may provide new promising biomarkers for predicting OS and RFS in patients with CRC and a novel strategy for the preoperative diagnosis and postoperative treatment in patients with CRC. Moreover, our signature model was independent of both age and gender in the prediction of the prognosis of patients with CRC. Thus it has the capability of practical usage in larger patient groups, regardless of gender or age.

The enrichment analysis, determining the biological functions of identified genes indicated that most of these genes are mainly involved in different types of autophagy through Gene Ontology (GO) terms. Also, we found that these genes are frequently observed in perturbations in fatty acid and lipid metabolism based on the KEGG pathway database. These results suggest there is a possible connection between the identified genes and progression or prognosis of CRC (17, 18).

A literature review showed the possible role of 15 of the 19 candidate genes have formerly been identified in the CRC and other cancer types. It is demonstrated that the ADH1B gene polymorphism has a direct effect on colorectal carcinogenesis (19). The AHI-1 gene encodes the Jouberin protein, and its mutation is associated with the development of leukemia and lymphoma (20). The AKAP12 gene is one of the A-kinase scaffold proteins, and studies have indicated its suppressive role in the CRC (21). The frequent occurrence of BNIP3 methylation in the CRC suggests its possible contribution to tumorigenesis (22). Hypermethylation of CLDN11 plays a crucial role in CRC metastasis and poor survival (23). An earlier study suggested that DPP10 may play a vital role in the CRC progression and may be an independent prognostic marker in the CRC (24). The high expression of HEBP2 has been reported in the CRC (25). A previous study has reported the overexpression of MMP21 is associated with poor survival in the CRC (26). The PAFAH1B2 is a gene overexpressed in some types of cancers (27). The PCOLCE2 is found to be overexpressed in sporadic CRC and may play a role in cancer cell metastasis (28). The PIM1 gene is overexpressed in a variety of cancers such as prostate cancer and lymphoma (29), suggesting the contribution of PIM1 to the tumor development and

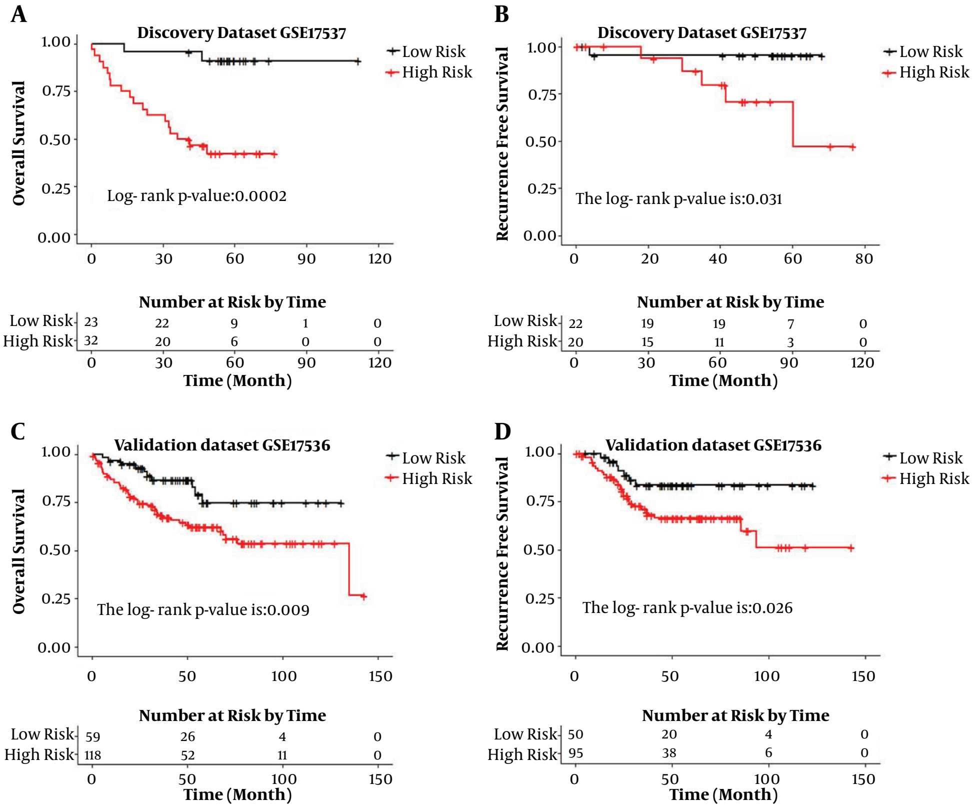

Figure 3. Kaplan-Meier analysis with two-sided log-rank test estimates of the OS and RFS of patients with CRC using the 19-mRNA signature. (A) Kaplan-Meier curves of OS for the discovery set patients; (B) Kaplan-Meier curves of RFS for the discovery set patients; (C) Kaplan-Meier curves of OS for the validation set patients; (D) Kaplan-Meier curves of RFS for the validation set patients.

progression of several types of tumors. One study has reported that the expression of TIPARP is decreased in ovarian cancer (30). Up-regulation of CST9L has been reported in pancreatic cancer (31). Low expression of FBXO33 gene is associated with a lower survival rate in renal cell carcinoma (32). Based on The Human Protein Atlas database (http://www.proteinatlas.org), the other genes, including INTS4, LIPJ, NGRN, TBKBP1, and TCEB3B, show a weak expression pattern in the CRC in immunohistochemistry staining.

The current study was a microarray-based transcriptome analysis. Analysis of microarray data is a challenging task because of a large number of genes measured ($p \sim 10^{3-4}$) vs. too small sample sizes ($n \sim 10^2$), and also the presence of group structure (high correlation) between genes (10). Penalized regression models such as LASSO (least absolute shrinkage and selection operator) are powerful tools to overcome the limitation of sample size in high dimensional data ($n < p$). In this study, we used a Bayesian penalized ordered probit model to analyze microarray data. Our model has some properties, which make it superior over other penalized regression models. It was performed in a Bayesian manner, which does not need to large sample assumption and could handle small datasets without losing the power and precision (33). It takes into account the group structure between genes in transcriptomics data and has the capability of variable selection at the group level. On the other hand, since the group structure between genes does not mean that all genes within a group are associated with the outcome; thus it is necessary to select genes within the groups. Consequently, variable the selection within a group results in the smallest set of the most important variables, which improves the prediction of the performance (9).

The main strength of our research is that we applied a more rational methodology to analyze microarray transcriptome data, and our signature model was derived from both tumor staging-related and survival-related mRNAs. Thus our finding may be more reliable for prognostic assessment of patients with CRC.

There are several limitations to this study. First, we investigated and discovered the biomarkers by using one mi-

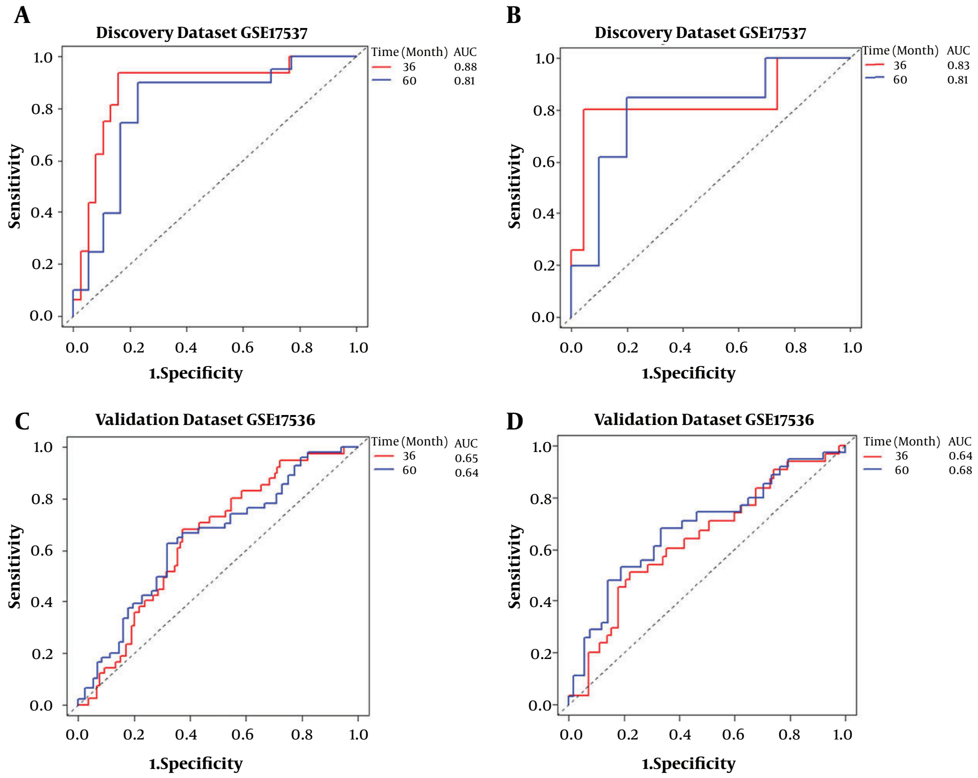

Figure 4. Time-dependent ROC curve analysis of the OS and RFS prediction based on the risk score with three and five years as the time points. (A) Time-dependent ROC curve of OS for the discovery set patients; (B) Time-dependent ROC curve of RFS for the discovery set patients; (C) Time-dependent ROC curve of OS for the validation set patients; (D) Time-dependent ROC curve of RFS for the validation set patients.

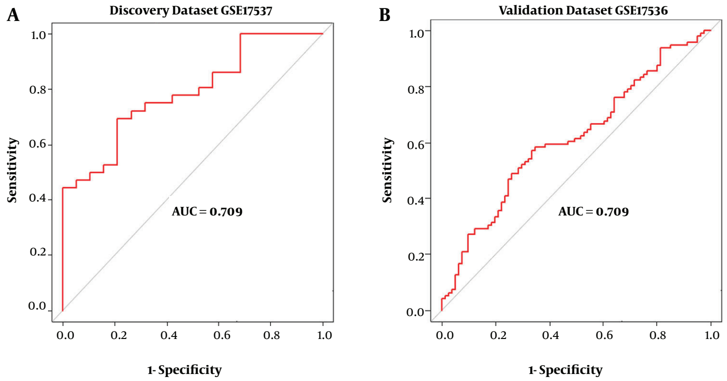

Figure 5. The ROC curve analysis of the sensitivity and specificity of discrimination of early- and late-stage by the 19-mRNA gene signature (A) for the discovery set and (B) the validation set

croarray data with limited sample size. Thus further validation using integrated microarray gene expression data to increase sample size may improve the statistical power and provide more robust results. Second, our study was conducted in a retrospective manner. Therefore, a comprehensive evaluation of these biomarkers with prospective randomized trials remains to be determined in future studies. Third, we defined the group structure of genes using the K-means approach. A biological clustering, when full pathway information is available, could improve the

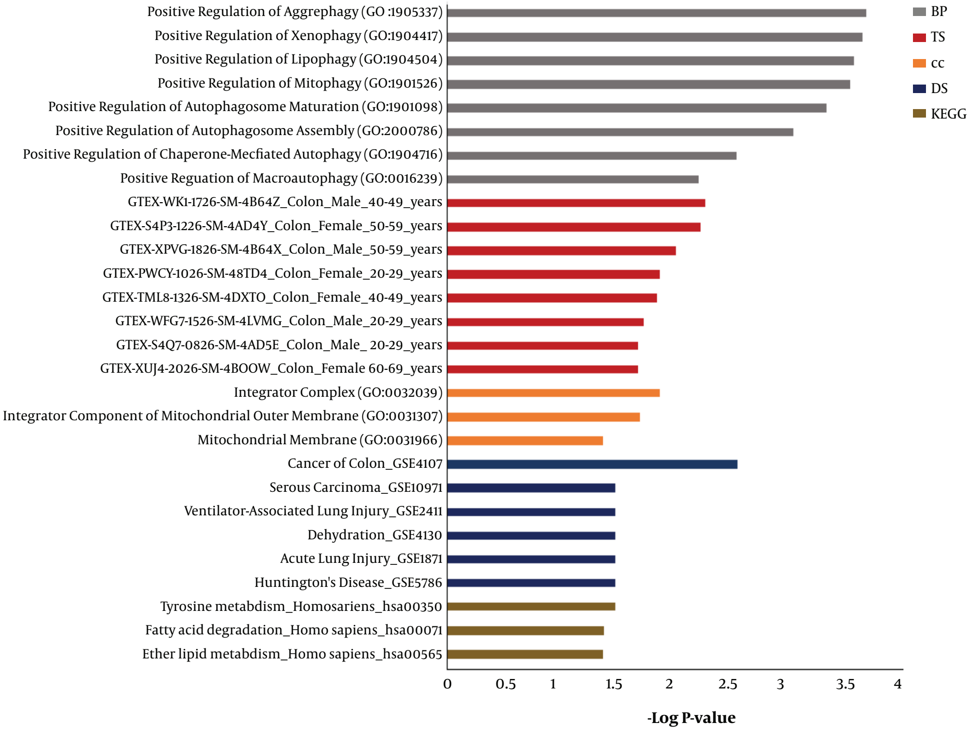

Figure 6. The GO and KEGG pathway enrichment analysis of the genes in the signature model. Abbreviations: GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; CC, cellular component; BP, biological process; DS, disease signature; TS, tissue sample.

results.

In conclusion, we identified a 19-mRNA signature model in this study, which may represent potential biomarkers for diagnosis and prognosis of CRC and helps to classify patients with CRC into low- or high-risk subgroups.

Footnotes

Authors' Contribution: Fatemeh Mohammadzadeh contributed in the design, concepts of the work, data acquisition, analysis, interpretation, drafting and revising the draft, approval of the final version of the manuscript, and agreed for all aspects of the work. Ebrahim Hajizadeh contributed in the design, concepts of the work, drafting and revising the draft, approval of the final version of the manuscript, and agreed for all aspects of the work. Aliakbar Rasekhi contributed in the analysis, revising the draft, approval of the final version of the manuscript, and agreed for all aspects of the work. Sadegh Azimzadeh Jamalkandi contributed in the analysis, interpretation, revising the draft, approval of the final version of the manuscript, and agreed for all aspects of the work.

Conflict of Interests: The authors declare that there is no conflict of interest.

Ethical Approval: This study exploited the collection or analysis of data and information freely available in the public domain and does not need ethical approval code.

Funding/Support: There was no financial support for this study.

References

1. Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global patterns and trends in colorectal cancer incidence and mortality. *Gut*. 2017;**66**(4):683–91. doi: 10.1136/gutjnl-2015-310912. [PubMed: 26818619].
2. Peters U, Bien S, Zubair N. Genetic architecture of colorectal cancer. *Gut*. 2015;**64**(10):1623–36. doi: 10.1136/gutjnl-2013-306705. [PubMed: 26187503]. [PubMed Central: PMC4567512].
3. Weitz J, Koch M, Debus J, Hohler T, Galle PR, Buchler MW. Colorectal cancer. *Lancet*. 2005;**365**(9454):153–65. doi: 10.1016/S0140-6736(05)17706-X. [PubMed: 15639298].
4. Woodward WA, Strom EA, Tucker SL, McNeese MD, Perkins GH, Schechter NR, et al. Changes in the 2003 American Joint Committee on Cancer staging for breast cancer dramatically affect stage-specific survival. *J Clin Oncol*. 2003;**21**(17):3244–8. doi: 10.1200/JCO.2003.03.052. [PubMed: 12947058].
5. Smith JJ, Deane NG, Wu F, Merchant NB, Zhang B, Jiang A, et al. Experimentally derived metastasis gene expression profile predicts

recurrence and death in patients with colon cancer. *Gastroenterology*. 2010;**138**(3):958–68. doi: 10.1053/j.gastro.2009.11.005. [PubMed: 19914252]. [PubMed Central: PMC3388775].

6. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;**4**(2):249–64. doi: 10.1093/biostatistics/4.2.249. [PubMed: 12925520].
7. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*. 2001;**98**(9):5116–21. doi: 10.1073/pnas.091062498. [PubMed: 11309499]. [PubMed Central: PMC33173].
8. Schwender H. *Siggenes: Multiple testing using SAM and Efron's empirical Bayes approaches*. R package version 1.12. 0; 2012.
9. Xu X, Ghosh M. Bayesian variable selection and estimation for group Lasso. *Bayesian Anal*. 2015;**10**(4):909–36. doi: 10.1214/14-ba929.
10. Ma S, Song X, Huang J. Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics*. 2007;**8**:60. doi: 10.1186/1471-2105-8-60. [PubMed: 17316436]. [PubMed Central: PMC1821041].
11. Tibshirani R, Hastie T, Eisen M, Ross D, Botstein D, Brown P. *Clustering methods for the analysis of DNA microarray data*. Stanford University; 1999.
12. Albert JH, Chib S. Bayesian analysis of binary and polychotomous response data. *J Am Stat Assoc*. 1993;**88**(422):669–79. doi: 10.1080/01621459.1993.10476321.
13. Sun X, Wang X, Feng W, Guo H, Tang C, Lu Y, et al. Gene signatures associated with drug resistance to irinotecan and oxaliplatin predict a poor prognosis in patients with colorectal cancer. *Oncol Lett*. 2017;**13**(4):2089–96. doi: 10.3892/ol.2017.5691. [PubMed: 28454366]. [PubMed Central: PMC5403337].
14. Qiu Z, Sun W, Gao S, Zhou H, Tan W, Cao M, et al. A 16-gene signature predicting prognosis of patients with oral tongue squamous cell carcinoma. *PeerJ*. 2017;**5**. e4062. doi: 10.7717/peerj.4062. [PubMed: 29158988]. [PubMed Central: PMC5695251].
15. Wan YW, Qian Y, Rathnagiriswaran S, Castranova V, Guo NL. A breast cancer prognostic signature predicts clinical outcomes in multiple tumor types. *Oncol Rep*. 2010;**24**(2):489–94. doi: 10.3892/or_00000883. [PubMed: 20596637]. [PubMed Central: PMC3095949].
16. Mettu RK, Wan YW, Habermann JK, Ried T, Guo NL. A 12-gene genomic instability signature predicts clinical outcomes in multiple cancer types. *Int J Biol Markers*. 2010;**25**(4):219–28. [PubMed: 21161944]. [PubMed Central: PMC3155635].
17. Mokarram P, Albokashy M, Zarghooni M, Moosavi MA, Sepehri Z, Chen QM, et al. New frontiers in the treatment of colorectal cancer: Autophagy and the unfolded protein response as promising targets. *Autophagy*. 2017;**13**(5):781–819. doi: 10.1080/15548627.2017.1290751. [PubMed: 28358273]. [PubMed Central: PMC5446063].
18. Nath A, Chan C. Genetic alterations in fatty acid transport and metabolism genes are associated with metastatic progression and poor prognosis of human cancers. *Sci Rep*. 2016;**6**:18669. doi: 10.1038/srep18669. [PubMed: 26725848]. [PubMed Central: PMC4698658].
19. Crous-Bou M, Rennert G, Cuadras D, Salazar R, Cordero D, Saltz Rennert H, et al. Polymorphisms in alcohol metabolism genes ADH1B and ALDH2, alcohol consumption and colorectal cancer. *PLoS One*. 2013;**8**(11). e80158. doi: 10.1371/journal.pone.0080158. [PubMed: 24282520]. [PubMed Central: PMC3839967].
20. Ringrose A, Zhou Y, Pang E, Zhou L, Lin AE, Sheng G, et al. Evidence for an oncogenic role of AHI-1 in Sezary syndrome, a leukemic variant of human cutaneous T-cell lymphomas. *Leukemia*. 2006;**20**(9):1593–601. doi: 10.1038/sj.leu.2404321. [PubMed: 16838023].
21. Liu W, Guan M, Hu T, Gu X, Lu Y. Re-expression of AKAP12 inhibits progression and metastasis potential of colorectal carcinoma in vivo and in vitro. *PLoS One*. 2011;**6**(8). e24015. doi: 10.1371/journal.pone.0024015. [PubMed: 21918680]. [PubMed Central: PMC3168868].
22. Agesen TH, Sveen A, Merok MA, Lind GE, Nesbakken A, Skotheim RI, et al. ColoGuideEx: A robust gene classifier specific for stage II colorectal cancer prognosis. *Gut*. 2012;**61**(11):1560–7. doi: 10.1136/gutjnl-2011-301179. [PubMed: 22213796].
23. Li J, Zhou C, Ni S, Wang S, Ni C, Yang P, et al. Methylated claudin-11 associated with metastasis and poor survival of colorectal cancer. *Oncotarget*. 2017;**8**(56):96249–62. doi: 10.18632/oncotarget.21997. [PubMed: 29221203]. [PubMed Central: PMC5707097].
24. Park HS, Yeo HY, Chang HJ, Kim KH, Park JW, Kim BC, et al. Dipeptidyl peptidase 10, a novel prognostic marker in colorectal cancer. *Yonsei Med J*. 2013;**54**(6):1362–9. doi: 10.3349/ymj.2013.54.6.1362. [PubMed: 24142639]. [PubMed Central: PMC3809881].
25. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, et al. The harmonizome: A collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database (Oxford)*. 2016;**2016**. doi: 10.1093/database/baw100. [PubMed: 27374120]. [PubMed Central: PMC4930834].
26. Huang Y, Li W, Chu D, Zheng J, Ji G, Li M, et al. Overexpression of matrix metalloproteinase-21 is associated with poor overall survival of patients with colorectal cancer. *J Gastrointest Surg*. 2011;**15**(7):1188–94. doi: 10.1007/s11605-011-1519-5. [PubMed: 21590459].
27. Ma C, Guo Y, Zhang Y, Duo A, Jia Y, Liu C, et al. PAFAH1B2 is a HIF1a target gene and promotes metastasis in pancreatic cancer. *Biochem Biophys Res Commun*. 2018;**501**(3):654–60. doi: 10.1016/j.bbrc.2018.05.039. [PubMed: 29758199].
28. Gutierrez ML, Corchete LA, Sarasquete ME, Del Mar Abad M, Bengoechea O, Ferminan E, et al. Prognostic impact of a novel gene expression profile classifier for the discrimination between metastatic and non-metastatic primary colorectal cancer tumors. *Oncotarget*. 2017;**8**(64):107685–700. doi: 10.18632/oncotarget.22591. [PubMed: 29296198]. [PubMed Central: PMC5746100].
29. Magnuson NS, Wang Z, Ding G, Reeves R. Why target PIM1 for cancer diagnosis and treatment? *Future Oncol*. 2010;**6**(9):1461–78. doi: 10.2217/fon.10.106. [PubMed: 20919829]. [PubMed Central: PMC3057053].
30. Goode EL, Chenevix-Trench G, Song H, Ramus SJ, Notaridou M, Lawrenson K, et al. A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. *Nat Genet*. 2010;**42**(10):874–9. doi: 10.1038/ng.668. [PubMed: 20852632]. [PubMed Central: PMC3020231].
31. Ansari D, Andersson R, Bauden MP, Andersson B, Connolly JB, Welinder C, et al. Protein deep sequencing applied to biobank samples from patients with pancreatic cancer. *J Cancer Res Clin Oncol*. 2015;**141**(2):369–80. doi: 10.1007/s00432-014-1817-x. [PubMed: 25216700].
32. Xinying HE, Wang S, Yu LI, Zhang G. Study on the relationship of MiR-2 5 targeting FBXO33 with cell apoptosis and prognosis in renal cell carcinoma. *J Mod Lab Med*. 2017;**32**(1):38–40.
33. van de Schoot R, Broere JJ, Perryck KH, Zondervan-Zwijnenburg M, van Loey NE. Analyzing small data sets using Bayesian estimation: The case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *Eur J Psychotraumatol*. 2015;**6**:25216. doi: 10.3402/ejpt.v6.25216. [PubMed: 25765534]. [PubMed Central: PMC4357639].