Am J Cancer Res 2023;13(7):3221-3233
www.ajcr.us /ISSN:2156-6976/ajcr0150391

Original Article

Predicting prognosis through the discovery of specific biomarkers according to colorectal cancer lymph node metastasis

Sung Won Park[1*], Junho Kang[2*], Hyung-Sik Kim[4], Sik Yoon[5], Byoung Soo Kim[6], Chaeseong Lim[7], Dongjun Lee[1], Yun Hak Kim[3,5]

[1]Department of Convergence Medical Science, School of Medicine, Pusan National University, Yangsan-si, Gyeongsangnam-do 50612, Republic of Korea; [2]Medical Research Institute, Pusan National University, Busan 46241, Republic of Korea; [3]Department of Biomedical Informatics, School of Medicine, Pusan National University, Yangsan-si, Gyeongsangnam-do 50612, Republic of Korea; [4]Department of Life Science in Dentistry, School of Dentistry, Pusan National University, Yangsan-si, Gyeongsangnam-do 50612, Republic of Korea; [5]Department of Anatomy, School of Medicine, Pusan National University, Yangsan-si, Gyeongsangnam-do 50612, Republic of Korea; [6]School of Biomedical Convergence Engineering, Pusan National University, Yangsan-si, Gyeongsangnam-do 50612, Republic of Korea; [7]Occupational and Environmental Medicine, Kosin University Gospel Hospital, Busan 46241, Republic of Korea. [*]Equal contributors.

Received March 28, 2023; Accepted July 8, 2023; Epub July 15, 2023; Published July 30, 2023

Abstract: Colorectal cancer (CRC) is a prevalent cancer worldwide, ranking as the third most common cancer and the second leading cause of cancer-related deaths. The presence or absence of lymph node metastases is one of the representative markers for predicting CRC prognosis, but often yields heterogeneous results. In this study, we conducted an integrative molecular analysis of CRC using publicly available data from The Cancer Genome Atlas database and NCBI's Gene Expression Omnibus. Through our analysis, we identified 372 upregulated genes that were differentially expressed in CRC patients. Additionally, Kyoto Encyclopedia of Genes and Genomes analysis revealed five significant pathways, including Hippo, FC-gamma, and forkhead box O signaling pathways, which are known to be associated with cancer. Survival analysis of 28 genes involved in these pathways led to the identification of 13 genes with prognostic significance ($P < 0.05$). To validate our findings, logistic regression models were generated and tested in multiple cohorts, demonstrating significant accuracy. Moreover, we identified six genes (*BNIP3*, *CD63*, *RDX*, *RGCC*, *WASF1*, and *WASF3*) whose combination predicted the best prognosis based on survival analysis. This predictive model holds promise as a potential biomarker for prognosis, survival, and treatment efficacy. In conclusion, our study provides valuable insights into the molecular characteristics of CRC and identifies prognostic biomarkers. The combination of differentially expressed genes and their involvement in cancer-related pathways enhances our understanding of CRC pathogenesis and opens avenues for personalized treatment approaches and improved patient outcomes.

Keywords: CRC, lymph node metastasis, TCGA, KEGG pathway, multivariate analysis, risk model

Introduction

Colorectal cancer (CRC) was the third most common malignancy worldwide and the second leading cause of cancer-related deaths in 2018 [1]. Currently, the aging population and risk factors for colorectal cancer, such as obesity, smoking, and lack of exercise, are expected to increase continuously, which, in turn, is expected to increase the incidence and mortality of CRC [2]. The American Joint Committee on Cancer tumor-node-metastasis (TNM) staging system is the standard for determining the prognosis of patients with CRC and is highly correlated with 5-year overall survival (OS). According to the TNM staging system, the 5-year survival rate of patients with stage I CRC is approximately 93%, which is reduced to approximately 80% for patients with stage II disease and 60% for patients with stage III [3]. The TNM staging system has reduced accuracy in patient groups with different prognoses, such as those receiving adjuvant chemotherapy, and the 5-year OS varies between 50% and 90% [4]. Although chemotherapy is universally recommended for all patients with stage III, patients

with stage IIIA have a significantly higher survival rate than those with stage IIB [3]. This highlights the need for a more accurate risk stratification of patients with stage III receiving adjuvant chemotherapy.

Microarray analysis can simultaneously evaluate the expression levels of approximately 25,000 genes and is one of the most common tools used to account for changes in gene expression levels [5-7]. Several studies from the early 2000s to the present have shown the potential of microarray analysis for predicting patient prognosis. For example, Arango et al. [8] have showed that the prognosis of patients with Dukes' C CRC is better than that of *TP53* and *KRAS* gene mutations through microarray analysis of patients with Dukes' C CRC, and Chang et al. [9] constructed the signatures of *GRB2*, *PTPN11*, *ITGB1*, and *POSTN* to confirm the predictability of risk groups in postoperative chemotherapy patients. In addition, commercial Oncotype Dx CRC, a relapse prediction signature based on the expression values of 18 genes, has been released and used in the past [10]. However, its application in actual clinical practice is limited due to limitations, such as overfitting of the discovery dataset, lack of sufficient validation, and heterogeneity between sequencing platforms [11, 12]. Therefore, when constructing a gene expression signature for application in clinical practice, it is essential to reduce the heterogeneity of expression values owing to the sequencing platform and validation in multiple cohorts.

In this study, we constructed a new prognostic signature to distinguish between low-risk and high-risk patients with stage III CRC using gene expression profiling data on the same sequencing platform from Gene Expression Omnibus (GEO), an open database, and validated it in The Cancer Genome Atlas (TCGA) and GEO cohorts.

Methods

Data collection and flowchart summarizing the study design

A flowchart summarizing the study design is shown in **Figure 1**. CRC gene expression data were downloaded from the open database GEO (https://gdc-portal.nci.nih. gov/) and the GDC data portal of TCGA (https://portal.gdc.cancer.gov). The datasets included in this study were GSE161158 [13], GSE14333 [14], GSE17536 [15], GSE40967 [16], and GSE17538 [17]; only datasets with at least 100 samples were included in our study. GSE161158, GSE14333, and GSE17536 were used for differentially expressed gene (DEG) screening, and GSE40967 and GSE17538 were used for risk score construction and validation. The GSE17538 and TCGA-COAD gene expression data and clinicopathological information were used as risk score validation datasets. There was inevitably a heterogeneity in the GEO data. To minimize heterogeneity between datasets, only the datasets used as the same sequencing platform were included in this study, and the sequencing platform used was GPL570 (HG-U133_Plus_2 Affymetrix Human Genome U133 Plus 2.0 Array). Detailed information on the datasets included in our study is presented in **Table 1**.

Transcriptome data preprocessing and DEG screening

Although various methods exist to reduce the batch effect, the Robust Multichip Average (RMA) method, which performs both background correction and normalization, is the currently accepted method for microarray data preprocessing. We performed RMA normalization on the microarray datasets included in this study. Differentially Expressed Gene (DEG)s were screened using the R package "*limma*" [18]. The criterion for DEGs was an adjusted *P*-value < 0.05; if it was met, these were considered as DEGs.

Prognosis-related gene selection through functional enrichment and survival analysis

To select prognosis-related genes, we performed survival analysis using Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and gene expression values. First, the gene set was configured based on the biological function, which was performed using the R package "*Clusterprofiler*" [19]. The biological function term selected in our study had an adjusted *P*-value < 0.05, and genes included in the top five terms were considered potential target genes. The prognostic relevance of the genes included in the top five terms was then evaluated. Genes were evaluated by dividing them into high-expression and low-expression groups. Cutoff Finder was used to determine the

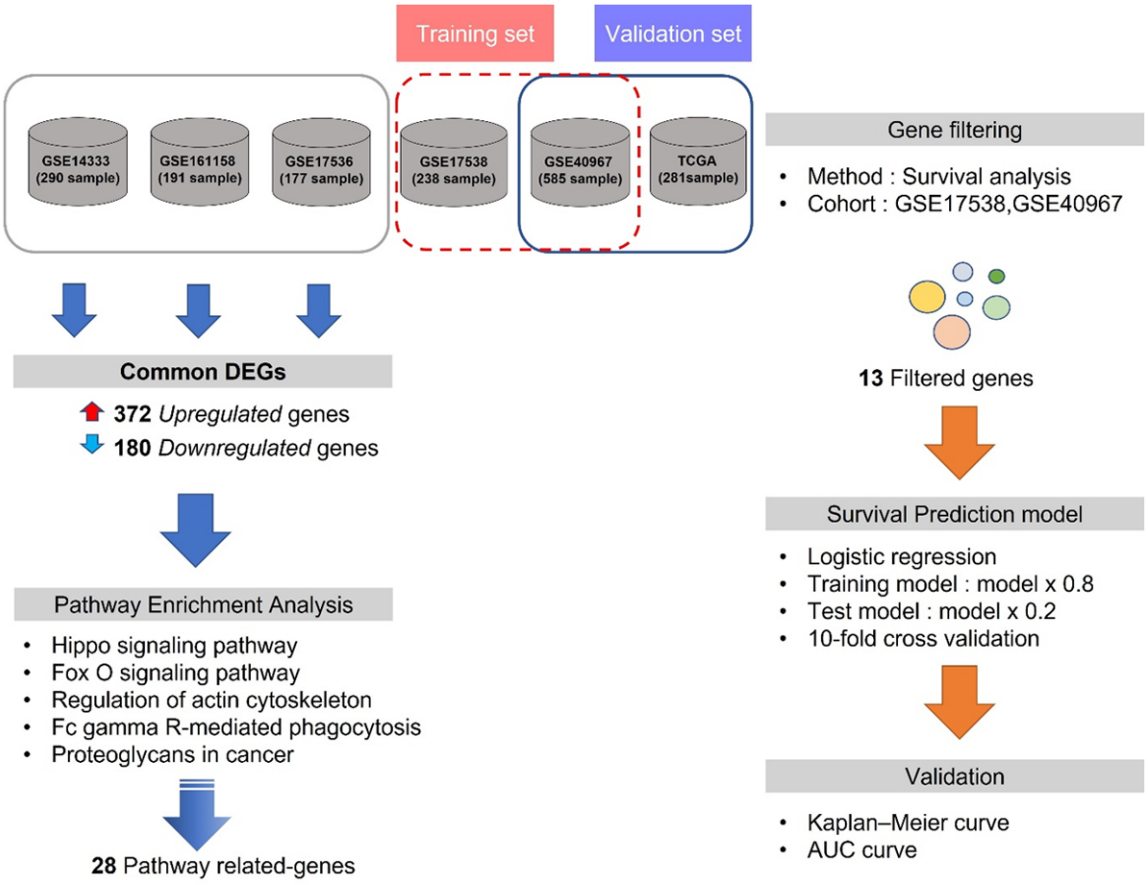

Figure 1. Study design and prognostic prediction model construction. This flowchart summarizes the study design and the construction of the prognostic prediction model. CRC gene expression data were obtained from the GEO database, including GSE14333, GSE161158, GSE17536, GSE17538, and GSE40967, as well as the TCGA dataset. Differential gene expression analysis identified a total of 552 differentially expressed genes, with 372 genes up-regulated and 180 genes down-regulated. Functional enrichment and survival analysis identified 13 candidate genes associated with prognosis. From these, a prognostic prediction model was constructed using the six genes with the best predictive rate. The model's performance was validated using cross-validation on GEO data and the TCGA dataset.

cutoff point for each gene expression value [20]. In addition, as survival analysis of individual genes has continuous variable data over time, Cox regression analysis was used for survival analysis. The analysis was performed using the Kaplan-Meier "*survival ROC*" package of the R package, and genes with $P < 0.05$ and $HR > 1$ in the high expression group versus the low expression group were included in the subsequent analysis [21].

Construction of prognosis prediction model through prognosis-related genes

To construct a prognosis prediction model, the GSE17538 cohort was divided into a training set and a validation set 8:2 through random sampling using the "caret" package [22]. Genes associated with prognosis were subjected to univariate and multivariate analysis using logistic regression, which was screened using the R packages "*glmnet*" and "*pROC*" [23, 24]. Logistic regression is a method of estimating the probability of data belonging to a certain category as a value between 0 and 1 and classifying the data into a more likely category according to that probability. This regression analysis method was chosen because it is commonly used to create predictive models using categorical data. In addition, to evaluate the discrimination power of gene combinations, ROC curves were created and AUROCs (Area Under the ROC Curves) were measured and compared. A risk model was constructed using the regression coefficients and expression values of genes significantly related to prognosis.

Table 1. Patients' characteristics

		GSE14333 n = 290	GSE161158 n = 191	GSE17536 n = 177	GSE40967 (validation) n = 585	TCGA (validation) n = 281
Stage	I	44 (Duke)	33	24 (ajcc)	-	43
	II	94 (Duke)	76	57 (ajcc)	-	104
	III	91 (Duke)	82	57 (ajcc)	-	81
Age	< 65	125	86	78	216	129
	≥ 65	165	105	99	369	152
Gender	Female	126	-	81	263	132
	Male	164	-	96	322	149
TNM stage	T0	-	-	-	1	6
	T1	-	-	-	12	42
	T2	-	-	-	49	197
	T3	-	-	-	379	35
	T4	-	-	-	119	-
	N0	138	109	81	314	162
	N1	152	82	96	137	73
	N2	--	-	-	100	46

The risk score was calculated as follows: Risk score = (0.6547 × *BNIP3*) + (4.9617 × *CD63*) + (1.7481 × *RDX*) + (1.7481 × *RGCC*) + (1.0393 × *WASF1*) + (-1.1305 × *WASF3*).

To validate the risk score, cross-validation was performed on GEO data with survival and TCGA data.

Results

Commonly upregulated genes and enriched pathways in CRC

To minimize heterogeneity arising from sequencing platform variations, we exclusively utilized three independent research datasets conducted on the same platform. This approach allowed us to ensure consistency in our analysis. Through our investigation, we discovered 3,776 genes (**Figure 2A**) in the GSE17536 dataset, 1,801 genes (**Figure 2B**) in GSE161158, and 1,914 genes (**Figure 2C**) in GSE14333. Subsequently, we identified genes that exhibited differential expression across all three datasets, resulting in a total of 552 common DEGs. Among these, 372 genes were commonly up-regulated (**Figure 2D**), while 180 genes were down-regulated (**Figure 2E**). Following the identification of these common DEGs, our objective was to determine their associated biological functions and the signaling pathways they participate in. To accomplish this, we conducted functional enrichment analysis. The top five enriched pathways of upregulated genes were the Hippo signaling pathway, forkhead box O (FOXO) signaling pathway, regulation of the actin cytoskeleton, Fc gamma R-mediated phagocytosis, and proteoglycans in cancer (**Figure 2F**). Notably, Hippo and FOXO signaling are pathways involved in cell physiological events, such as cell proliferation, apoptosis, and cell cycle regulation, indicating that their regulation is a key genomic event in CRC-associated tumors [25, 26]. In addition, Fc-gamma receptor signaling plays an immunomodulatory role as it is involved in the adaptive immune response by promoting antigen presentation or stimulating the secretion of inflammatory mediators [27]. Interestingly, Phosphatidylinositol 3-kinase is a gene involved in all pathways except Hippo signaling and is known as an oncogene in many studies [28]. These results led us to hypothesize the presence of general prognostic genomic biomarkers that control CRC-associated tumorigenesis. Proteoglycans in cancer and regulation of the actin cytoskeleton pathway, a pathway at the protein level, was abundant in CRC. Thus, we were able to identify the tendency for regulation to occur in various pathways, from the cellular level to proteins and immunity. Next, the top five enrichment pathways for downregulated genes included long-term potentiation, amphetamine addiction, and neurodegeneration pathways involved in neuronal signaling, along with

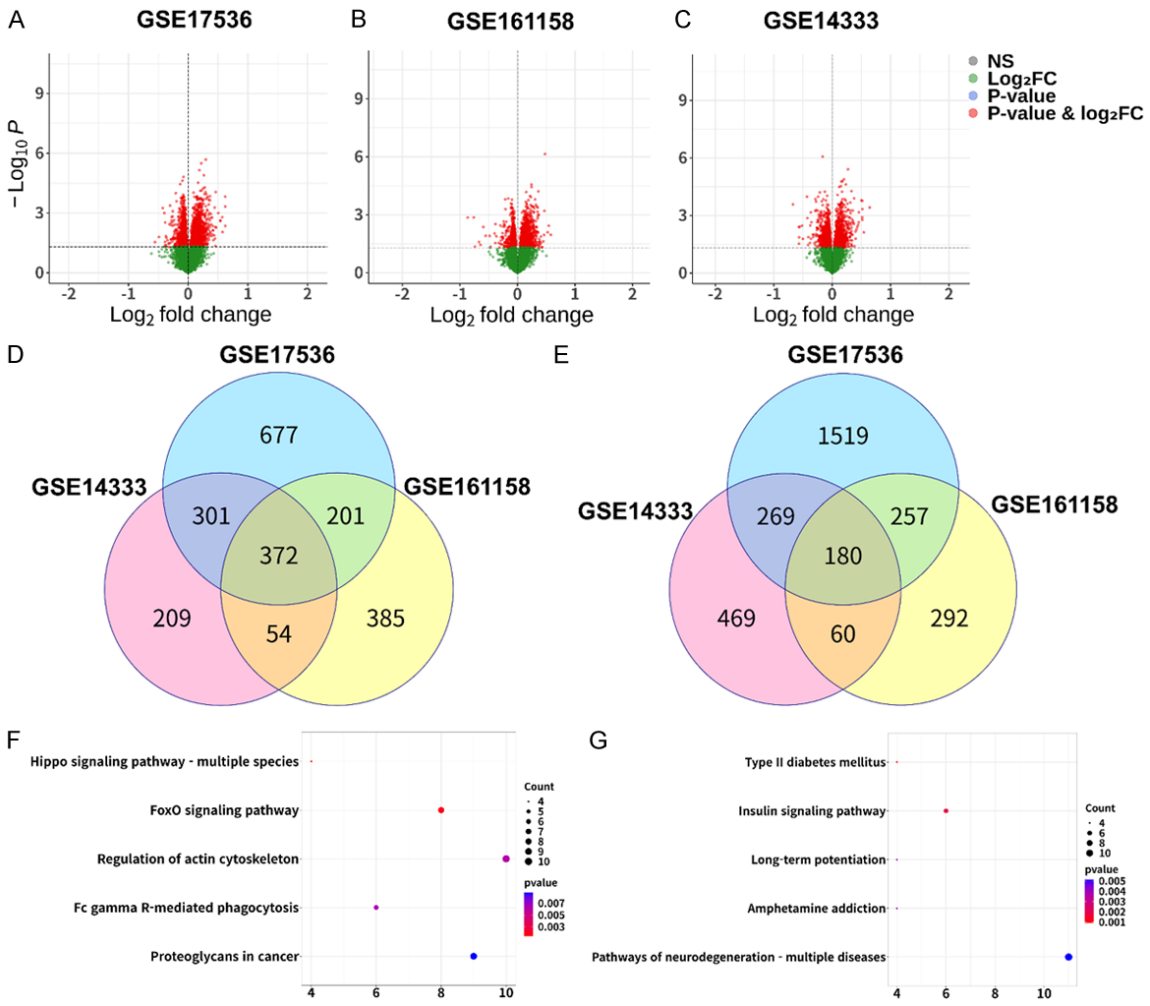

Figure 2. Commonly differentially expressed genes (DEGs) and enhanced pathways in colorectal cancer (CRC). A-C. Volcano plot visualizing DEGs between CRC lymph node-negative and CRC lymph node-positive data sets. D and E. 372 up-regulated genes and 180 down-regulated genes based on $P < 0.05$. F and G. Enhanced pathway using KEGG analysis for DEGs.

the type 2 diabetes and insulin signaling pathways involved in blood sugar control (**Figure 2G**). However, upregulated gene pathways are known to be more involved in tumorigenesis and progression than downregulated gene pathways, and 28 genes were selected from the upregulated gene pathways, where more often selected common genes were included in the pathway.

Variable selection and model development based on CRC data

Primary gene filtering was performed via survival analysis to identify prognostic genetic signatures. Thirteen of the initially selected 28 genes were selected based on a *P*-value < 0.05. Selected genes and their regression coefficients are listed in **Table 2**. Univariate Cox regression analysis was performed to identify hazard ratios (HRs) and confidence intervals (CIs) for selected genes. Logistic regression was used to select the prognostic variables related to the stage of CRC lymph node metastasis. To construct a prognosis prediction model, we used the regression coefficients and expression values of the selected genes in logistic regression analysis. In summary, we constructed a predictive model that exhibited the highest predictive power. The model incorporated a subset of 6 genes out of the initial 28 genes. The genes included in the model were *BNIP3*, *CD63*, *RDX*, *RGCC*, *WASF1*, and *WASF3*. **Figure 3** depicts the Kaplan-Meier plot result-

Table 2. Gene filter using survival analysis

	GE040967			GE017538		
	HR	95% CI	*P*-value	HR	95% CI	*P*-value
Gene						
BNIP3	0.64	0.47-0.88	0.0051	11.89	1.63-86.48	0.0019
CD63	0.66	0.49-0.9	0.0074	2.49	1.33-4.67	0.0034
FBXO32	1.85	1.12-3.04	0.014	2.31	1.3-4.13	0.0035
ITGB5	1.94	1.25-3.04	0.0029	2.32	1.08-4.95	0.026
PTK2	2.74	1.53-4.92	0.00043	1.99	1.22-3.54	0.0017
RDX	1.67	1.2-2.31	0.0018	2.33	1.32-4.13	0.0027
RGCC	1.99	1.31-3.01	0.00093	3.59	1.29-10.01	0.009
SDC2	2.7	1.33-5.49	0.0042	2.17	1.21-3.89	0.008
STK3	1.36	1.01-1.83	0.0045	2.61	1.45-4.73	0.00096
THBS1	4.42	2.07-9.43	0.00003	2.48	1.41-4.38	0.0012
WASF1	1.63	1.09-2.44	0.015	3.13	1.59-6.16	0.00051
WASF3	0.6	0.39-0.93	0.021	2.16	1.07-4.33	0.027
WWTR1	1.72	1.21-2.43	0.0029	3.03	1.6-5.73	0.00034

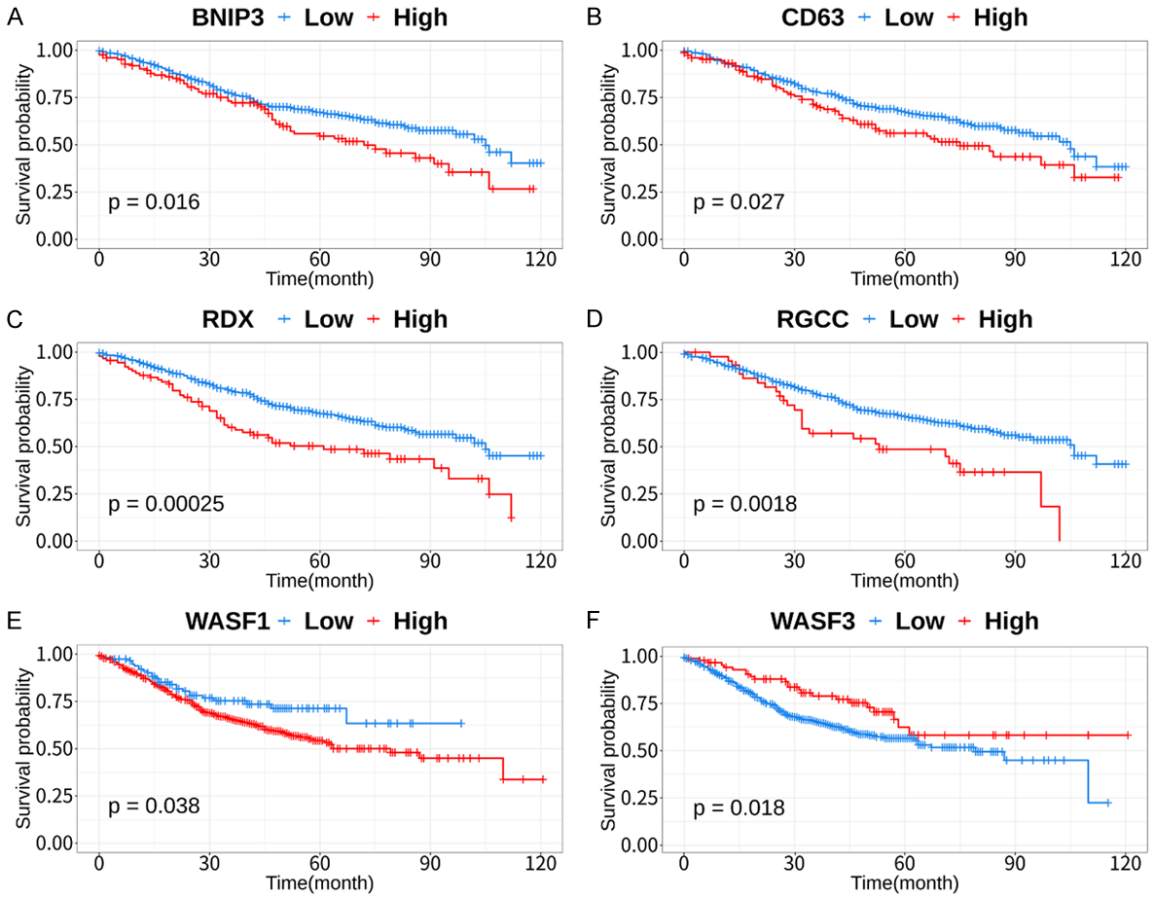

Figure 3. Survival analysis of individual genes included in the CRC survival prediction model. In the figure, red lines represent groups with high expression values for the respective gene, while blue lines indicate groups with low expression values. The x-axis represents the survival probability, and the y-axis represents the survival time. A. BNIP3; B. CD63; C. RDX; D. RGCC; E. WASF1; F. WASF3.

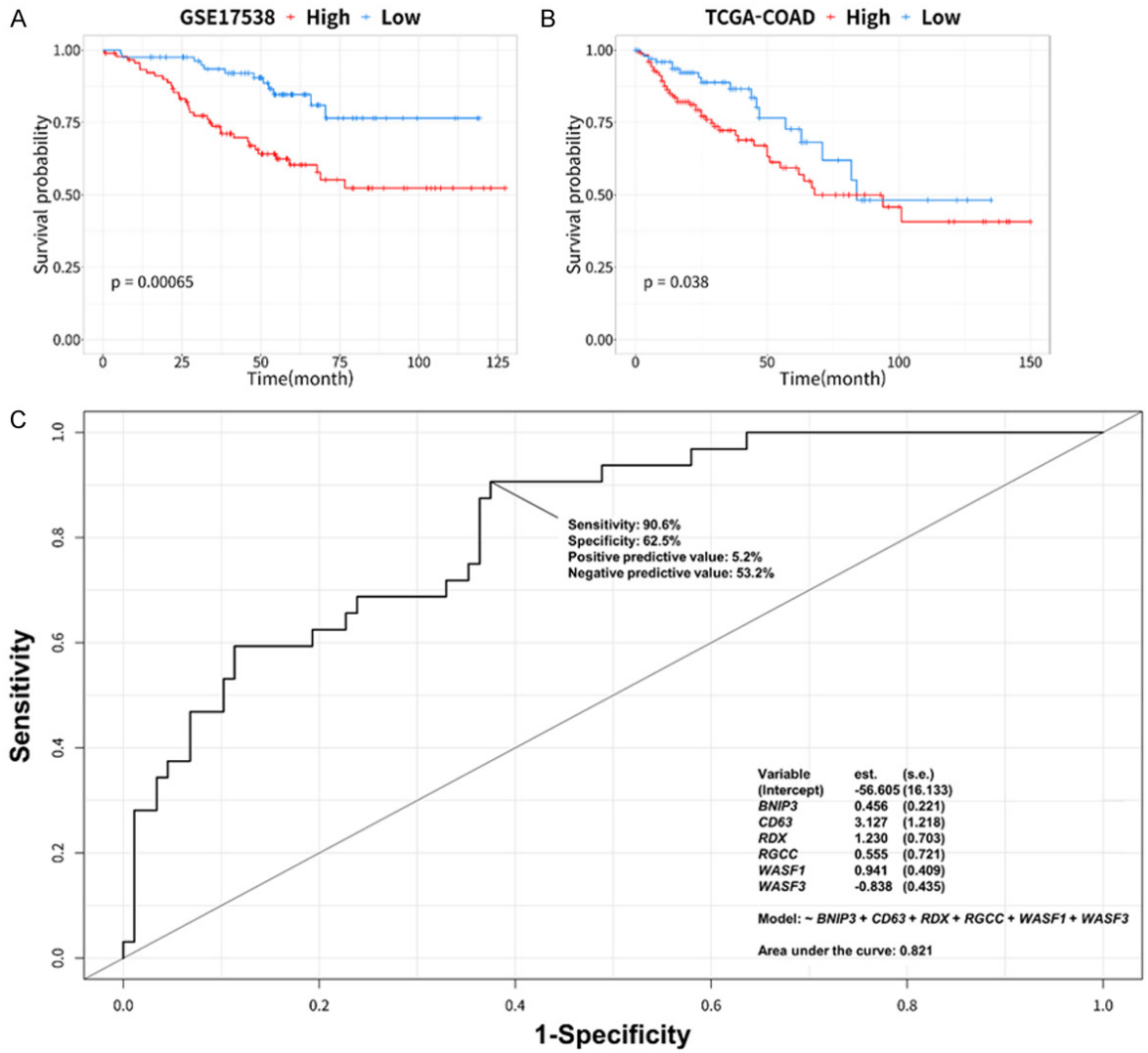

Figure 4. Validation of prognostic prediction models using external data. The result of survival analysis of the model created in the development cohort in the validation cohort. A. GEO data. B. TCGA-COAD data. C. ROC graph of the final model.

ing from the univariate survival analysis of these selected genes. The final risk model, using the calculated risk score, had a value of 78.59.

Validation of prognostic prediction models using external data

To confirm the prognostic significance, Kaplan-Meier survival analyses were performed using log-rank tests on one additional GSE17538 and TCGA dataset. The results of the validation cohort were significant, with P = 0.00065 for GSE17538 (**Figure 4A**) or P = 0.038 for TCGA (**Figure 4B**). In addition, as a result of checking the AUC by creating an ROC curve to evaluate the result, it improved to 0.821 (**Figure 4C**). These results indicated that our predictive model based on the identified prognostic genetic features had selective predictive potential along the stages of CRC to lymph node metastasis.

Validation of protein expression levels in CRC of prognostic genes

We used the Human Protein Atlas to examine the protein levels of six genes used in our prognostic prediction model for CRC. For comparison, we observed the stained results in glandular cells and tumor cells. In normal cell types, WASF1 showed a moderate intensity, while in

tumor cells, it exhibited a weak intensity in six cases, moderate intensity in seven cases, and strong intensity in nine cases. RDX was not detected in any normal cell types, but in tumor cells, it showed a negative intensity in ten cases, weak intensity in seven cases, and moderate intensity in two cases. BNIP3 exhibited a weak intensity in all normal cell types, whereas in tumor cells, it showed a weak intensity in five cases, moderate intensity in twelve cases, and strong intensity in two cases. RGCC exhibited a strong intensity in all normal cell types, while in tumor cells, it showed a moderate intensity in six cases and a strong intensity in thirteen cases. CD63 exhibited a weak intensity in all normal cell types, while in tumor cells, it showed a negative intensity in ten cases, weak intensity in eight cases, and moderate intensity in eight cases. WASF3 was not detected in any cells (**Figure 5**; **Table 3**).

Discussion

The global burden of CRC is projected to increase by 60%, resulting in 2.2 million new cases and 1.1 million deaths by 2030 [29]. In addition, since lymph flow in the primary tumor site was first identified, many studies have attempted to classify metastatic lymph nodes to accurately predict CRC [30]. Prognosis prediction using microarray analysis, which is one of the methods, is most often used through a change in gene expression level [5-7]. Examination of the expression patterns of CRC lymph node-associated genes in previous studies has identified unique molecular characteristics between lymph node-positive and negative tumors [31, 32]. However, the stage specificity of CRC lymph node-positive tumors remains unknown. Thus, we developed a prognostic gene signature based on the stage of CRC lymph node metastasis using multiple cohorts containing CRC lymph node information.

Cell bioactivity is regulated by complex networks that maintain a steady state from the cell cycle to proliferation [33]. When these pathways are damaged, cancer develops through cell damage. KEGG pathway enrichment analysis showed that genes upregulated in the CRC lymph node metastasis-positive group were included in the Hippo signaling pathway and Fox O signaling pathway. The Hippo signaling pathways are frequently deregulated in human cancers by controlling several cellular functions that are central to tumorigenesis, including proliferation and apoptosis [33]. In a subgroup of the AKT signaling pathway, the FOXO signaling pathway was found to be involved in tumorigenesis by phosphorylating and inactivating the FOXO transcription factor, thereby mediating the expression of genes important for apoptosis, such as the Fas ligand gene [25]. We also found that Fc-gamma receptor signaling was involved in the modulation of subsequent immune responses [34, 35]. Fc gamma receptor signaling is known to be involved in antitumor activity through the modulation of immunomodulatory antibody activity. This result is probably because the above pathway is associated with the accumulation of intracellular damage and the progression of tumorigenesis during the lymph node metastasis stage. Misregulation of the Insulin signaling pathway, particularly the pathway of downregulated genes, causes type 2 diabetes mellitus. Factors associated with insulin resistance, such as hyperinsulinemia, hyperglycemia, and hypertriglyceridemia, are also associated with CRC carcinogenesis [36]. These pathways are characterized by increased insulin concentrations during the early stages of the disease. Many prognostic predictors of patient survival have been developed for the gene expression profiles of patients with CRC based on clinical data, including the presence or absence of lymph node metastasis [36-39]. The six genes included in our prediction model, *BNIP3*, *CD63*, *RDX*, *RGCC*, *WASF1*, and *WASF3*, are crucial players in cell cycle progression, apoptosis regulation, migration, and adhesion, and have significant associations with various types of cancer. *BNIP3*, a member of the BCL2 family, is involved in apoptosis and autophagy regulation [40]. Altered *BNIP3* expression is strongly linked to clinical outcomes in cancer. Reduced *BNIP3* expression is associated with poor prognosis, aggressive tumor behavior, and decreased patient survival [41]. *CD63* plays a crucial role in cancer metastasis, enabling the spread of cancer cells from the primary tumor to distant sites [42]. Changes in *CD63* expression levels are closely associated with clinical outcomes and prognosis across various cancer types [43, 44]. *RDX*, a member of the ERM (ezrin-radixin-moesin) family of cytoskeletal proteins, is significantly associated with cancer metastasis and invasion [45]. A previous paper profiling

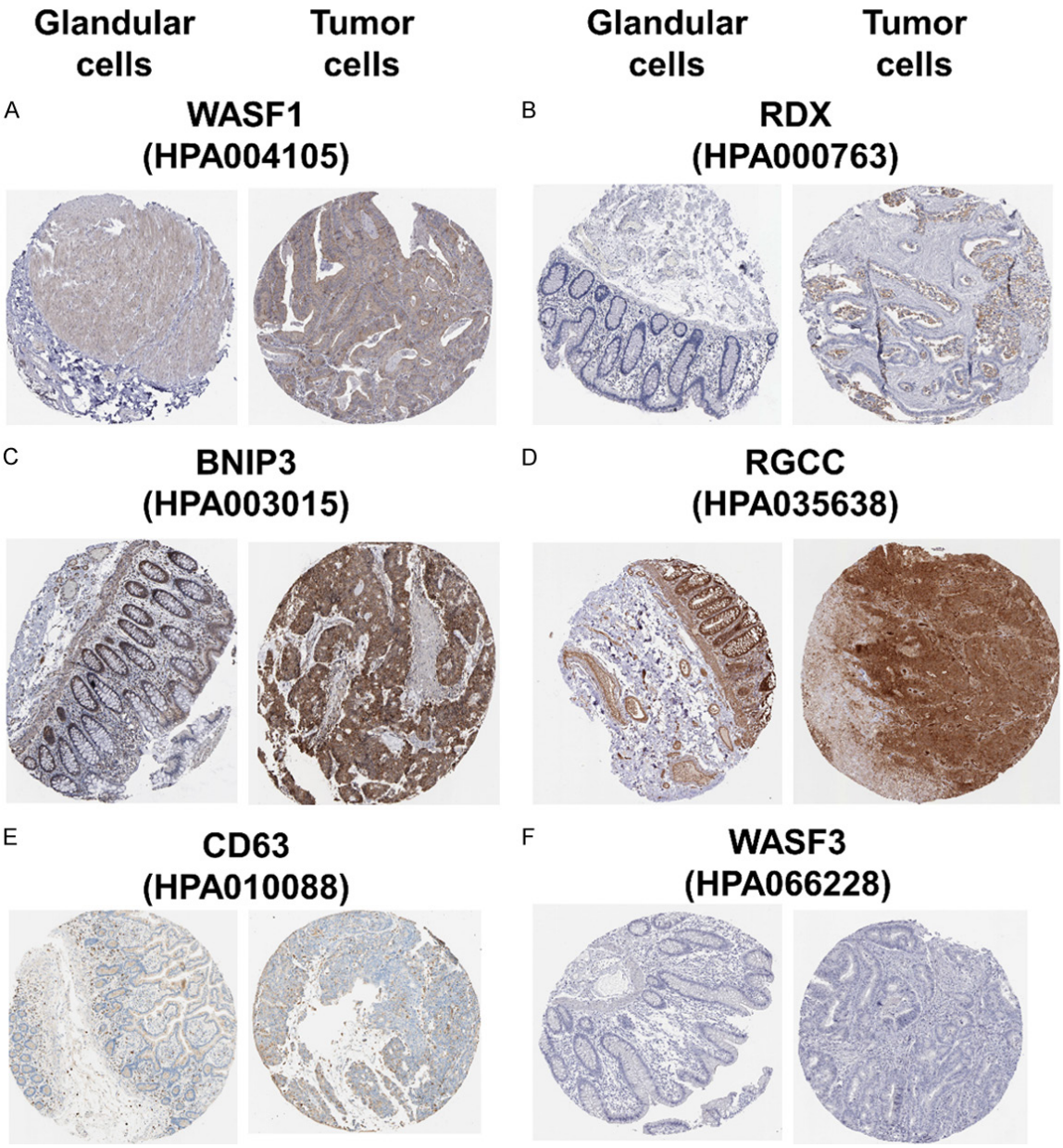

Figure 5. Protein expression patterns of six prognostic predictive genes from normal colon and primary colorectal tumor origin. A. Immunohistochemical staining of WASF1 protein showed moderate expression in normal colon tissue and moderate to high expression in CRC, representing 73% of cases. The highest staining intensity observed was strong, accounting for 33% of cases. B. Immunohistochemical staining of RDX protein showed no staining in normal colon tissue, while weak staining was predominant in CRC, accounting for 37% of cases. Moderate intensity staining was observed in 11% of cases. C. Immunohistochemical staining of BNIP3 protein showed weak expression in normal colon tissue, whereas moderate intensity staining was the most prevalent in CRC, accounting for 63% of cases. Strong intensity staining was observed in 11% of cases. D. Immunohistochemical staining of RGCC protein showed strong expression in normal colon tissue, while moderate intensity staining was the most prevalent in CRC, accounting for 31% of cases. The highest staining intensity observed was strong, accounting for 68% of cases. E. Immunohistochemical staining of CD63 protein showed weak expression in normal colon tissue, while weak intensity staining was the most prevalent in CRC, accounting for 33% of cases. Moderate intensity staining was observed in 25% of cases. F. Immunohistochemical staining of WASF3 protein showed negative staining in both normal and CRC tissues.

pancreatic cancer according to the presence or absence of lymph node metastasis confirmed that radixin had a significantly higher expression level at the protein level [46]. These results

Table 3. Intensity of protein staining in glandular cells and tumor cells of genes predicting CRC prognosis

Gene	Glandular cells		Tumor cells		Antibody
	Staining	Intensity	Staining	Intensity	
WASF1	Medium: 3	Moderate: 3	Not detected: 4 Low: 7 Medium: 8 High: 3	Weak: 6 Moderate: 7 Strong: 9	HPA004105
RDX	Not detected: 3	Negative: 3	Not detected: 17 Medium: 2	Negative: 10 Weak: 7 Moderate: 2	HPA000763
BNIP3	Low: 3	Weak: 3	Low: 5 Medium: 12 High: 2	Weak: 5 Moderate: 12 Strong: 2	HPA003015
RGCC	High: 2	Strong: 2	Medium: 6 High: 13	Moderate: 6 Strong: 13	HPA035638
CD63	Low: 3	Weak: 3	Not detected: 10 Low: 6 Medium: 4	Negative: 10 Weak: 8 Moderate: 6	HPA010088
WASF3	Not detected: 3	Negative: 3	Not detected: 24	Negative: 24	HPA066228

and our results suggest the possibility of *RDX* as a potential biomarker to predict the presence or absence of lymph node metastasis. *RGCC* (Response Gene to Complement-32) has garnered attention for its involvement in cancer [47]. Expression levels of *RGCC* have been studied as potential prognostic markers across multiple cancers [48]. Increased *RGCC* expression is correlated with unfavorable clinical outcomes, including shorter overall survival and disease-free survival in some cancer types. *WASF1*, also known as *WAVE1*, is a member of the WASP family and plays a critical role in tumor cell migration and invasion [49]. It contributes to the formation of invadopodia, specialized protruding structures that facilitate cancer cell invasion by breaking down the extracellular matrix. High *WASF1* expression has been suggested as a potential prognostic marker, as it is associated with worse prognosis, advanced cancer stage, and reduced overall survival in certain cases [50]. *WASF3*, also known as *WAVE3*, is another member of the WASP family and is involved in metastasis across various cancer types [49]. WASF3 drives the process of epithelial-mesenchymal transition (EMT), allowing cancer cells to acquire invasiveness and metastatic capabilities. High *WASF3* expression is associated with worse prognosis, advanced disease stage, and, in some cases, reduced overall survival [51]. These genes, *BNIP3*, *CD63*, *RDX*, *RGCC*, *WASF1*, and *WASF3*, play pivotal roles in cancer biology and have demonstrated significant clinical relevance as potential prognostic markers and targets for therapeutic intervention in metastatic cancers. However, studies that can predict the prognosis at each stage in patients with positive lymph node metastasis are lacking. Therefore, we developed a set of prognostic genes for lymph node metastasis. Various regression analyses have been performed to develop prognostic gene signatures. Among them, we used logistic regression, which is the most commonly used method for risk analysis. Risk scores were calculated for 6 out of 13 significant gene combinations. The validation results using the risk scores were mostly validated using additional validation sets.

Conclusion

This study utilizesd multiple cohorts to establish and validate the prognostic genetic characteristics of lymph node metastasis in patients with CRC. A prognostic predictive model based on six gene combination features suggests that it may play an important role in developing a step-by-step treatment strategy, even in the lymph node-positive status of patients with CRC. This model ensures a similar reproducibility in other patients with CRC. Collectively, these findings are expected to provide effective results as potential biomarkers for prognosis, survival, and treatment.

Acknowledgements

This work was supported by grants from the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021-R1A2C4001466 and 2022R1A5A2027161 for D.L, 2018R1A5A2023879 for Y.H.K, and 2021R1A6A3A01087108 for JK), the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) (HI22C1377), funded by the Ministry of Health & Welfare, Republic of Korea, the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2023-00223764), and the Manufacturing Human Cell-based Artificial Blood and Platform Technology Development for Transfusion, funded by the Multi-Ministrial Research Project, Republic of Korea (grant number: HX23C1692).

Disclosure of conflict of interest

None.

Abbreviations

CRC, Colorectal cancer; TNM, Tumor-Node-Metastasis; TCGA, The Cancer Genome Atlas; GEO, Gene Expression Omnibus; RMA, Robust Multichip Average; DEG, Differentially Expressed Gene; KEGG, Kyoto Encyclopedia of Genes and Genomes; AUROCs, Area Under the ROC Curves; FOXO, forkhead box O; HRs, Hazard Ratios; CIs, Confidence Intervals.

Address correspondence to: Dongjun Lee, Department of Convergence Medical Science, School of Medicine, Pusan National University, Rm#502, 49 Busandaehak-ro, Mulgeum-eup, Yangsan-si, Gyeongsangnam-do 50612, Republic of Korea. Tel: +82-51-510-8042; Fax: +82-51-510-8526; E-mail: lee.dongjun@pusan.ac.kr; Yun Hak Kim, Department of Biomedical Informatics, School of Medicine, Pusan National University, Rm#701, 49 Busandaehak-ro, Mulgeum-eup, Yangsan-si, Gyeongsangnam-do 50612, Republic of Korea. Tel: +82-51-510-8091; Fax: +82-51-510-8049; E-mail: yunhak10510@pusan.ac.kr

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA and Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018; 68: 394-424.

[2] Dekker E, Tanis PJ, Vleugels JLA, Kasi PM and Wallace MB. Colorectal cancer. Lancet 2019; 394: 1467-1480.

[3] O'Connell JB, Maggard MA and Ko CY. Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging. J Natl Cancer Inst 2004; 96: 1420-1425.

[4] Howlader N. Seer cancer statistics review, 1975-2008, national cancer institute, bethesda, md. http://seer. cancer. gov/csr/1975_2008/, based on November 2010 SEER data submission, posted to the SEER web site 2011.

[5] Stoughton RB. Applications of DNA microarrays in biology. Annu Rev Biochem 2005; 74: 53-82.

[6] Ramsay G. DNA chips: state-of-the art. Nat Biotechnol 1998; 16: 40-44.

[7] Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H and Brown EL. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996; 14: 1675-1680.

[8] Arango D, Laiho P, Kokko A, Alhopuro P, Sammalkorpi H, Salovaara R, Nicorici D, Hautaniemi S, Alazzouzi H, Mecklin JP, Jarvinen H, Hemminki A, Astola J, Schwartz S Jr and Aaltonen LA. Gene-expression profiling predicts recurrence in Dukes' C colorectal cancer. Gastroenterology 2005; 129: 874-884.

[9] Chang W, Gao X, Han Y, Du Y, Liu Q, Wang L, Tan X, Zhang Q, Liu Y, Zhu Y, Yu Y, Fan X, Zhang H, Zhou W, Wang J, Fu C and Cao G. Gene expression profiling-derived immunohistochemistry signature with high prognostic value in colorectal carcinoma. Gut 2014; 63: 1457-1467.

[10] Webber EM, Lin JS and Whitlock EP. Oncotype DX tumor gene expression profiling in stage II colon cancer. Application: prognostic, risk prediction. PLoS Curr 2010; 2: RRN1177.

[11] Qi L, Chen L, Li Y, Qin Y, Pan R, Zhao W, Gu Y, Wang H, Wang R, Chen X and Guo Z. Critical limitations of prognostic signatures based on risk scores summarized from gene expression levels: a case study for resected stage I non-small-cell lung cancer. Brief Bioinform 2016; 17: 233-242.

[12] Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K and Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 2010; 11: 733-739.

[13] Szeglin BC, Wu C, Marco MR, Park HS, Zhang Z, Zhang B, Garcia-Aguilar J, Beauchamp RD, Chen XS and Smith JJ. A SMAD4-modulated gene profile predicts disease-free survival in stage II and III colorectal cancer. Cancer Rep (Hoboken) 2022; 5: e1423.
[14] Jorissen RN, Gibbs P, Christie M, Prakash S, Lipton L, Desai J, Kerr D, Aaltonen LA, Arango D, Kruhoffer M, Orntoft TF, Andersen CL, Gruidl M, Kamath VP, Eschrich S, Yeatman TJ and Sieber OM. Metastasis-associated gene expression changes predict poor outcomes in patients with dukes stage B and C colorectal cancer. Clin Cancer Res 2009; 15: 7642-7651.
[15] Smith JJ, Deane NG, Wu F, Merchant NB, Zhang B, Jiang A, Lu P, Johnson JC, Schmidt C, Bailey CE, Eschrich S, Kis C, Levy S, Washington MK, Heslin MJ, Coffey RJ, Yeatman TJ, Shyr Y and Beauchamp RD. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. Gastroenterology 2010; 138: 958-968.
[16] Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, Vescovo L, Etienne-Grimaldi MC, Schiappa R, Guenot D, Ayadi M, Kirzin S, Chazal M, Fléjou JF, Benchimol D, Berger A, Lagarde A, Pencreach E, Piard F, Elias D, Parc Y, Olschwang S, Milano G, Laurent-Puig P and Boige V. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med 2013; 10: e1001453.
[17] Williams CS, Bernard JK, Demory Beckler M, Almohazey D, Washington MK, Smith JJ and Frey MR. ERBB4 is over-expressed in human colon cancer and enhances cellular transformation. Carcinogenesis 2015; 36: 710-718.
[18] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W and Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015; 43: e47.
[19] Yu G, Wang LG, Han Y and He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012; 16: 284-287.
[20] Budczies J, Klauschen F, Sinn BV, Győrffy B, Schmitt WD, Darb-Esfahani S and Denkert C. Cutoff Finder: a comprehensive and straightforward Web application enabling rapid biomarker cutoff optimization. PLoS One 2012; 7: e51862.
[21] Kamarudin AN, Cox T and Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: current methods and applications. BMC Med Res Methodol 2017; 17: 53.
[22] Kuhn M. Caret: classification and regression training. Astrophysics Source Code Library 2015; ascl: 1505.1003.
[23] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC and Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011; 12: 77.
[24] Ma S, Meng Z, Chen R and Guan KL. The Hippo pathway: biology and pathophysiology. Annu Rev Biochem 2019; 88: 577-604.
[25] Farhan M, Wang H, Gaur U, Little PJ, Xu J and Zheng W. FOXO signaling pathways as therapeutic targets in cancer. Int J Biol Sci 2017; 13: 815-827.
[26] Bournazos S, Gupta A and Ravetch JV. The role of IgG Fc receptors in antibody-dependent enhancement. Nat Rev Immunol 2020; 20: 633-643.
[27] Vanhaesebroeck B, Perry MWD, Brown JR, André F and Okkenhaug K. Author correction: PI3K inhibitors are finally coming of age. Nat Rev Drug Discov 2021; 20: 798.
[28] Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A and Bray F. Global patterns and trends in colorectal cancer incidence and mortality. Gut 2017; 66: 683-691.
[29] Kim HJ and Choi GS. Clinical implications of lymph node metastasis in colorectal cancer: current status and future perspectives. Ann Coloproctol 2019; 35: 109-117.
[30] Hayashi N, Ito I, Nakamura Y, Yanagisawa A, Kato Y, Nakamori S, Imaoka S, Watanabe H and Ogawa M. Genetic diagnosis of lymph-node metastasis in colorectal cancer. Lancet 1995; 345: 1257-1259.
[31] Beaton C, Twine CP, Williams GL and Radcliffe AG. Systematic review and meta-analysis of histopathological factors influencing the risk of lymph node metastasis in early colorectal cancer. Colorectal Dis 2013; 15: 788-797.
[32] Harvey KF, Zhang X and Thomas DM. The Hippo pathway and human cancer. Nat Rev Cancer 2013; 13: 246-257.
[33] Altomare DA and Testa JR. Perturbations of the AKT signaling pathway in human cancer. Oncogene 2005; 24: 7455-7464.
[34] Mellor JD, Brown MP, Irving HR, Zalcberg JR and Dobrovic A. A critical review of the role of Fc gamma receptor polymorphisms in the response to monoclonal antibodies in cancer. J Hematol Oncol 2013; 6: 1.
[35] Larsson SC, Orsini N and Wolk A. Diabetes mellitus and risk of colorectal cancer: a meta-analysis. J Natl Cancer Inst 2005; 97: 1679-1687.
[36] Shida H, Ban K, Matsumoto M, Masuda K, Imanari T, Machida T and Yamamoto T. Prognostic significance of location of lymph node metastases in colorectal cancer. Dis Colon Rectum 1992; 35: 1046-1050.
[37] Glasgow SC, Bleier JI, Burgart LJ, Finne CO and Lowry AC. Meta-analysis of histopathological

features of primary colorectal cancers that predict lymph node metastases. J Gastrointest Surg 2012; 16: 1019-1028.
[38] Chen K, Collins G, Wang H and Toh JWT. Pathological features and prognostication in colorectal cancer. Curr Oncol 2021; 28: 5356-5383.
[39] Andres-Terre M, McGuire HM, Pouliot Y, Bongen E, Sweeney TE, Tato CM and Khatri P. Integrated, multi-cohort analysis identifies conserved transcriptional signatures across multiple respiratory viruses. Immunity 2015; 43: 1199-1211.
[40] Zhang J and Ney PA. Role of BNIP3 and NIX in cell death, autophagy, and mitophagy. Cell Death Differ 2009; 16: 939-946.
[41] Shimizu S, Iida S, Ishiguro M, Uetake H, Ishikawa T, Takagi Y, Kobayashi H, Higuchi T, Enomoto M, Mogushi K, Mizushima H, Tanaka H and Sugihara K. Methylated BNIP3 gene in colorectal cancer prognosis. Oncol Lett 2010; 1: 865-872.
[42] Chen Z, Mustafa T, Trojanowicz B, Brauckhoff M, Gimm O, Schmutzler C, Köhrle J, Holzhausen HJ, Kehlen A, Klonisch T, Finke R, Dralle H and Hoang-Vu C. CD82, and CD63 in thyroid cancer. Int J Mol Med 2004; 14: 517-544.
[43] Kaprio T, Hagström J, Andersson LC and Haglund C. Tetraspanin CD63 independently predicts poor prognosis in colorectal cancer. Histol Histopathol 2020; 35: 887-892.
[44] Koh HM, Jang BG and Kim DC. Prognostic value of CD63 expression in solid tumors: a meta-analysis of the literature. In Vivo 2020; 34: 2209-2215.
[45] Jiang QH, Wang AX and Chen Y. Radixin enhances colon cancer cell invasion by increasing MMP-7 production via Rac1-ERK pathway. ScientificWorldJournal 2014; 2014: 340271.
[46] Cui Y, Wu J, Zong M, Song G, Jia Q, Jiang J and Han J. Proteomic profiling in pancreatic cancer with and without lymph node metastasis. Int J Cancer 2009; 124: 1614-1621.
[47] Zhang H, Su Y, Jia J and Wang Q. RGCC is a prognostic biomarker and correlates with immune infiltrates in breast cancer. 2023.
[48] Jarnuczak AF, Najgebauer H, Barzine M, Kundu DJ, Ghavidel F, Perez-Riverol Y, Papatheodorou I, Brazma A and Vizcaino JA. An integrated landscape of protein expression in human cancer. Sci Data 2021; 8: 115.
[49] Alekhina O, Burstein E and Billadeau DD. Cellular functions of WASP family proteins at a glance. J Cell Sci 2017; 130: 2235-2241.
[50] Sowalsky AG, Sager R, Schaefer RJ, Bratslavsky G, Pandolfi PP, Balk SP and Kotula L. Loss of Wave1 gene defines a subtype of lethal prostate cancer. Oncotarget 2015; 6: 12383-91.
[51] Limaye AJ, Whittaker MK, Bendzunas GN, Cowell JK and Kennedy EJ. Targeting the WASF3 complex to suppress metastasis. Pharmacol Res 2022; 182: 106302.