Computational and Structural Biotechnology Journal 19 (2021) 3922–3929

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Using bioinformatics approaches to investigate driver genes and identify BCL7A as a prognostic gene in colorectal cancer

Jeffrey Yung-chuan Chao [a,b], Hsin-Chuan Chang [c], Jeng-Kai Jiang [d,e], Chih-Yung Yang [f,g,h], Fang-Hsin Chen [i,j,k], Yo-Liang Lai [l,m], Wen-Jen Lin [n], Chia-Yang Li [o], Shu-Chi Wang [p], Muh-Hwa Yang [a,p], Yu-Feng Lin [q,*], Wei-Chung Cheng [m,r,s,*]

[a] *Institute of Clinical Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan*
[b] *Department of Radiation Oncology, Taichung Veterans General Hospital, Taichung, Taiwan*
[c] *Institute of Microbiology and Immunology, National Yang Ming Chiao Tung University, Taipei, Taiwan*
[d] *Division of Colon & Rectal Surgery, Department of Surgery, Taipei Veterans General Hospital, Taipei, Taiwan*
[e] *School of Medicine, National Yang-Ming University, Taipei, Taiwan*
[f] *Department of Teaching and Research, Taipei City Hospital, Taipei, Taiwan*
[g] *Commission for General Education, National United University, Miaoli, Taiwan*
[h] *General Education Center, University of Taipei, Taipei, Taiwan*
[i] *Department of Medical Imaging and Radiological Sciences, Chang Gung University, Taoyuan, Taiwan*
[j] *Department of Radiation Oncology, Chang Gung Memorial Hospital at Linkou, Taoyuan, Taiwan*
[k] *Institute for Radiological Research, Chang Gung University and Chang Gung Memorial Hospital, Taoyuan, Taiwan*
[l] *Department of Radiation Oncology, China Medical University Hospital, Taichung, Taiwan*
[m] *Graduate Institute of Biomedical Science, China Medical University, Taichung, Taiwan*
[n] *Graduate Institute of Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan*
[o] *Department of Medical Laboratory Science and Biotechnology, College of Health Sciences, Kaohsiung Medical University, Kaohsiung 80708, Taiwan*
[p] *Division of Medical Oncology, Taipei Veterans General Hospital, Taipei 112, Taiwan*
[q] *Department of Medical Laboratory Science and Biotechnology, College of Medical and Health Science, Asia University, Taichung, Taiwan*
[r] *The Ph.D. Program for Cancer Biology and Drug Discovery, China Medical University and Academia Sinica, Taichung 404, Taiwan*
[s] *Research Center for Cancer Biology, China Medical University, Taichung, Taiwan*

ARTICLE INFO

Article history:
Received 10 March 2021
Received in revised form 19 June 2021
Accepted 28 June 2021
Available online 01 July 2021

Keywords:
Driver genes
Colorectal cancer
Prognostic genes
Next generation sequencing
Cancer panel

ABSTRACT

Colorectal cancer (CRC) results from the uncontrolled growth of cells in the colon, rectum, or appendix. The 5-year relative survival rate for patients with CRC is 65% and is correlated with the stage at diagnosis (being 91% for stage I at diagnosis versus 12% for stage IV). This study aimed to identify CRC driver genes to assist in the design of a cancer panel to detect gene mutations during clinical early-stage screening and identify genes for use in prognostic assessments and the evaluation of appropriate treatment options. First, we utilized bioinformatics approaches to analyze 354 paired sequencing profiles from The Cancer Genome Atlas (TCGA) to identify CRC driver genes and analyzed the sequencing profiles of 38 patients with >5 years of follow-up data to search for prognostic genes. The results revealed eight driver genes and ten prognostic genes. Next, the presence of the identified gene mutations was verified using tissue and blood samples from Taiwanese CRC patients. The results showed that the set identified gene mutations provide high coverage for driver gene screening, and *APC, TP53, PIK3CA*, and *FAT4* could be detected in blood as ctDNA test targets. We further found that *BCL7A* gene mutation was correlated with prognosis in CRC (log-rank p-value = 0.02), and that mutations of *BCL7A* could be identified in ctDNA samples. These findings may be of value in clinical early cancer detection, disease monitoring, drug development, and treatment efforts in the future.

* Corresponding authors at: Graduate Institute of Biomedical Science, China Medical University, No. 100, Sec. 1, Jingmao Rd., Beitun Dist., Taichung 406040, Taiwan (Wei-Chung Cheng); Department of Medical Laboratory Science and Biotechnology, College of Medical and Health Science, Asia University, No. 500, Lioufeng Rd., Wufeng, Taichung 41354, Taiwan (Yu-Feng Lin).

E-mail addresses: yflin@asia.edu.tw (Y.-F. Lin), wccheng@mail.cmu.edu.tw (W.-C. Cheng).

https://doi.org/10.1016/j.csbj.2021.06.044

1. Introduction

The prevalence of colorectal cancer (CRC) is quite high, with 776,120 men and 768,650 women diagnosed with CRC in the United States as of January 1, 2019 [1]. New case numbers were estimated at 104,610 and 43,340 for colon cancer and rectal cancer, respectively, diagnosed in 2020 in the US, according to an American Cancer Society report [2]. CRC, which is also known as bowel cancer, results from uncontrolled cell growth in the colon, rectum, or appendix. Most CRCs arise from adenomatous polyps. These neoplasms are usually benign, but some will further develop into cancer. The 5-year relative survival rate for patients with CRC is 65% (67% for rectal cancer and 64% for colon cancer). For CRC patients diagnosed with stage I and stage II disease, the 5-year relative survival rates are 91% and 82%, respectively. However, for patients diagnosed with stage IV CRC, the 5-year survival rate declines to 12% [1]. In other words, most long-term CRC survivors are diagnosed in the early stages, indicating the importance of early diagnosis and continuous monitoring. Regular screening can prevent death caused by CRC.

The current CRC screening methods have various limitations. For example, stool-based tests have a high false-positive rate (50%), and guaiac-based fecal occult blood tests (gFOBT) have sensitivity levels of only 24–40% [3]. One more advanced method, fecal immunochemical tests (FITs), can detect sources of lower gastrointestinal bleeding [4] but cannot prove cancer. A variety of serum markers have been associated with CRC, particularly carcinoembryonic antigen (CEA) and carbohydrate antigen 19–9 (CA 19–9). However, they have a low diagnostic ability to detect primary CRC due to significant overlap with benign disease, as well as low sensitivity for early-stage CRC [5]. Colonoscopy is considered the gold standard for CRC screening; however, studies have indicated that colonoscopy will miss 2% to 6% of CRCs [6], some of which may be challenging to detect due to a proximal location or depressed appearance. Furthermore, colonoscopy is an invasive procedure that requires patients to undergo rigorous bowel preparation and the risk of pore formation, underscoring the need for a non-invasive method for diagnosing CRC.

Next-generation sequencing (NGS) is currently widely used in cancer genomics research. Many large-scale cancer projects, such as The Cancer Genome Atlas (TCGA), the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) project, the International Cancer Genome Consortium (ICGC), and the Pediatric Cancer Genome Project (PCGP), use NGS techniques to generate large amounts of cancer genome data that can be used to form the basis for the development of precision medicines [7]. In recent studies, scientists have coined the term "driver mutations" to describe mutations known to cause cancer development and clonal expansion. By contrast, mutations that accompany cancer progression but do not cause cancer development and do not affect the cancer cells are referred to as "passenger mutations" [8–10]. Identifying driver mutations in cancer-related genes associated with colorectal cancer progression has been a central aim for the development of clinical early-stage detection techniques.

Based on the results of fundamental research studies combined with new mutation detection technologies, the effectiveness of colon cancer screening panels has significantly increased, but their use is still currently limited to scientific research and personal testing, with only a few panels having received or still awaiting FDA approval. For scientific research purposes, Qiagen applies GeneRead DNASeq Targeted Panels V2 for CRC (<100 genes), while ThermoFisher Scientific applies the Ion AmpliSeq™ Colon and Lung Cancer Research Panel. For personal testing, Ambry Genetics™ applies ColoNext, which includes a 20-gene panel[11]. The Invitae company (https://www.invitae.com/en/), meanwhile, offers tests that can help people detect whether they have genetic mutations that are risk factors for several types of cancers, but not CRC.

Otherwise, HalioDx, a new French company, has developed a cancer assay, Immunoscore, that is an in vitro risk-assessment tool that provides independent and superior prognostic value and can predict the risk of relapse in early-stage colon cancer patients by measuring the host immune response at the tumor site [12]. However, as alluded to above, only two colon tests have obtained FDA approval thus far, namely, Cologuard [13], which was approved by the FDA in August 2014, and Epigenomics' proColon, which is the first and only FDA-approved blood-based test for the detection of CRC [14]. Cologuard is a stool-based triple test method that includes a FIT test, a DNA mutation and methylation test, and a β-actin test, plus a hemoglobin immunoassay. Meanwhile, proColon is a blood-based CRC screening method that detects methylated SEPT9 in the blood. However, the coverage and the false-positive rates of these two tests remain unknown. We continue to require a more convenient and accurate screening method for the identification of early-stage CRC. Using gene mutations to detect cancer can improve the precision and specificity of cancer screening techniques. In addition to driver genes, the correlation between cancer prognostication and gene mutations has been studied [15–17]. The detection of the prognostic genes can aid drug selection and surgery evaluations in determining medical treatment strategies.

This study sought to identify driver mutations to develop an effective, early-stage cancer detection panel for CRC patients and identify prognostic genes for the evaluation of treatment options and prognostic assessments. First, we collected 354 CRC tumor-normal pairs sequencing profiles from TCGA and used the DriverDB [18] and Ingenuity Pathways Analysis (IPA) software (Ingenuity Systems, Redwood City, CA, USA; http://www.ingenuity.com/) to find an eight-driver gene signature in CRC. To identify prognostic genes, we utilized the sequencing profiles of CRC patients in combination with prognostic data from TCGA, which resulted in ten genes that we classified as prognostic genes. Next, these eighteen genes, including eight driver genes and ten prognostic genes, were verified using tissue and blood samples from Taiwanese CRC patients. The results showed that the identified set of genes provides high coverage for driver gene screening, and one of the identified prognostic genes, *BCL7A*, was associated with a significant difference (log-rank p-value = 0.02) in survival. These findings could be used to improve CRC screening and provide guidance for drug development and treatment efforts in the future.

2. Material and methods

2.1. Data collection and mutation annotation

We collected data for 354 CRC patients, including 266 colon adenocarcinoma (COAD) and 88 rectum adenocarcinoma (READ) patients, with both tumor and normal sequencing data from our own database, DriverDB [18], in 2015. In brief, the mutation data, including single-nucleotide variants and insertions/deletions, of these patients were downloaded from the Cancer Genomics Hub [19] and mapped to published databases, including the dbSNP [20], 1000 genomes [21], NHGRI GWAS catalog [22], COSMIC [23], NHLBI GO ESP [24], ClinVar [25], and OMIM [26] databases, in order to annotate known variants. To predict the impacts of various mutations, such as synonymous mutations, frame-shift mutations, and stop/gain mutations, we used the SnpEff [27] and VEP [28] tools.

2.2. Identification of CRC driver genes

Based on mutation data from TCGA, we utilized 3 bioinformatics tools, as described in our previous study [18]: MutsigCV [29], Dendrix [30], and ActiveDriver [31], using default parameters. For ActiverDriver, the genes with a FDR adjusted P value < 0.05 are statistically unexpected mutated in protein phosphorylation sites or protein kinase domains. For MutsigDV, the genes with an FDR-adjusted P-value < 0.05 are significantly mutated. Dendrix identifies large numbers of genes in different K, where K represents a geneset within a module/pathway. The genes reported in at least 10% of modules in any K are identified as significant, based on our previous study [18]. The common genes identified by the three tools being selected for further analysis. For the COAD and READ dataset in TCGA, there were 177 and 144 genes commonly identified by the 3 tools, respectively. Forty-seven genes were identified in both of the two datasets.

2.3. Identification of CRC prognostic genes

We identified 38 CRC patients with at least 5 years of follow-up data or who died within 5 years of data collection by TCGA in 2015. We divided these 38 patients into 2 groups, according to >5-year and <5-year survival times, into "Good" and "Poor" survival groups, respectively. To identify prognostic genes in CRC, we used the 3 described bioinformatics tools to identify significantly mutated genes in the two groups. Ten significant genes identified in the good prognosis group but not the poor prognosis group were regarded as prognostic genes for CRC in this study.

2.4. Clinical samples collection

There were 56 tissue samples and nine blood samples collected from 56 Taiwan CRC patients who served as the validation cohort in this study. The appropriate clinical samples for molecular analysis need to be considered in both qualitative and quantitative terms. This study was approved by the Institutional Review Board (IRB) of Taipei Veterans General Hospital (CF13041A-2), and all the participating patients signed and provided their informed consent before sample collection. During surgery, the tumor samples and adjacent normal parts were collected simultaneously by Dr. Jeng-Kae Jiang at Taipei Veterans General Hospital.

2.5. Genetic network analysis and canonical pathway analysis

The differential gene expression profiles were imported into the Ingenuity Pathways Analysis (IPA) software (Qiagen) to obtain functional regulatory networks and canonical pathways. The "core function" can interpret gene interaction networks based on its Ingenuity Knowledge Base (IKB), which is constructed by integrating findings from the past literature. It can in turn generate the gene-gene interaction/relationships and the pathways of focused genes by importing interesting genes to IPA. For canonical pathway analysis, IPA can help us discover genes incorporated into biological functions/pathways and generate a list of statistical scores as P values, indicating the connection strength between gene sets and pathways.

2.6. Target sequencing of the driver genes and prognostic genes

Utilizing Illumina TruSeq Custom Amplicon (TSCA), the amplicon design service provided by Illumina Inc., San Diego, CA, USA, we designed a target sequencing gene set consisting of the 8 driver genes and 10 prognostic genes (Fig. 1). Then, we performed target sequencing on the DNA from the tissue and blood samples of the aforementioned 56 Taiwanese CRC patients. According to the manufacturer's protocol, libraries were prepared with TSCA. Preparation of these libraries was performed starting with up to 50 ng of genomic DNA, the adjacent normal tissue sample corresponding to the normal reference, or hg19 when germline DNA from the patient was not available. Briefly, the gDNA samples from different patients were amplified by numerous PCR primer sets designed by Illumina DesignStudio (http://www.illumina.com/informatics/research/experimental-design/designstudio.html), and the amplified DNA fragments were simultaneously added to adaptor sequences. The size distribution of the purified libraries was assessed using the 2100 Bioanalyzer with DNA assay kit (Agilent Technologies, Santa Clara, CA). The fluorescent method and Q-PCR (KAPA, USA, cat. KK4903, and KK4603) were used to quantify the libraries. The libraries were then sequenced using Illumina MiSeq sequencer (Illumina Inc., San Diego, CA, USA).

Fastq files generated by an Illumina sequencer for each sample were mapped to the human reference genome (assembly version hg19) using BWA mapper [32] with default parameters. The reads' alignment was saved as BAM files; these were subsequently sorted depending on their mapping position with the "sort" command in the software package SAMtools [33]. PCR duplicates were removed from the BAM files using the "rmdup" command of the SAMtools package. Mutations were annotated by a custom script printing out gene information and the variation effect (synonymous, non-synonymous, stop-loss/gain, frame-shift). The raw data of the Taiwanese cohort has been deposited into the SRA database [34], with accession number PRJNA711905.

3. Results

3.1. Identifying driver genes in CRC

We collected tumor-normal paired sequencing data for 266 colon adenocarcinoma (COAD) and 88 rectum adenocarcinoma (READ) patients from the TCGA portal. The tumor-normal paired sequencing data of these 354 patients was then used to find CRC mutated genes, including those predicting variant effects, and to find variants that might cause amino acid changes or lead to protein dysfunction (the "Driver Genes Identification" module in Fig. 1). These mutations were further mapped to the gene-level to predict driver genes. After identifying driver genes, 177 genes and 134 genes were annotated from the COAD and READ patients, respectively, and 47 genes that were annotated from both the COAD and READ patients were considered the common driver genes in CRC (Fig. 2A). These 47 genes are listed in Table S1. By taking advantage of IPA, we examined the relationships among these common driver genes in CRC. Eight genes related to COAD were recognized, namely, *APC, BRAF, CDH8, FAT4, KRAS, NEFH, PIK3CA*, and *TP53* (P-value = 7.12E-07) (Fig. 2B). These 8 driver gene candidates were mapped back to the original TCGA data and showed a high coverage rate in COAD (93.98%) and READ (94.32%). The coverage rates of these eight genes in the 354 patients were as follows: *APC* (300/354), *TP53* (249/354), *PIK3CA* (177/354), *KRAS* (172/354), *NEFH* (108/354), *BRAF* (96/354), *FAT4* (67/354) and *CDH8* (33/354) (Fig. 2D). These results indicated that these genes could cover most of the CRC patients. The gene network analysis of the aforementioned 47 genes showed that 5 of the 8 driver genes, including *APC, BRAF, KRAS, TP53,* and *PIK3CA*, had hub characteristics that were connected by more than three edges in our IPA results (Fig. 2C).

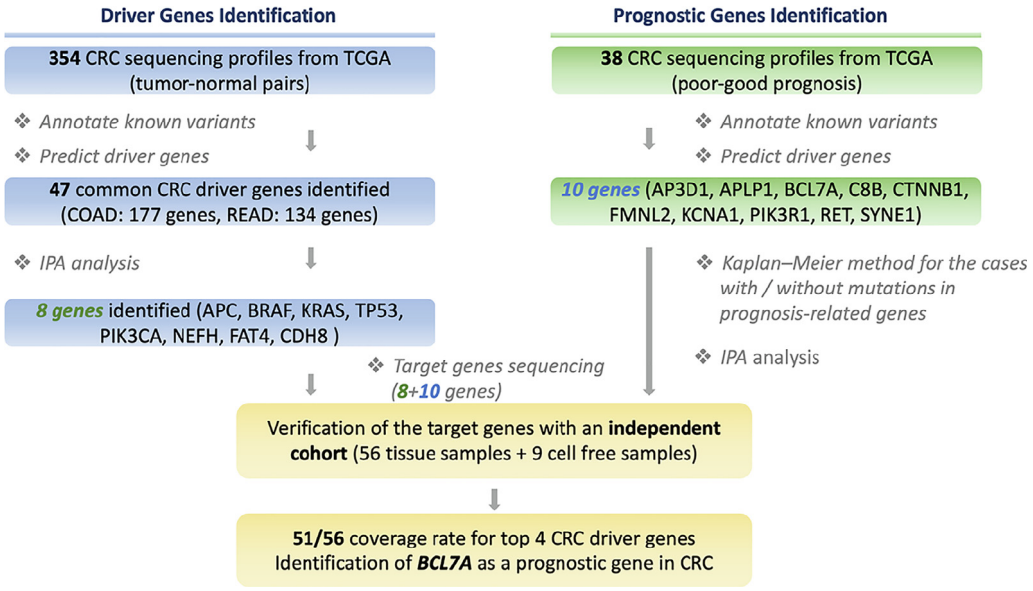

Fig. 1. Flowchart showing the identification of driver genes and prognostic genes. The left "Driver Gene Identification" module showed that the eight CRC driver genes were identified from 354 CRC paired sequencing profiles. The ten prognostic genes were selected by taking advantage of 38 CRC sequencing profiles from TCGA, as shown in the right "Prognostic Genes Identification" module. An independent cohort (56 Taiwanese patients) was used as the validation cohort.

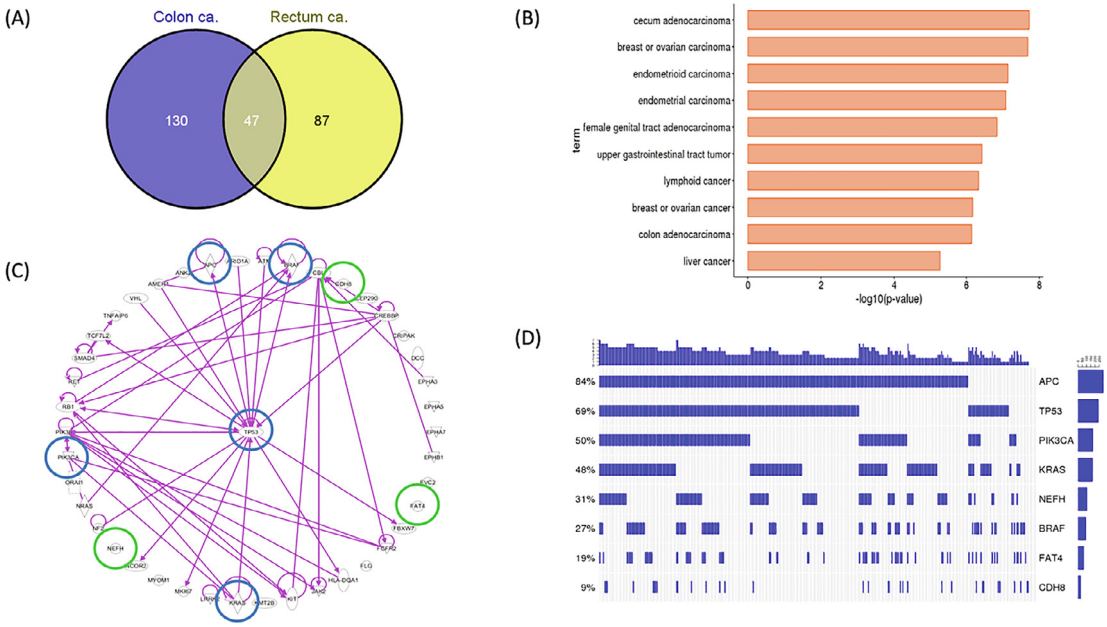

Fig. 2. The results of CRC driver gene identification. (A) Forty-seven driver genes intersected between colon cancer and rectal cancer. (B) Bar chart for functional analysis by IPA. (C) The genes *APC, TP53, KRAS, PIK3CA,* and *BRAF* play central roles in CRC (blue circles). *NEFH, FAT4*, and *CDH8* were also chosen as CRC driver genes due to their potential to contribute to CRC development (green circles). (D) The Driver Gene Coverage Rate in Caucasians. The coverage rate of the eight selected genes in 354 CRC patients was 94.07% (333/354). Each gene's coverage rate: *APC* (84.75%), *TP53* (70.34%), *PIK3CA* (50%), *KRAS* (48.59%), *NEFH* (30.51%), *BRAF* (27.12%), *FAT4* (18.93%), and *CDH8* (9.32%).

3.2. Validation of identified CRC driver genes by using Taiwanese samples

We collected clinical samples from 56 CRC patients at Taipei Veterans General Hospital in Taiwan. The baseline information of these 56 CRC patients is listed in Table 1. In terms of gender, 57.1% of the patients were male, and 42.9% were female, indicating no substantial gender imbalance in this data. The pathological examination results showed that most of these patients had T3N0M0 stage adenocarcinomas. After identifying the CRC driver genes, we designed a mutated gene sequencing set containing the eight genes related to CRC development and performed targeted sequencing with this focused gene set. The results revealed a high coverage rate of 96.4% in the 56 patient tissue samples and coverage of eight out of nine patient blood samples (Fig. 3). The coverage rates for the eight genes in the tissue samples were 33/56 (*APC*), 26/56 (*FAT4*), 23/56 (*KRAS*), 21/56 (*TP53*), 16/56 (*CDH8*), 12/56 (*PIK3CA*), 10/56 (*NEFH*), and 5/56 (*BRAF*). Thirty of the 56 patients had at least three mutations while 15 of the patients only had one gene mutation in this mutated gene sequencing set. Among these 15 patients, the *KRAS* gene was mutated in five patients, three patients had *APC* gene mutations, and two patients each had *TP53* and *FAT4* mutations. For the nine blood samples, the circulating cell-free tumor DNA (ctDNA) was

Table 1
The baseline information of colorectal cancer patients from Taiwan cohort.

Characteristics		Numbers (%)
Gender		
	Male	32 (57.1)
	Female	24 (42.9)
T Stage		
	1	1 (1.8)
	2	2 (3.6)
	3	51 (91.1)
	4	2 (3.6)
N Stage		
	0	51 (91.1)
	1	2 (3.6)
	2	3 (5.4)
M stage		
	0	54 (96.4)
	1	2 (3.6)
Histologic Type		
	Adenocarcinoma	55 (98.2)
	Carcinoid	1(1.8)
Histologic differentiation		
	Well differentiated	1(1.8)
	Moderately differentiated	54 (96.4)
	Undifferentiated	1(1.8)
Recurrence		
	YES	24 (42.9)
	NO	32 (57.1)
Vital status		
	DECEASED	34 (60.7)
	LIVING	22 (39.3)
BCL7A mutation		
	YES	12 (21.4)
	NO	44 (78.6)

tested, and the results showed that the driver gene mutations could be detected in the blood samples (Fig. 3B). The mutated eight-gene sequencing results showed that four gene mutations (*APC, TP53, PIK3CA,* and *FAT4*) could be detected in plasma. The *APC* and *TP53* genes have been used as target genes in cancer screening using the ColoNext method of Ambry Genetics™, but the *PIK3CA* and *FAT4* genes have not been used in that approach. Therefore, based on our findings, *PIK3CA* and *FAT4* could be new targets for ctDNA analysis development in CRC.

3.3. Identifying prognostic genes in CRC

To find the genes correlated to prognostication, we collected characteristic clinical data from TCGA. According to their survival times, the sequencing profiles of 38 CRC patients were classified into two groups: a good prognosis group and a poor prognosis group. By taking advantage of DriverDB, we annotated the known genes and predicted the driver genes. We looked for genes in the poor prognosis group and filtered out the genes that intersected in the good and poor prognosis groups simultaneously. In the end, we identified ten prognostic genes, namely, *AP3D1, APLP1, BCL7A, C8B, CTNNB1, FMNL2, KCNA1, PIK3R1, RET,* and *SYNE1* (the "Prognostic Genes Identification" module in Fig. 1). To evaluate the power of the ten prognostic genes, the 38 patients were divided into patients with the prognostic gene mutations and patients without the prognostic gene mutations. The Kaplan–Meier survival curve showed a significant difference (log-rank P = 0.0027) between these two groups (Fig. 4A). This result meant that the survival rate of the patients with the prognostic gene mutations was significantly decreased. The IPA showed that nine out of these ten genes were identified as abdominal adenocarcinoma-related genes (P = 1.26E-04) and digestive system cancer-related genes (P = 1.33E-03), while four of the genes (*CTNNB1, PIK3R1, RET,* and *SYNE1*) were correlated with rectal cancer (P = 1.47E-04) (Fig. 4B).

3.4. CRC prognostic genes validation in Taiwanese samples

The ten genes identified as prognostic genes were then tested in the 56 patient samples we collected. The results showed a coverage rate of 71.4% in these 56 patient tissue samples (Fig. 4C) and coverage of six out of nine blood patient samples (Fig. S1). The mutations of seven of the ten genes could be detected in ctDNA analysis, including *KCNA1, RET, APLP1, C8B, SYNE1, BCL7A,* and *FMNL2.*

Next, we inspected the Kaplan–Meier survival plot for the ten prognostic genes in 56 Taiwanese patients (Fig. S1). No significant difference in survival was observed between patients with and without these ten mutated genes. We then examined the survival rate for each prognostic gene. Among these ten genes, *BCL7A* was significantly correlated with the survival rate (P = 0.0204) (Fig. 4D). This is the first study in which *BCL7A* has been identified as a prognostic gene in CRC, and besides identifying its correlation with the survival rate. The mutation in *BCL7A* can be detected in ctDNA from blood samples (Fig. S1).

4. Discussion

This study performed driver gene identification for CRC by combining the DriverDB and IPA, analyzing 354 paired sequencing profiles from TCGA. A total of eight driver genes correlated to CRC were thus revealed, including *APC, BRAF, CDH8, FAT4, KRAS, NEFH, PIK3CA,* and *TP53.* In a recent study identifying CRC mutations, six genes were classified as CRC driver genes: *APC, TP53, PIK3CA, KRAS, BRAF,* and *SMAD4* [35]. The most frequently mutated of those genes, namely, *APC, TP53, PIK3CA, KRAS*, were consistent with our findings.

A series of mutations accumulate in critical driver genes over time in CRC development [36,37]. *APC* plays a principal role in CRC development and contributes to tumor progression. A defect of the APC protein induces beta-catenin accumulation and the acti-

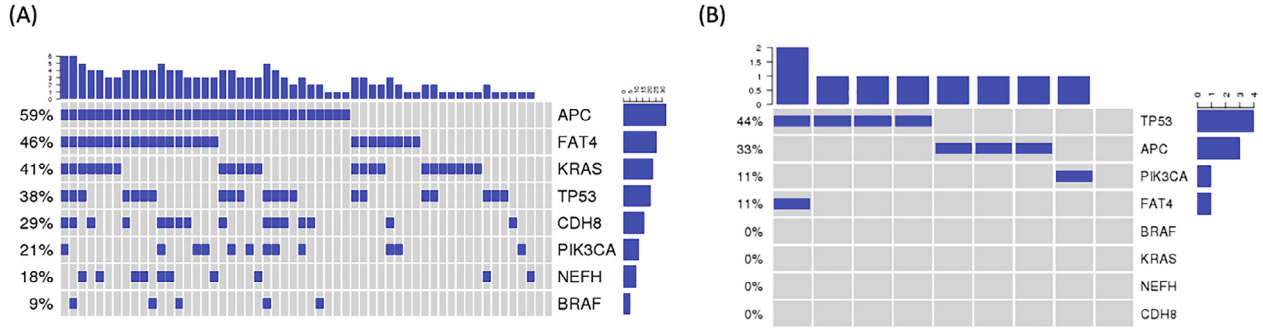

Fig. 3. Oncoprint representation of the eight mutated genes in Taiwanese patients. The coverage of the eight driver genes showed in (A) 56 patient tissue samples and (B) nine blood samples.

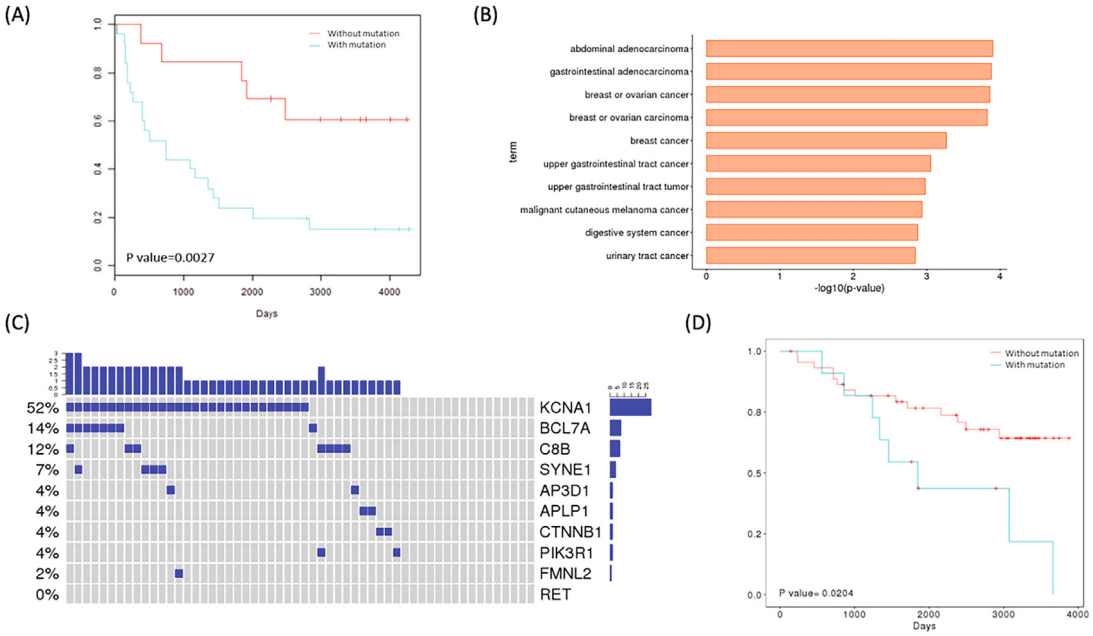

Fig. 4. The results of CRC prognostic gene identification. (A) The Kaplan–Meier survival plot for the ten prognostic genes. (B) Bar chart of functional enrichment analysis by IPA. (C) Oncoprint representation of the ten mutation genes in 56 patient tissue samples. The gene names are labeled on the right side, and the coverage rates are listed on the left side. (D) The Kaplan–Meier survival plot for the *BCL7A* gene.

vation of TCF-responsive genes in transcription [38]. *KRAS* and *BRAF* are involved in the Ras-Raf-MEK-ERK signaling pathway that affects cell growth and proliferation [39]. Mutated *TP53* leads to p53 function loss, which has been reported to be one of the major incidents in CRC development [40]. Moreover, mutations in the *PIK3CA* and *TP53* genes have been found to results in the emergence of a malignant tumor with invasion potential [41,42]. In this genetic development model of CRC, mutations of *APC*, *KRAS*, and *BRAF* are involved in transforming normal epithelial cells into adenoma cells, while mutated *PIK3CA* and *TP53* genes have been correlated with the late stage of CRC development, that is, the transition to carcinoma.

We found five genes with hub characteristics in our IPA results, namely, *APC, BRAF, KRAS, TP53,* and *PIK3CA*, a finding consistent with previous studies. The other three genes identified as prognostic genes in this study (*FAT4, CDH8,* and *NEFH*) have previously been reported to be correlated with other cancers besides CRC. In previous investigations, *NEFH* was found to serve as a tumor suppressor in hepatocellular carcinoma, esophageal squamous cell carcinoma, and breast cancer [10,43,44]; *FAT4* was found to be involved in the Wnt/PCP pathway [45] and to play a role in HIF1-α regulation [46]; and *CDH8* expression was found to be correlated with renal cell carcinoma and lung squamous cell carcinoma [47,48]. In this study, the *FAT4* gene coverage rates were 18.9% and 46.4% for the 354 profiles from the TCGA dataset and the 56 profiles from the Taiwanese dataset, respectively, while the *CDH8* coverage rates were 9.3% and 28.6%, respectively. The *NEFH* coverage rate for the 354 profiles from the TCGA dataset was 30.5% higher than the 17.9% coverage rate for the 56 profiles from the Taiwanese dataset. The differences between these two datasets could have been caused by racial differences and differences in the number of samples. In any case, the *FAT4, CDH8,* and *NEFH* genes could be potential targets for early-stage detection of CRC.

This study also attempted to identify prognostic genes from the 38 profiles with associated survival data obtained from the TCGA dataset. After classifying these 38 sequencing profiles into a poor prognosis group and a good prognosis group, we used DriverDB to identify driver genes by searching for genes that were expressed in the poor prognosis group but not the good prognosis group. Ten genes (*AP3D1, APLP1, BCL7A, C8B, CTNNB1, FMNL2, KCNA1, PIK3R1, RET,* and *SYNE1*) were identified, and the Kaplan–Meier analysis showed that patients with mutations in these ten genes mutated had significantly poorer overall survival (log-rank P = 0.0027).

Kaplan–Meier analysis was applied to each of the ten identified prognostic genes in 56 Taiwanese patients. The *KCNA1* gene, which encodes a voltage-gated delayed potassium channel, was mutated in 29 tissue samples from 56 Taiwanese patients and was detected in ctDNA derived from blood samples. A recent study revealed that the hypermethylation of the *KCNA1* promoter might serve as a potential novel diagnostic biomarker for early CRC patients [49]. Combined with our findings, this previous result suggests that *KCNA1* might play an important role in CRC, although the mutation of *KCNA1* was not associated with patient survival in the Taiwanese cohort. Patients with *BCL7A* mutations showed a worse survival rate (log-rank P = 0.0204). After Bonferroni correction, the conservative correction method which multiplies the raw P values by the number of tests, the posthoc adjusted p-value of BCL7A gene is 0.1836. Although the adjusted p-value is not small than 0.05, we believe that it's due to the small sample size of the Taiwanese cohort. The *BCL7A* gene encodes a subunit of the SWI/SNF complex and has previously been reported to serve as a prognostic gene for early-stage cutaneous T-cell lymphoma and ovarian cancer [50,51]. The result still provides insight of BCL7A as a prognostic gene. This study is the first to identify *BCL7A* as a prognostic gene in CRC and the first to report *BCL7A* ctDNA detection in plasma.

In recent years, the detection of circulating cell-free tumor DNA (ctDNA) has arisen as an attractive and minimally invasive method for detecting cancer-specific genetic differences in blood. This value of this approach has benefitted from the rapid development and application of NGS technology in cancer research and diagnostics. In our study, the highest coverage rate in the blood samples was for the *TP53* gene (44%), followed by that for the *APC* gene (33%). The ColoNext (https://www.ambrygen.com/providers/genetic-testing/6/oncology/colonext) is a 20-gene guideline-based panel identifying patients with inherited risk for CRC. The only two genes included among both the 20 ColoNext genes and our

eight driver genes were *APC* and *TP53*, which showed that mutations of these two genes could be detected through blood screening. Moreover, the validated prognostic gene, *BCL7A*, could also be detected in the ctDNA samples. Owing to the rapid development of NGS technology, genetic tests will become cheaper and faster, and more and more patients will benefit from this technology. Relatedly, developing precise gene panels with reasonable prices is an urgent need for the protection and improvement of human health. For the accurate early-stage detection of cancer, there is still a large amount of research that must be performed, and our findings could provide guidance for future research efforts.

5. Conclusion

The results of the current study showed that eight driver gene mutations were identified in CRC. Of those genes, *APC, TP53, PIK3CA*, and *FAT4* could be detected in blood as ctDNA analysis targets. Moreover, this study discovered 10 prognostic gene mutations in CRC. Among these genes, *BCL7A* gene mutations are correlated with prognosis in CRC, and that these mutations could be found through the ctDNA sequencing. However, while our results and those of many previous studies have indicated several potential targets for ctDNA analysis, further large-scale research studies and further technological development are required. Furthermore, the plasma and ctDNA separation protocols need to be standardized for comparison between different studies in order to determine the clinical situations in which ctDNA analysis is superior to other methods and to further optimize these protocols for application in clinical early cancer detection or disease monitoring. Finally, the findings of the present study could provide valuable guidance for future studies involving disease screening, drug development, and treatment suggestions.

6. Author statement

W-C.C. and Y-F.L. conceived and designed the experiments. JY-C.C, H-C.C., J-K.J., and C-Y.Y performed the experiments. W-C.C., F-H.C., Y-L.L, W-J.L, and C-Y.L analyzed the data. JY-C.C, M-H.Y and S-C.W. contributed reagents, materials, and/or analysis tools. W-C.C., Y-F.L. and JY-C.C wrote the paper. All authors reviewed the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by Ministry of Science and Technology [MOST 108-2622-E-039-005-CC2; MOST 109-2622-E-039-004-CC2; MOST 109-2314-B-182-078-MY3; MOST 109-2327-B-039-002; MOST 109-2628-B-182-008; MOST 109-2628-E-039-001-MY3]; China Medical University [CMU 108-S-22; CMU 108-Z-02; CMU 109-MF-61]; China Medical University Hospital [DMR-109-055; DMR-109-223; DMR-110-072; DMR-110-244]; Chang Gung Memorial Hospital at Linkou [CMRPD1J0321; CMRPD1H0472].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.06.044.

References

[1] Miller KD, Nogueira L, Mariotto AB, Rowland JH, Yabroff KR, Alfano CM, et al. Cancer treatment and survivorship statistics, 2019. CA Cancer J Clin 2019;69(5):363–85.
[2] Society AC: Colorectal Cancer Facts & Figures 2020-202In. Atlanta: American Cancer Society.
[3] Brenner H, Hoffmeister M, Birkner B, Stock C. Diagnostic performance of guaiac-based fecal occult blood test in routine screening: state-wide analysis from Bavaria, Germany. Am J Gastroenterol 2014;109(3):427–35.
[4] Lee JK, Liles EG, Bent S, Levin TR, Corley DA. Accuracy of fecal immunochemical tests for colorectal cancer: systematic review and meta-analysis. Ann Intern Med 2014;160(3):171-181. https://doi.org/10.7326/M13-1484.
[5] van der Schouw YT, Verbeek ALM, Wobbes Th, Segers MFG, Thomas CMG. Comparison of four serum tumour markers in the diagnosis of colorectal carcinoma. Br J Cancer 1992;66(1):148–54.
[6] Bressler B, Paszat LF, Chen Z, Rothwell DM, Vinden C, Rabeneck L. Rates of new or missed colorectal cancers after colonoscopy and their risk factors: a population-based analysis. Gastroenterology 2007;132(1):96–102.
[7] Strano S, Muti P, Blandino G. What biomarkers (if any) for precise medicine?. Aging (Albany NY) 2015;7(8):533–4.
[8] Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009;458(7239):719–24.
[9] Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz Jr LA, Kinzler KW. Cancer genome landscapes. Science 2013;339(6127):1546–58.
[10] Bozic I, Antal T, Ohtsuki H, Carter H, Kim D, Chen S, et al. Accumulation of driver and passenger mutations during tumor progression. PNAS 2010;107(43):18545–50.
[11] Cragun D, Radford C, Dolinsky JS, Caldwell M, Chao E, Pal T. Panel-based testing for inherited colorectal cancer: a descriptive study of clinical testing performed by a US laboratory. Clin Genet 2014;86(6):510–20.
[12] Pages F, Mlecnik B, Marliot F, Bindea G, Ou FS, Bifulco C, et al. International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. Lancet 2018;391(10135):2128–39.
[13] A stool DNA test (Cologuard) for colorectal cancer screening. Med Lett Drugs Ther 2014, 56(1453):100-101.
[14] Johnson DA, Barclay RL, Mergener K, Weiss G, König T, Beck J, et al. Plasma Septin9 versus fecal immunochemical testing for colorectal cancer screening: a prospective multicenter study. PLoS ONE 2014;9(6):e98238.
[15] Kim S-Y, Kim YS. A gene sets approach for identifying prognostic gene signatures for outcome prediction. BMC Genomics 2008;9(1):177. https://doi.org/10.1186/1471-2164-9-177.
[16] Willis S, Villalobos VM, Gevaert O, Abramovitz M, Williams C, Sikic BI, et al. Single gene prognostic biomarkers in ovarian cancer: a meta-analysis. PLoS ONE 2016;11(2):e0149183.
[17] Zhu Y, Wu J, Zhang C, Sun S, Zhang J, Liu W, et al. BRCA mutations and survival in breast cancer: an updated systematic review and meta-analysis. Oncotarget 2016;7(43):70113–27.
[18] Cheng W-C, Chung I-F, Chen C-Y, Sun H-J, Fen J-J, Tang W-C, et al. DriverDB: an exome sequencing database for cancer driver gene identification. Nucl Acids Res 2014;42(D1):D1048–54.
[19] Wilks C, Cline MS, Weiler E, Diehkans M, Craft B, Martin C, Murphy D, Pierce H, Black J, Nelson D et al: The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. Database (Oxford) 2014, 2014.
[20] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucl Acids Res 2001;29(1):308–11.
[21] Trager RS: DNA sequencing. Venter's next goal: 1000 human genomes. Science 2002, 298(5595):947.
[22] Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 2009;106(23):9362–7.
[23] Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer 2004;91(2):355–8.
[24] Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 2012;337(6090):64–9.
[25] Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucl Acids Res 2014;42(D1):D980–5.
[26] Schorderet DF. Using OMIM (On-line Mendelian Inheritance in Man) as an expert system in medical genetics. Am J Med Genet 1991;39(3):278–84.
[27] Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 2012;6(2):80–92.
[28] McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 2010;26(16):2069–70.
[29] Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499(7457):214–8.
[30] Vandin F, Upfal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. Genome Res 2012;22(2):375–85.

[31] Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol Syst Biol 2013;9(1):637. https://doi.org/10.1038/msb:2012.68.
[32] Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 2010;26(5):589–95.
[33] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. Genome Project Data Processing S: The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) 2009;25(16):2078–9.
[34] Kodama Y, Shumway M, Leinonen R. International nucleotide sequence database C: the sequence read archive: explosive growth of sequencing data. Nucl Acids Res 2012;40(D1):D54–6.
[35] Schell MJ, Yang M, Teer JK, Lo FY, Madan A, Coppola D, et al. A multigene mutation classification of 468 colorectal cancers reveals a prognostic role for APC. Nat Commun 2016;7(1). https://doi.org/10.1038/ncomms11743.
[36] Nowell P. The clonal evolution of tumor cell populations. Science 1976;194 (4260):23–8.
[37] Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell 1990;61(5):759–67.
[38] Dow L, O'Rourke K, Simon J, Tschaharganeh D, van Es J, Clevers H, et al. Apc restoration promotes cellular differentiation and reestablishes crypt homeostasis in colorectal cancer. Cell 2015;161(7):1539–52.
[39] Eklöf V, Wikberg ML, Edin S, Dahlin AM, Jonsson B-A, Öberg Å, et al. The prognostic role of KRAS, BRAF, PIK3CA and PTEN in colorectal cancer. Br J Cancer 2013;108(10):2153–63.
[40] Rechsteiner M, von Teichman A, Rüschoff JH, Fankhauser N, Pestalozzi B, Schraml P, et al. KRAS, BRAF, and TP53 deep sequencing for colorectal carcinoma patient diagnostics. J Mol Diagn 2013;15(3):299–311.
[41] Samuels Y, Velculescu VE. Oncogenic mutations of PIK3CA in human cancers. Cell Cycle 2004;3(10):1221–4.
[42] Baker SJ, Fearon ER, Nigro JM, Hamilton SR, Preisinger AC, Jessup JM, et al. Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. Science 1989;244(4901):217–21.
[43] Revill K, Wang T, Lachenmayer A, Kojima K, Harrington A, Li J, et al. Genome-wide methylation analysis and epigenetic unmasking identify tumor suppressor genes in hepatocellular carcinoma. Gastroenterology 2013;145 (6):1424–1435.e25.
[44] Jeschke J, Van Neste L, Glöckner SC, Dhir M, Calmon MF, Deregowski V, et al. Biomarkers for detection and prognosis of breast cancer identified by a functional hypermethylome screen. Epigenetics 2012;7(7):701–9.
[45] Saburi S, Hester I, Fischer E, Pontoglio M, Eremina V, Gessler M, et al. Loss of Fat4 disrupts PCP signaling and oriented cell division and leads to cystic kidney disease. Nat Genet 2008;40(8):1010–5.
[46] Wang H-W, Huang T-S, Lo H-H, Huang P-H, Lin C-C, Chang S-J, et al. Deficiency of the microRNA-31-microRNA-720 pathway in the plasma and endothelial progenitor cells from patients with coronary artery disease. Arterioscler Thromb Vasc Biol 2014;34(4):857–69.
[47] Blaschke S, Mueller CA, Markovic-Lipkovski J, Puch S, Miosge N, Becker V, et al. Expression of cadherin-8 in renal cell carcinoma and fetal kidney. Int J Cancer 2002;101(4):327–34.
[48] Choi M, Kadara H, Zhang J, Parra ER, Rodriguez-Canales J, Gaffney SG, et al. Mutation profiles in early-stage lung squamous cell carcinoma with clinical follow-up and correlation with markers of immune function. Ann Oncol 2017;28(1):83–9.
[49] Uhan S, Zidar N, Tomažič A, Hauptman N. Hypermethylated promoters of genes UNC5D and KCNA1 as potential novel diagnostic biomarkers in colorectal cancer. Epigenomics 2020;12(19):1677–88.
[50] Litvinov IV, Zhou Y, Kupper TS, Sasseville D. Loss of BCL7A expression correlates with poor disease prognosis in patients with early-stage cutaneous T-cell lymphoma. Leuk Lymphoma 2013;54(3):653–4.
[51] Sun Z, Sun L, He M, Pang Y, Yang Z, Wang J. Low BCL7A expression predicts poor prognosis in ovarian cancer. J Ovarian Res 2019;12(1):41.