Cancer Medicine

Open Access

ORIGINAL RESEARCH

Analysis of potential genes and pathways associated with the colorectal normal mucosa–adenoma–carcinoma sequence

Zhuoxuan Wu[1], Zhen Liu[1], Weiting Ge[2], Jiawei Shou[1], Liangkun You[1], Hongming Pan[1] & Weidong Han[1]

[1]Department of Medical Oncology, Sir Run Run Shaw Hospital, College of Medicine, Zhejiang University, Hangzhou, Zhejiang, China
[2]Cancer Institute, The Second Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou, Zhejiang, China

Keywords
Colorectal normal mucosa–adenoma–carcinoma sequence, differentially expressed genes, functional analysis, microarray analysis, prognosis

Correspondence
Weidong Han and Hongming Pan, Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University, 3# East Qingchun Road, Hangzhou, Zhejiang 310016, China.
Tel: +86-571-86006926 (WH);
+86-571-86006922 (HP);
Fax: +86-571-86436673 (WH);
+86-571-86436673 (HP);
E-mails: hanwd@zju.edu.cn;
panhongming@zju.edu.cn

Funding Information
This work was supported by the National Natural Science Foundation of China (81572592, 81772543), the Zhejiang Province Preeminence Youth Fund (LR16H160001), the Zhejiang Natural Sciences Foundation Grant (LZ15H160001, LY17H160029, and Q17H160042), the National Health and Family Planning Commission Fund (2015112271), and the Zhejiang medical innovative discipline construction project-2016.

Received: 27 December 2017; Revised: 10 March 2018; Accepted: 15 March 2018

Cancer Medicine **2018; 7(6):2555–2566**

doi: 10.1002/cam4.1484

Abstract

This study aimed to identify differentially expressed genes (DEGs) related to the colorectal normal mucosa–adenoma–carcinoma sequence using bioinformatics analysis. Raw data files were downloaded from Gene Expression Omnibus (GEO) and underwent quality assessment and preprocessing. DEGs were analyzed by the limma package in R software (R version 3.3.2). Gene Ontology (GO) enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis were performed with the DAVID online tool. In a comparison of colorectal adenoma (n = 20) and colorectal cancer (CRC) stage I (n = 31), II (n = 38), III (n = 45), and IV (n = 62) with normal colorectal mucosa (n = 19), we identified 336 common DEGs. Among them, seven DEGs were associated with patient prognosis. Five (*HEPACAM2*, *ITLN1*, *LGALS2*, *MUC12*, and *NXPE1*) of the seven genes presented a sequentially descending trend in expression with tumor progression. In contrast, *TIMP1* showed a sequentially ascending trend. *GCG* was constantly downregulated compared with the gene expression level in normal mucosa. The significantly enriched GO terms included extracellular region, extracellular space, protein binding, and carbohydrate binding. The KEGG categories included HIF-1 signaling pathway, insulin secretion, and glucagon signaling pathway. We discovered seven DEGs in the normal colorectal mucosa–adenoma–carcinoma sequence that was associated with CRC patient prognosis. Monitoring changes in these gene expression levels may be a strategy to assess disease progression, evaluate treatment efficacy, and predict prognosis.

Introduction

Colorectal cancer (CRC) is the third leading cause of cancer and cancer-related death in patients with cancer worldwide, accounting for more than 134,000 estimated new cases and 49,000 estimated deaths in 2016 [1]. The five-year survival rate of CRC is approximately 65% in high-income countries but is <50% in low-income countries [2–4]. From 2000 to 2014, the mortality of CRC decreased by 18% in individuals aged ≥50 years due to the extensive use of traditional screening methods, including flexible

colonoscopy, barium enema X-ray, and fecal blood testing [5]. However, these tests have non-negligible shortcomings, including bleeding, perforation, and acute diverticulitis, as well as a variable sensitivity ranging from 37% to 80% [6, 7]. Therefore, sensitive and specific biomarkers are urgently needed to improve the rate of early diagnosis, to help manage individual therapy and to predict the prognosis of patients in different stages of the disease.

Most CRC cases develop slowly through the normal mucosa–adenoma–carcinoma sequence over several years [8]. During this multistep process of colorectal tumorigenesis, many factors play important roles, including old age, smoking, alcohol, a high-fat diet, and lack of physical exercise [9]. In recent decades, multiple genes and signaling pathways have been shown to participate in the initiation and development of CRC. Kinzler and Fearon et al. reported that *APC* inactivation was an early event in more than 70% of colorectal adenomas and carcinomas [10, 11]. *KRAS* and *TP53* mutations participated in the adenoma–carcinoma sequence [11]. Liu et al. [12] found that low miR-126 and high CXCR4 protein expression were associated with poor prognosis in colorectal cancer. Tsukamoto et al. [13] reported that overexpression of osteoprotegerin in human colorectal cancer might be a predictive biomarker of CRC recurrence and a potential target for individual treatment of this disease. Dynamic changes of genes in different stages have important roles in the occurrence and development of CRC, as well as the treatment and prognosis of this disease [14–17]. These differentially expressed genes (DEGs) may show changes that correspond to their functions in the different stages of CRC, which lead to different survival outcomes [18]. Stage at diagnosis is an important prognostic factor for patients with CRC. Siegel et al. [2] found that the five-year survival rate of patients diagnosed with CRC ranges from 90.1% in stage I to 11.7% in stage IV. Thus, it is important to identify DEGs during the normal mucosa–adenoma–carcinoma sequence, which will help elucidate the molecular mechanisms involved in the occurrence and development of CRC, provide potential biomarkers for diagnosis at the early stage, and suggest potential targets for individual therapy.

Bioinformatics is a newly emerging scientific field that combines biology, mathematics, and information technology, making it possible to analyze large and increasingly complex molecular datasets. Microarray assays can acquire expression information on thousands of genes simultaneously and explore the genomic alterations associated with the progress of colorectal initiation and development [19]. Extensive genetic information is available online due to the development of public cancer databases, such as The Cancer Genome Atlas (TCGA), Oncomine, Gene Expression Omnibus (GEO), and others, which are repositories for microarray data retrieval and deposit. Online datasets can help enlarge the sample size and increase the statistical power. For example, Fu et al. [20] identified 72 miRNA–mRNA pairs along with 22 dysregulated miRNAs and their 58 target mRNAs that were involved in CRC tumorigenesis by a combination of expression profiling and bioinformatics analysis. Robles et al. [21] found that the CRC that developed in patients with IBD had different genetic characteristics from sporadic CRC with whole-exome sequencing analysis, providing possible genetic biomarkers for diagnosis and treatment of patients with IBD and CRC.

In our study, we aimed to identify DEGs related to the colorectal normal mucosa–adenoma–carcinoma sequence. Original data were downloaded from GEO and analyzed with R software (R version 3.3.2). Gene expression levels in colorectal adenoma and CRC stage I, II, III, and IV were compared with those in normal colorectal mucosa. We eventually identified seven potential DEGs related to CRC patient prognosis and explored their function by performing Gene Ontology (GO) analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis.

Materials and Methods

Screening microarray data

The GEO database was systematically searched without language, race, region, and time restrictions (up to 7 January 2017). The advanced search strategy insured the comprehensiveness of the search results (see Table S1 for search strategy details). These inclusion criteria were as follows: (1) total RNA was extracted from frozen colorectal tissue sections; (2) datasets had the original gene expression data files. RNA extracted from frozen tissues shows little degradation. This was the basis for the qualified microarray. In addition, the original gene expression data files could realistically indicate the microarray quality. Evaluating chip quality and rejecting unqualified chip data insured the accuracy of the subsequent analysis. The exclusion criteria were as follows: (1) the genome-wide gene expression profile was not generated by Affymetrix Human Genome U133 Plus 2.0 Array; (2) No disease staging information was present; (3) Frozen tissue sections came from patients who might have received antitumor treatments previously. At present, there is no readymade dataset containing normal colorectal mucosa, adenoma, and carcinoma with all four stages in one dataset. However, there were several datasets containing either colorectal adenoma or carcinoma at different stages. The integration of these different datasets combined the normal colorectal mucosa, adenoma, and carcinoma at all four stages together,

making it possible to analyze genetic changes in the progression of colorectal cancer. Different datasets had heterogeneity, but the heterogeneity was much smaller if the datasets were generated from one platform. The Affymetrix Human Genome U133 Plus 2.0 Array was selected because this platform generated the most available datasets for further analysis. At the same time, it was necessary to exclude the influence of the therapeutic factors on the gene expression level. Thus, patients who previously received antitumor treatments were excluded.

Evaluation of microarray quality

The selected gene expression data were downloaded from the GEO database. These raw data files were from the Affymetrix platform, which could be analyzed by the affy package [22] in R software (R version 3.3.2). Before data preprocessing, all the microarrays were evaluated for quality by quality control (QC), relative logarithmic expression (RLE), normalized unscaled standard errors (NUSE), and RNA degradation curve [23, 24]. QC is an overall assessment of the microarray quality, which primarily consists of the present percentage, background noise, scale factor, GAPDH 3'/5' ratio, and actin 3'/5' ratio. RLE and NUSE can both evaluate the consistency of the data, while NUSE is more sensitive. The RNA degradation curve plays an important role in evaluating the degradation of the microarray. Through quality assessment, poor quality data were removed.

Processing of microarray data

Data preprocessing was performed with a standard robust multiarray average (RMA) method, including background correction, normalization, and logarithmic conversion [25]. The raw data were converted to probe-level data after the RMA algorithm and were then transformed to gene symbols in R software [26]. The gene expression levels were the mean of the probes when multiple probes corresponded to one gene symbol. The batch effect could be due to different experimental times, methods, experimenters, datasets, platforms, and many other unpredictable factors, which might affect the accuracy of the data analysis. However, the datasets generated from one platform have a much smaller batch effect. In addition, the batch effect was evaluated by the expression level of housekeeping genes in each dataset to judge whether batch effects have a significant impact on our conclusions. The DEGs were identified by the limma package in R software [27]. Only genes with $|\log_2 FC| > 1$ (FC: fold change) and an adjusted *P*-value <0.05 were considered DEGs. Then, all DEGs underwent prognostic analysis with the survival information from TCGA. TCGA database had no separate colorectal adenoma data, and thus, this information was only used to validate the DEGs in colorectal cancer with four stages with RNA-seq data.

Functional and pathway enrichment analysis

GO annotates and classifies genes based on three categories, including biology process, molecular function, and cellular component [28, 29]. KEGG pathway interprets pathway maps of molecular interactions, reactions, and relation networks [30]. In our analysis, the DAVID online tool was used to perform GO enrichment and KEGG pathway analysis of the identified DEGs and many other related background DEGs with threshold *P*-values <0.05.

Results

Basic characteristics of the microarray data

A total of 592 search results were obtained by our search strategy (See Table S1 for search strategy details). Thirty-eight datasets met the two inclusion criteria. Their RNAs were all extracted from frozen colorectal tissue sections. In addition, these 38 datasets had the original gene expression data files. Based on the exclusion criteria, 19 datasets were excluded because the Affymetrix Human Genome U133 Plus 2.0 Array was not used. One dataset was excluded because it had no staging information. Thirteen datasets with patients who might have received antitumor treatments shortly before were also excluded. Thus, only five datasets (GSE4183, GSE14333, GSE39582, GSE8671, and GSE10714) were finally eligible for analysis (Fig. 1, Table S2).

These five datasets had a total of 971 raw data files, which were from the Affymetrix platform. According to the staging information, data were divided into six groups: normal mucosa, adenoma, and CRC stage I, II, III, and IV. Quality assessment was performed for all these raw data files by QC, RLE, NUSE, and the RNA degradation curve (Fig. S1). At the same time, the consistency of the data volume of each group and the microarray quality were taken into account, and 215 data profiles were finally included in the analysis (Table 1). In addition, the batch effect was evaluated with the expression level of GAPDH across the different datasets, and heterogeneity was not significant (Fig. S2).

Identification of DEGs and prognosis analysis

The gene expression levels in the colorectal adenoma and CRC stage I, II, III, and IV were compared to

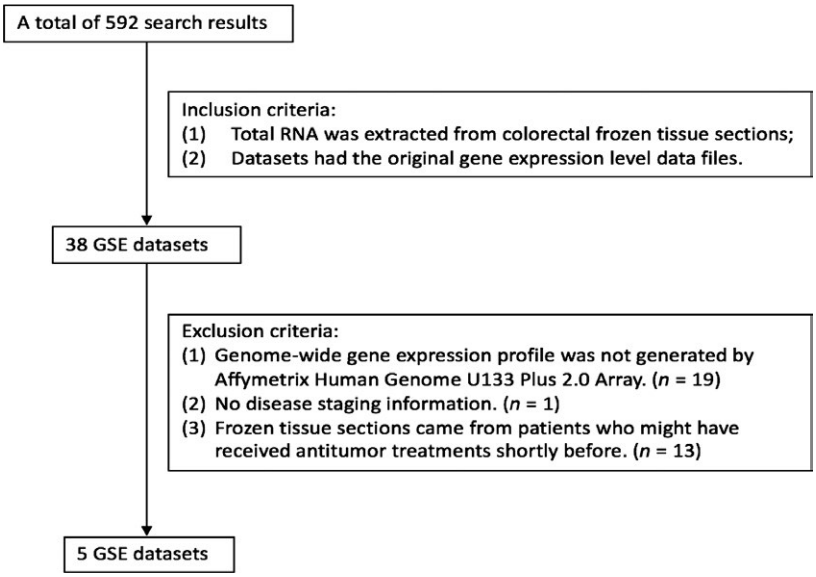

Figure 1. Flowchart of selecting eligible datasets.

those in normal colorectal mucosa. The five comparison groups were normal mucosa–adenoma, normal mucosa–CRC stage I, normal mucosa–CRC stage II, normal mucosa–CRC stage III, and normal mucosa–CRC stage IV. With a threshold of $|\log_2 FC| > 1$ and an adjusted *P*-value <0.05, 645 DEGs were identified between the normal mucosa and the adenoma. In addition, there were 1059, 1183, 1195, and 1100 DEGs corresponding to normal mucosa–CRC stage I, normal mucosa–CRC stage II, normal mucosa–CRC stage III, and normal mucosa–CRC stage IV, respectively. Then, 336 *common* DEGs were extracted from these five comparison groups (Fig. 2A). Among these 336 *common* DEGs, 87 genes were identified with an ascending or descending trend through the normal mucosa–adenoma–carcinoma sequence (Table S3). Using the survival information of TCGA, we eventually selected six DEGs related to patient prognosis. They were *HEPACAM2, ITLN1, LGALS2, MUC12, NXPE1*, and *TIMP1* (Fig. 3A–F, Table 2). Five of the six genes presented a descending trend in expression with tumor progression, while one gene presented an ascending trend.

We also obtained another six common DEGs from these five comparison groups with a threshold of $|\log_2 FC| > 4$ and an adjusted *P*-value <0.05. They were *AQP8, CLCA4, CLDN8, GCG, GUCA2A*, and *MS4A12* (Fig. 2B, Table 3). These genes were all downregulated compared with the gene expression levels in normal mucosa. Among them, only *GCG* was considered related to patient prognosis (Fig. 3G, Table 3).

A total of seven DEGs from the GEO database associated with prognosis were obtained by bioinformatics analysis. They were *HEPACAM2, ITLN1, LGALS2, MUC12, NXPE1, TIMP1*, and *GCG*. The expression patterns of these seven DEGs were also confirmed by 672 RNA-seq data from TCGA. Their expression levels all presented the same trend as that in the GEO database except *MUC12* (Table 4). However, *MUC12* was also downregulated in colorectal cancer tissues compared with normal colorectal mucosa. The expression of these seven DEGs in colorectal adenomas could not be verified because TCGA database had no separate adenoma data.

Table 1. GSE datasets included in our study.

Sample stage	Quantity	GSE datasets
Colorectal normal mucosa	19	GSE4183 + GSE8671
Adenoma	20	GSE4183 + GSE8671
CRC stage 1	31	GSE14333 + GSE39582
CRC stage 2	38	GSE14333
CRC stage 3	45	GSE14333
CRC stage 4	62	GSE14333 + GSE39582

CRC, colorectal cancer; GEO, Gene Expression Omnibus; GSE, GEO series.

GO term enrichment analysis

Using a threshold of $|\log_2 FC| > 1$ and an adjusted *P*-value <0.05, we identified 645, 1059, 1183, 1195, and 1100 DEGs in the comparisons of normal mucosa–colorectal adenoma, normal mucosa–CRC stage I, normal mucosa–CRC stage II, normal mucosa–CRC stage III,

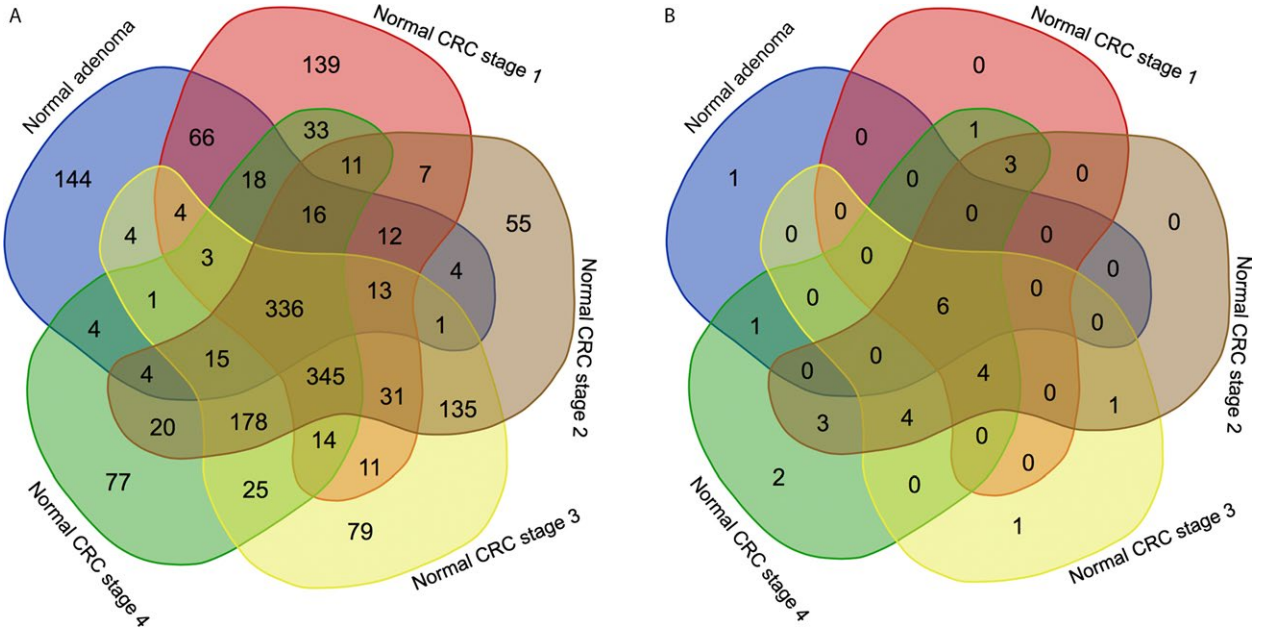

Figure 2. Venn diagram for (A) the 336 common DEGs with threshold of $|\log_2 FC| > 1$ and an adjusted *P*-value <0.05 and for (B) the six common DEGs with threshold of $|\log_2 FC| > 4$ and an adjusted *P*-value <0.05 extracted from normal mucosa–adenoma, normal mucosa–CRC stage I, normal mucosa–CRC stage II, normal mucosa–CRC stage III, and normal mucosa–CRC stage IV. FC, fold change.

and normal mucosa–CRC stage IV, respectively. Combining these genes together, we obtained a total of 1805 background DEGs in a union including the seven identified DEGs.

These 1805 DEGs were uploaded to the DAVID online tool to perform GO analysis and KEGG pathway analysis to explore the possible biological functions and signaling pathways of the DEGS. The results of the seven identified DEGs were extracted separately. In the biology process GO category, the functional enrichment of the seven DEGs was scattered so that no common biology process was found among these seven DEGs. In the cellular component GO category, *TIMP1*, *GCG*, and *NXPE1* were related to extracellular region, and *TIMP1* and *GCG* were related to extracellular space. In the molecular function GO category, *HEPACAM2*, *LGALS2*, *TIMP1*, and *GCG* were associated with protein binding, and *ITLN1* and *LGALS2* were associated with carbohydrate binding. Additionally, these DEGs had some other specific classifications (Table 5).

KEGG pathway analysis

The KEGG pathway enrichment analysis indicated that *TIMP1* might participate in the HIF-1 signaling pathway, and *GCG* might play a role in the insulin secretion and glucagon signaling pathway (Table 6).

Discussion

Sequential changes of gene expression in different stages play essential roles in the colorectal normal mucosa–adenoma–carcinoma sequence [31]. Many studies have confirmed that DEGs participated in the progress of CRC. Heijink et al. [32] reported that caspase-8 and cellular flice-like inhibitory protein (cFLIP) expression induced colorectal carcinogenesis independently in sporadic and hereditary nonpolyposis colorectal cancer (HNPCC)-associated adenomas and carcinomas. Galamb et al. [24] showed that downregulated amnionless homolog (AMN) and prostaglandin-D2 receptor (PTGDR) and upregulated osteopontin and osteonectin were potential biomarkers of colorectal carcinogenesis and disease progression. However, most studies only focused on the colorectal carcinogenesis process, from normal colorectal mucosa to CRC [33, 34]. Many other studies have examined the macro classification, such as from normal colorectal mucosa to adenoma and then to CRC with all stages mixed [35, 36]. Comparatively speaking, our study had a more detailed grouping because there were five comparison groups in total. The gene expression levels in the colorectal adenoma and CRC stage I, II, III, and IV were compared with those of the normal colorectal mucosa. This has not been performed previously in the current literature. By analyzing public gene data from GEO and TCGA in R software

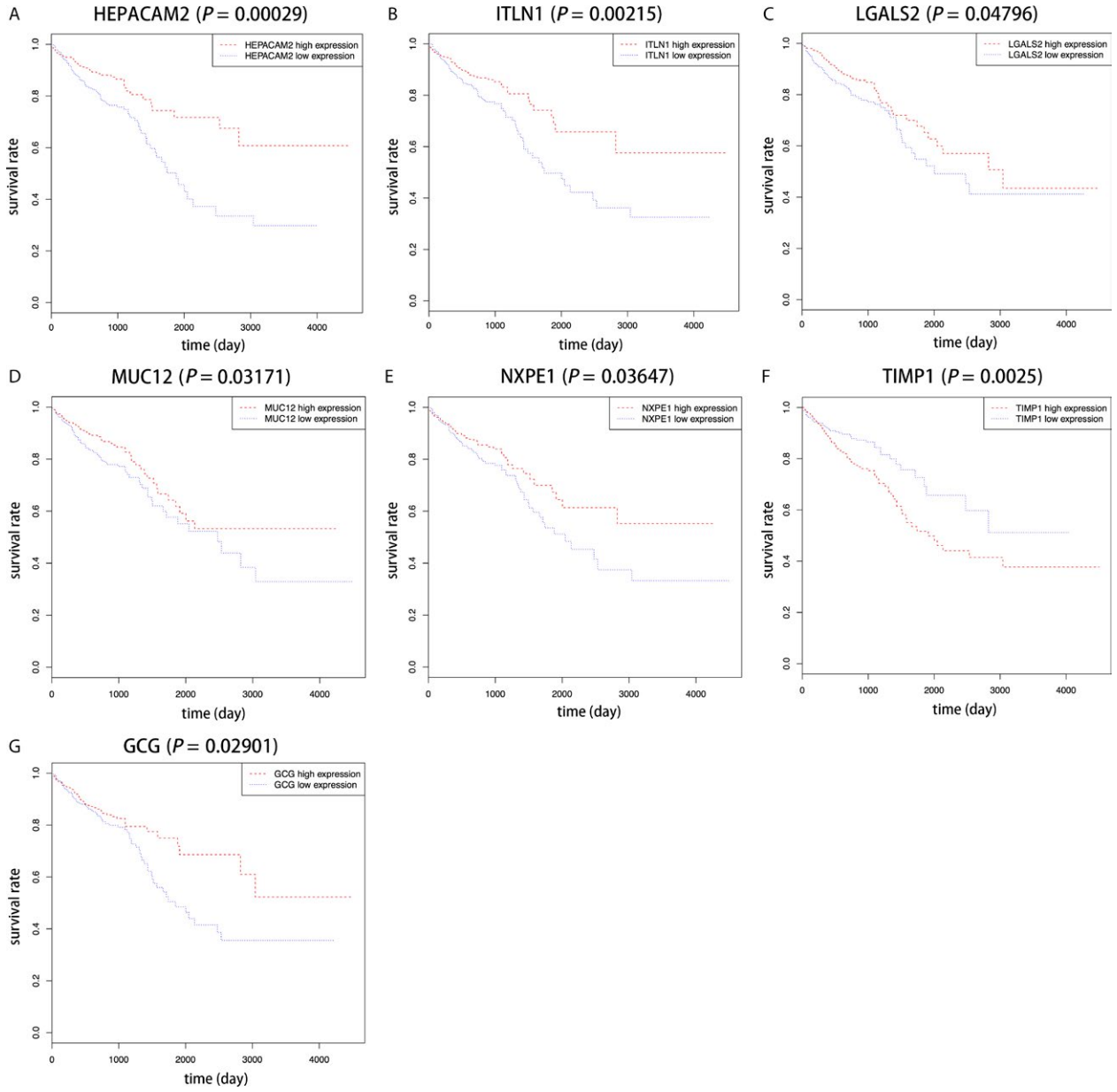

Figure 3. Survival curve of (A) HEPACAM2, (B) ITLN1, (C) LGALS2, (D) MUC12, (E) NXPE1, (F) TIMP1, and (G) GCG.

Table 2. Six genes presenting sequentially expression level changes through normal colorectal mucosa–adenoma–carcinoma sequence.

Genes	Normal adenoma |FC|	Normal stage 1 |FC|	Normal stage 2 |FC|	Normal stage 3 |FC|	Normal stage 4 |FC|	Five-year survival rate (*P*-value)
HEPACAM2	−1.21761	−3.85613	−4.0706	−4.06757	−4.57233	0.00029
ITLN1	−2.18651	−4.00026	−4.54605	−4.00429	−4.82109	0.00215
LGALS2	−1.86415	−2.7824	−3.50298	−3.01946	−3.08688	0.04796
MUC12	−1.08214	−2.12846	−2.39063	−2.5668	−2.54186	0.03171
NXPE1	−1.00081	−2.57598	−2.93155	−2.77359	−3.2236	0.03647
TIMP1	1.561751	1.690323	2.036284	2.050785	2.217349	0.0025

FC, fold change.

Table 3. Six genes maintained continuous downregulated compared with gene expression level in normal colorectal mucosa.

Genes	Normal adenoma |FC|	Normal stage 1 |FC|	Normal stage 2 |FC|	Normal stage 3 |FC|	Normal stage 4 |FC|	Five-year survival rate (*P*-value)
AQP8	−5.08128	−5.59115	−5.85151	−5.60485	−6.51036	0.13895
CLCA4	−4.0168	−5.75406	−5.75458	−5.46981	−6.55938	0.16333
CLDN8	−4.71209	−5.19946	−5.58836	−5.06006	−5.65663	0.9914
GCG	−5.4652	−5.40381	−5.95615	−5.65407	−5.70174	0.02901
GUCA2A	−4.11537	−4.40829	−4.32434	−4.24492	−4.73345	0.20505
MS4A12	−4.48105	−5.56404	−5.52912	−5.06747	−6.16997	0.74113

FC, fold change.

(R version 3.3.2), we eventually identified seven DEGs potentially related to CRC patient prognosis and explored their function by performing GO analysis and KEGG pathway analysis. These findings may help elucidate the molecular mechanisms involved in the initiation and development of CRC, provide potential biomarkers of early diagnosis, help manage potential targets in individual therapy, and predict the prognosis of patients at different stages of the disease.

In our study, we downloaded GSE4183, GSE14333, GSE39582, GSE8671, and GSE10714 from the public database GEO, which is an international public repository for microarray data retrieval and deposit. These data were submitted by the investigators, and there was a lack of quality review and evaluation [37]. Thus, evaluating the quality of microarray assays is important. QC is an overall assessment of the microarray quality, which primarily consists of present percentage, background noise, scale factor, GAPDH 3'/5' ratio, and actin 3'/5' ratio [24]. RLE and NUSE can both evaluate the consistency of the data, while NUSE is more sensitive. The RNA degradation curve plays an important role in evaluating the degradation of the microarray [24]. For RLE, the box chart center of each data profile in the high-quality dataset should be close to the position of the ordinate 0. For NUSE, it should be close to 1. If the slope of the RNA degradation curve is close to 0, it indicates that the degradation of the microarray is serious, and these data should be removed. The raw data must go through these quality evaluations, and only qualified data can be entered into the next data processing step to insure the reliability of the subsequent analysis [38].

We obtained seven DEGs of interest, which were *HEPACAM2, ITLN1, LGALS2, MUC12, NXPE1, TIMP1*, and *GCG*, and all were associated with patient prognosis. *HEPACAM2, ITLN1, LGALS2, MUC12*, and *NXPE1* presented a sequentially descending trend in expression with tumor progression. In contrast, *TIMP1* presented a sequentially ascending trend. Furthermore, *GCG* showed constant downregulation compared with the gene expression level in normal mucosa. Among these seven DEGs, *TIMP1* and *GCG* have been studied extensively. *TIMP1* is a member of the tissue inhibitors of metalloproteinase (*TIMP*) family that regulates matrix metalloproteinases (MMPs) and disintegrin metalloproteinases [39]. Recent studies reported that the dysregulated activity of *TIMP1* was associated with cancer progression [40]. Increased expression of *TIMP1* was shown to predict worse prognosis of laryngeal carcinoma [41] and melanoma [42]. Many studies have reported the *TIMP1* was upregulated in both early and advanced CRC [43, 44], and it possibly acted as a prognostic biomarker involved in liver metastasis of CRC [45, 46]. In our study, we found that *TIMP1* presented a sequentially ascending trend through the normal colorectal mucosa–adenoma–carcinoma sequence, and the upregulation of *TIMP1* indicated a poor survival prognosis, consistent with previous studies. *GCG* is involved in the regulation of incretin synthesis, secretion, inactivation, and RET signaling. Diseases related to *GCG* are diabetes [47] and other metabolic diseases [48]. Drucker [49] reported that the protein encoded by *GCG* was a preproprotein, which could be cleaved into four mature peptides and regulated cell proliferation, differentiation, and apoptosis. However, few studies have focused on the role of *GCG* in CRC progression. We discovered that *GCG* expression was downregulated in both adenomas and carcinomas, which was also confirmed by Spisak et al. [50].

Table 4. The expression pattern of seven DEGs on the TCGA.

Genes	Normal stage 1 |FC|	Normal stage 2 |FC|	Normal stage 3 |FC|	Normal stage 4 |FC|
HEPACAM2	−4.6630	−4.4554	−4.8929	−5.5554
ITLN1	−5.3944	−5.3083	−5.4875	−6.0155
LGALS2	−3.5660	−3.8652	−3.5149	−3.9496
MUC12	−2.7603	−2.8897	−2.4149	−2.5861
NXPE1	−3.6871	−3.7573	−4.0759	−4.0087
TIMP1	1.1477	1.4022	1.5606	1.4898
GCG	−7.1766	−7.2416	−6.8744	−6.0155

DEGs, differentially expressed genes.
$P < 0.05$.

There are few studies about *HEPACAM2, ITLN1, LGALS2, MUC12*, and *NXPE1*, even rare in CRC. These

Table 5. GO term enrichment analysis of seven DEGs.

Genes	Species	Biology process	Cellular component	Molecular function
HEPACAM2	Homo sapiens	GO:0007067~mitotic nuclear division, GO:0051301~cell division,	GO:0005819~spindle, GO:0030496~midbody,	GO:0005515~protein binding,
ITLN1	Homo sapiens	GO:0001934~positive regulation of protein phosphorylation,	GO:0031225~anchored component of membrane, GO:0031526~brush border membrane, GO:0070062~extracellular exosome,	GO:0030246~carbohydrate binding,
LGALS2	Homo sapiens			GO:0005515~protein binding, GO:0030246~carbohydrate binding,
MUC12	Homo sapiens	GO:0001558~regulation of cell growth, GO:0016266~O-glycan processing,	GO:0005796~Golgi lumen, GO:0005887~integral component of plasma membrane,	
NXPE1	Homo sapiens		GO:0005576~extracellular region,	
TIMP1	Homo sapiens	GO:0002576~platelet degranulation, GO:0009725~response to hormone, GO:0022617~extracellular matrix disassembly, GO:0034097~response to cytokine, GO:0042060~wound healing, GO:0043066~negative regulation of apoptotic process, GO:0043434~response to peptide hormone, GO:0051216~cartilage development,	GO:0005576~extracellular region, GO:0005578~proteinaceous extracellular matrix, GO:0005581~collagen trimer, GO:0005604~basement membrane, GO:0005615~extracellular space, GO:0031093~platelet alpha granule lumen, GO:0070062~extracellular exosome,	GO:0002020~protease binding, GO:0005125~cytokine activity, GO:0005515~protein binding, GO:0008083~growth factor activity,
GCG	Homo sapiens	GO:0008283~cell proliferation, GO:0010800~positive regulation of peptidyl-threonine phosphorylation, GO:0043066~negative regulation of apoptotic process, GO:0070374~positive regulation of ERK1 and ERK2 cascade,	GO:0005576~extracellular region, GO:0005615~extracellular space, GO:0005788~endoplasmic reticulum lumen, GO:0005886~plasma membrane,	GO:0005102~receptor binding, GO:0005179~hormone activity, GO:0005515~protein binding,

DEGs, differentially expressed genes; GO, Gene Ontology analysis.

five genes all presented a sequentially descending trend in expression through the normal colorectal mucosa–adenoma–carcinoma sequence. *HEPACAM2* is a member of the immunoglobulin family of adhesion genes. The clinical importance of *HEPACAM2* in CRC remains unclear. *ITLN1* encodes intelectin-1, which functions as a receptor for both bacterial arabinogalactans and lactoferrin. Li et al. [51] noted that intelectin-1 suppressed tumor progression and was associated with improved survival in gastric cancer. However, there is no research exploring the function of *ITLN1* in CRC. *LGALS2* encodes galectin-2, which participates in non-small-cell lung cancer [52], coronary heart disease [53], and ischemic stroke [54] instead of CRC. *MUC12* is a member of mucin family. Matsuyama et al. [55] reported that *MUC12* mRNA expression was an independent marker of prognosis in stage II and stage III colorectal cancer. For *NXPE1*, we found two studies, which were bioinformatics studies, but they did not determine the function of this gene. Although there are few studies on the functions of these genes, we used GO analysis and KEGG pathway analysis to predict the possible function of the genes and the possible signaling

Table 6. KEGG pathway analysis of seven DEGs.

Genes	Species	KEGG pathway
HEPACAM2	Homo sapiens	/
ITLN1	Homo sapiens	/
LGALS2	Homo sapiens	/
MUC12	Homo sapiens	/
NXPE1	Homo sapiens	/
TIMP1	Homo sapiens	hsa04066: HIF-1 signaling pathway,
GCG	Homo sapiens	hsa04911: Insulin secretion, hsa04922: Glucagon signaling pathway,

DEGs, differentially expressed genes; KEGG, Kyoto Encyclopedia of Genes and Genomes Pathway.

Table 7. The expression level and the relationship with prognosis of seven DEGs from patients in our hospital for medical treatment.

Genes	Normal stage 1 |FC|	Normal stage 2 |FC|	Normal stage 3 |FC|	Normal stage 4 |FC|	Three-year survival rate (*P*-value)
GCG	−5.32297	−3.3835	−4.15322	−5.56661	0.95118
HEPACAM2	−3.09828	−1.72218	−2.58992	−1.76615	0.09793
ITLN1	−1.86113	−1.51937	−1.64402	−1.13096	0.8532
LGALS2	−3.18707	−2.26365	−2.63189	−2.53429	0.0765
MUC12	−1.37528	−2.07949	−2.08844	−2.87224	0.14525
NXPE1	−2.06956	−2.71137	−2.60975	−2.77593	0.0068
TIMP1	1.00443	2.180932	2.418588	2.083604	0.2753

DEGs, differentially expressed genes; FC, fold change.

pathways, providing a direction for subsequent functional research.

The expression pattern of these seven DEGs was confirmed by 672 RNA-seq data on TCGA. Their expression levels presented the same trend as that in the GEO database, except for *MUC12*. However, *MUC12* was also downregulated in colorectal cancer tissues compared with normal colorectal mucosa, and its expression levels in our own patient validation cases were consistent with the predictions of the GEO database, indicating that *MUC12* is a promising marker. However, one-third of cases in TCGA were extracted again for further prognosis validation, and all seven DEGs showed a close relationship with patient prognosis (Table S4).

We also confirmed the expression levels and the relationship with prognosis of these seven DEGs using patients in our hospital for medical treatment. Gene sequencing was performed on the surgical samples of 28 patients with CRC who had not received antitumor treatment. Consistent with the bioinformatics predictions, *MUC12* and *NXPE1* presented a sequentially descending trend in expression with tumor progression, and *TIMP1* presented a sequentially ascending trend (Table 5). Among these genes, only *NXPE1* was considered related to patient prognosis of the three-year survival rate (Fig. S3, Table 7). There was a much larger mismatch between the validation results of our own patient cases and TCGA RNA-seq data, probably because of the scarcity of patients compared with the large number of validation cases in TCGA. Nevertheless, our findings suggest that DEGs play an important role in the development of CRC.

There are several limitations in our study. First, the amount of data we obtained from the GEO database and our validation were still not sufficient. However, these were all the data we could obtain from the GEO while still insuring data quality. In the future, more qualified patients in our hospital for medical treatment should be followed up. Second, the data in the GEO database are based on different experimental studies, and there was a lack of uniform standards, which would add heterogeneity to our findings. Thus, we attempted to minimize heterogeneity and to insure the rigor of the data by only including GPL570 platform data, performing quality assessment of the microarray and preprocessing the data. Although the batch effect was evaluated across the different datasets, and the heterogeneity was not significant, it still exists.

In conclusion, we discovered seven DEGs through the normal colorectal mucosa–adenoma–carcinoma sequence associated with CRC patient prognosis. Six genes that present sequential expression level changes with different stages might reflect the degree of tumor progression. One downregulated gene might play a key role in the early stages of neoplasia. Monitoring changes in these gene expression levels will allow us to assess disease progression, evaluate treatment efficacy, and predict prognosis. In addition, our study provides a set of useful targets for further functional research exploring the molecular mechanisms and uncovering new therapeutic targets.

Ethical Approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. For this type of study, formal consent is not required.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (81572592, 81772543), the Zhejiang Province Preeminence Youth Fund (LR16H160001), the Zhejiang Natural Sciences Foundation Grant (LZ15H160001, LY17H160029, and Q17H160042), the National Health and Family Planning Commission Fund (2015112271), and the Zhejiang medical innovative discipline construction project-2016.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Siegel, R. L., K. D. Miller, and A. Jemal. 2016. Cancer statistics, 2016. CA Cancer J. Clin. 66:7–30.
2. Siegel, R., C. DeSantis, K. Virgo, K. Stein, A. Mariotto, T. Smith, et al. 2012. Cancer treatment and survivorship statistics, 2012. CA Cancer J. Clin. 62:220–241.
3. Brenner, H., A. M. Bouvier, R. Foschi, M. Hackl, I. K. Larsen, V. Lemmens, et al. 2012. Progress in colorectal cancer survival in Europe from the late 1980s to the early 21st century: the EUROCARE study. Int. J. Cancer 131:1649–1658.
4. Sankaranarayanan, R., R. Swaminathan, H. Brenner, K. Chen, K. S. Chia, J. G. Chen, et al. 2010. Cancer survival in Africa, Asia, and Central America: a population-based study. Lancet Oncol. 11:165–173.
5. Siegel, R. L., K. D. Miller, S. A. Fedewa, D. J. Ahnen, R. G. S. Meester, A. Barzi, et al. 2017. Colorectal cancer statistics, 2017. CA Cancer J. Clin. 67:177–193.
6. Rutter, C. M., E. Johnson, D. L. Miglioretti, M. T. Mandelson, J. Inadomi, and D. S. Buist. 2012. Adverse events after screening and follow-up colonoscopy. Cancer Causes Control 23:289–296.
7. Nannini, M., M. A. Pantaleo, A. Maleddu, A. Astolfi, S. Formica, and G. Biasco. 2009. Gene expression profiling in colorectal cancer using microarray technologies: results and perspectives. Cancer Treat. Rev. 35:201–209.
8. De Sousa, E. M. F., X. Wang, M. Jansen, E. Fessler, A. Trinh, L. P. de Rooij, et al. 2013. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. Nat. Med. 19:614–618.
9. Zauber, A. G., S. J. Winawer, M. J. O'Brien, I. Lansdorp-Vogelaar, M. van Ballegooijen, B. F. Hankey, et al. 2012. Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths. N. Engl. J. Med. 366:687–696.
10. Kinzler, K. W., and B. Vogelstein. 1996. Lessons from hereditary colorectal cancer. Cell 87:159–170.
11. Fearon, E. R. 2011. Molecular genetics of colorectal cancer. Annu. Rev. Pathol. 6:479–507.
12. Liu, Y., Y. Zhou, X. Feng, P. Yang, J. Yang, P. An, et al. 2014. Low expression of microRNA-126 is associated with poor prognosis in colorectal cancer. Genes Chromosom. Cancer 53:358–365.
13. Tsukamoto, S., T. Ishikawa, S. Iida, M. Ishiguro, K. Mogushi, H. Mizushima, et al. 2011. Clinical significance of osteoprotegerin expression in human colorectal cancer. Clin. Cancer Res. 17:2444–2450.
14. Alizadeh, A. A., M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, et al. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403:503–511.
15. Kihara, C., T. Tsunoda, T. Tanaka, H. Yamana, Y. Furukawa, K. Ono, et al. 2001. Prediction of sensitivity of esophageal tumors to adjuvant chemotherapy by cDNA microarray analysis of gene-expression profiles. Can. Res. 61:6474–6479.
16. Wang, Y., T. Jatkoe, Y. Zhang, M. G. Mutch, D. Talantov, J. Jiang, et al. 2004. Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer. J. Clin. Oncol. 22:1564–1571.
17. Arango, D., P. Laiho, A. Kokko, P. Alhopuro, H. Sammalkorpi, R. Salovaara, et al. 2005. Gene-expression profiling predicts recurrence in Dukes' C colorectal cancer. Gastroenterology 129:874–884.
18. Pesson, M., A. Volant, A. Uguen, K. Trillet, P. De La Grange, M. Aubry, et al. 2014. A gene expression and pre-mRNA splicing signature that marks the adenoma-adenocarcinoma progression in colorectal cancer. PLoS ONE 9:e87761.
19. Shih, W., R. Chetty, and M. S. Tsao. 2005. Expression profiling by microarrays in colorectal cancer (Review). Oncol. Rep. 13:517–524.
20. Fu, J., W. Tang, P. Du, G. Wang, W. Chen, J. Li, et al. 2012. Identifying microRNA-mRNA regulatory network in colorectal cancer by a combination of expression profile and bioinformatics analysis. BMC Syst. Biol. 6:68.
21. Robles, A. I., G. Traverso, M. Zhang, N. J. Roberts, M. A. Khan, C. Joseph, et al. 2016. Whole-exome sequencing analyses of inflammatory bowel disease-associated colorectal cancers. Gastroenterology 150:931–943.
22. Gautier, L., L. Cope, B. M. Bolstad, and R. A. Irizarry. 2004. affy–analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20:307–315.
23. Tumor Analysis Best Practices Working Group. 2004. Expression profiling–best practices for data generation and interpretation in clinical trials. Nat. Rev. Genet. 5:229–237.
24. Galamb, O., F. Sipos, S. Spisak, B. Galamb, T. Krenacs, G. Valcz, et al. 2009. Potential biomarkers of colorectal adenoma-dysplasia-carcinoma progression: mRNA expression profiling and in situ protein detection on TMAs reveal 15 sequentially upregulated and 2 downregulated genes. Cell Oncol. 31:19–29.
25. Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, et al. 2003. Exploration, normalization, and summaries of high

density oligonucleotide array probe level data. Biostatistics 4:249–264.
26. Kou, Y., S. Zhang, X. Chen, and S. Hu. 2015. Gene expression profile analysis of colorectal cancer to investigate potential mechanisms using bioinformatics. Onco Targets Ther. 8:745–752.
27. Liu, Y. J., S. Zhang, K. Hou, Y. T. Li, Z. Liu, H. L. Ren, et al. 2013. Analysis of key genes and pathways associated with colorectal cancer with microarray technology. Asian Pac. J. Cancer Prev. 14:1819–1823.
28. Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25:25–29.
29. Gene Ontology Consortium 2006. The Gene Ontology (GO) project in 2006. Nucleic Acids Res. 34:D322–D326.
30. Kanehisa, M., and S. Goto. 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28:27–30.
31. Sakai, E., M. Fukuyo, K. Ohata, K. Matsusaka, N. Doi, Y. Mano, et al. 2016. Genetic and epigenetic aberrations occurring in colorectal tumors associated with serrated pathway. Int. J. Cancer 138: 1634–1644.
32. Heijink, D. M., J. H. Kleibeuker, M. Jalving, W. Boersma-van Ek, J. J. Koornstra, J. Wesseling, et al. 2007. Independent induction of caspase-8 and cFLIP expression during colorectal carcinogenesis in sporadic and HNPCC adenomas and carcinomas. Cell Oncol. 29:409–419.
33. Liang, B., C. Li, and J. Zhao. 2016. Identification of key pathways and genes in colorectal cancer using bioinformatics analysis. Med. Oncol. 33:111.
34. Guo, Y., Y. Bao, M. Ma, and W. Yang. 2017. Identification of key candidate genes and pathways in colorectal cancer by integrated bioinformatical analysis. Int. J. Mol. Sci. 18:722.
35. Hewedi, I. H., R. M. Farid, K. F. Sidhom, M. I. Salman, and R. G. Mostafa. 2017. Differential expression of cytochrome c oxidase subunit I along the colorectal adenoma: carcinoma progression. Appl. Immunohistochem. Mol. Morphol. doi: 10.1097/PAI. 0000000000000509. [Epub ahead of print].
36. Andersen, V., L. K. Vogel, T. I. Kopp, M. Saebo, A. W. Nonboe, J. Hamfjord, et al. 2015. High ABCC2 and low ABCG2 gene expression are early events in the colorectal adenoma-carcinoma sequence. PLoS ONE 10:e0119255.
37. Clough, E., and T. Barrett. 2016. The gene expression omnibus database. Methods Mol. Biol. 1418:93–110.
38. Gao, S., J. Ou, and K. Xiao. 2014. R language and bioconductor in bioinformatics applications (Chinese Edition). Tianjin Science and Technology Translation Publishing Co., Ltd, Tianjin.
39. Brew, K., and H. Nagase. 2010. The tissue inhibitors of metalloproteinases (TIMPs): an ancient family with structural and functional diversity. Biochim. Biophys. Acta 1803:55–71.
40. Kim, Y. S., S. H. Kim, J. G. Kang, and J. H. Ko. 2012. Expression level and glycan dynamics determine the net effects of TIMP-1 on cancer progression. BMB Rep. 45:623–628.
41. Ma, J., J. Wang, W. Fan, X. Pu, D. Zhang, C. Fan, et al. 2014. Upregulated TIMP-1 correlates with poor prognosis of laryngeal squamous cell carcinoma. Int. J. Clin. Exp. Pathol. 7:246–254.
42. Tarhini, A. A., Y. Lin, O. Yeku, W. A. LaFramboise, M. Ashraf, C. Sander, et al. 2014. A four-marker signature of TNF-RII, TGF-alpha, TIMP-1 and CRP is prognostic of worse survival in high-risk surgically resected melanoma. J. Transl. Med. 12:19.
43. Lau, T. P., A. C. Roslani, L. H. Lian, H. C. Chai, P. C. Lee, I. Hilmi, et al. 2014. Pair-wise comparison analysis of differential expression of mRNAs in early and advanced stage primary colorectal adenocarcinomas. BMJ Open 4:e004930.
44. Holten-Andersen, M. N., H. J. Nielsen, S. Sorensen, V. Jensen, N. Brunner, and I. J. Christensen. 2006. Tissue inhibitor of metalloproteinases-1 in the postoperative monitoring of colorectal cancer. Eur. J. Cancer 42:1889–1896.
45. Weidle, U. H., F. Birzele, and A. Kruger. 2015. Molecular targets and pathways involved in liver metastasis of colorectal cancer. Clin. Exp. Metastasis 32:623–635.
46. Bunatova, K., M. Pesta, V. Kulda, O. Topolcan, J. Vrzalova, A. Sutnar, et al. 2012. Plasma TIMP1 level is a prognostic factor in patients with liver metastases. Anticancer Res. 32:4601–4606.
47. Zhang, L., M. Zhang, J. J. Wang, C. J. Wang, Y. C. Ren, B. Y. Wang, et al. 2016. Association of TCF7L2 and GCG gene variants with insulin secretion, insulin resistance, and obesity in new-onset diabetes. Biomed. Environ. Sci. 29:814–817.
48. Wang, J., G. Yan, J. Zhang, K. Gao, M. Zhang, L. Li, et al. 2015. Association of LRP5, TCF7L2, and GCG variants and type 2 diabetes mellitus as well as fasting plasma glucose and lipid metabolism indexes. Hum. Immunol. 76:339–343.
49. Drucker, D. J. 2003. Glucagon-like peptides: regulators of cell proliferation, differentiation, and apoptosis. Mol. Endocrinol. 17:161–171.
50. Spisak, S., A. Kalmar, O. Galamb, B. Wichmann, F. Sipos, B. Peterfia, et al. 2012. Genome-wide screening of genes regulated by DNA methylation in colon cancer development. PLoS ONE 7:e46215.
51. Li, D., X. Zhao, Y. Xiao, H. Mei, J. Pu, X. Xiang, et al. 2015. Intelectin 1 suppresses tumor progression and is associated with improved survival in gastric cancer. Oncotarget 6:16168–16182.

52. Uso, M., E. Jantus-Lewintre, S. Calabuig-Farinas, A. Blasco, E. Garcia Del Olmo, R. Guijarro, et al. 2017. Analysis of the prognostic role of an immune checkpoint score in resected non-small cell lung cancer patients. Oncoimmunology 6:e1260214.
53. Tian, J., S. Hu, F. Wang, X. Yang, Y. Li, and C. Huang. 2015. PPARG, AGTR1, CXCL16 and LGALS2 polymorphisms are correlated with the risk for coronary heart disease. Int. J. Clin. Exp. Pathol. 8:3138–3143.
54. Misra, S., P. Kumar, A. Kumar, R. Sagar, K. Chakravarty, and K. Prasad. 2016. Genetic association between inflammatory genes (IL-1 alpha, CD14, LGALS2, PSMA6) and risk of ischemic stroke: a meta-analysis. Meta Gene 8:21–29.
55. Matsuyama, T., T. Ishikawa, K. Mogushi, T. Yoshida, S. Iida, H. Uetake, et al. 2010. MUC12 mRNA expression is an independent marker of prognosis in stage II and stage III colorectal cancer. Int. J. Cancer 127:2292–2299.

Supporting Information

Additional supporting information may be found in the online version of this article:

Figure S1. Quality assessment.

Figure S2. Batch effect was evaluated with the expression level of GAPDH across the different datasets, and heterogeneity was not significant ($P > 0.05$).

Figure S3. Survival curve of **A.** HEPACAM2, **B.** ITLN1, **C.** LGALS2, **D.** MUC12, **E.** NXPE1, **F.** TIMP1 and **G.** GCG from patients in our hospital for medical treatment.

Table S1. Search strategies.

Table S2. Summary of thirty-eight datasets.

Table S3. Eighty-seven genes present sequentially expression level changes through normal colorectal mucosa-adenoma-carcinoma sequence.

Table S4. Further prognosis validation with one-third cases on TCGA.