IJC
International Journal of Cancer

High-risk Stage III colon cancer patients identified by a novel five-gene mutational signature are characterized by upregulation of IL-23A and gut bacterial translocation of the tumor microenvironment

Weiting Ge[1,*], Hanguang Hu[1,2,*], Wen Cai[1,2,*], Jinhong Xu[3], Wangxiong Hu[1], Xingyue Weng[1], Xin Qin[4], Yanqin Huang[1], Weidong Han[5], Yeting Hu[1,7], Jiekai Yu[1], Wufeng Zhang[1], Sisi Ye[1], Lina Qi[1], Pingjie Huang[6], Lirong Chen[1,3], Kefeng Ding[1,7], Li Dong Wang[8] and Shu Zheng[1]

[1]Cancer Institute (Key Laboratory of Cancer Prevention and Intervention, China National Ministry of Education), The Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China
[2]Department of Medical Oncology, The Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China
[3]Department of Pathology, The Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China
[4]Medical College, Hubei University of Arts and Science, Xiangyang, China
[5]Department of Medical Oncology, Biomedical Research Center, Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University, Hangzhou, China
[6]State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou, China
[7]Department of Surgical Oncology, The Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China
[8]Henan Key Laboratory for Esophageal Cancer Research of the First Affiliated Hospital, State Key Laboratory for Esophageal Cancer Prevention & Treatment, Zhengzhou University, Zhengzhou, Henan, China

The heterogeneities of colorectal cancer (CRC) lead to staging inadequately of patients' prognosis. Here, we performed a prognostic analysis based on the tumor mutational profile and explored the characteristics of the high-risk tumors. We sequenced 338 colorectal carcinomas as the training dataset, constructed a novel five-gene (*SMAD4*, *MUC16*, *COL6A3*, *FLG* and *LRP1B*) prognostic signature, and validated it in an independent dataset from The Cancer Genome Atlas (TCGA). Kaplan–Meier and Cox regression analyses confirmed that the five-gene signature is an independent predictor of recurrence and prognosis in patients with Stage III colon cancer. The mutant signature translated to an increased risk of death (hazard ratio = 2.45, 95% confidence interval = 1.15–5.22, $p = 0.016$ in our dataset; hazard ratio = 4.78, 95% confidence interval = 1.33–17.16, $p = 0.008$ in TCGA dataset). RNA and bacterial 16S rRNA sequencing of high-risk tumors indicated that mutations of the five-gene signature may lead to intestinal barrier integrity, translocation of gut bacteria and deregulation of immune response and extracellular related genes. The high-risk tumors overexpressed *IL23A* and *IL1RN* genes and enriched with cancer-related bacteria (*Bacteroides fragilis*, *Peptostreptococcus, Parvimonas, Alloprevotella and Gemella*) compared to the low-risk tumors. The signature identified the high-risk group characterized by gut bacterial translocation and upregulation of interleukins of the tumor microenvironment, which was worth further researching.

Tumor Markers and Signatures

*W.G., H.H. and W.C. contributed equally to this work

Additional Supporting Information may be found in the online version of this article.

Key words: colorectal cancer, prognostic signature, tumor mutational profile, tumor microenvironment

Abbreviations: CRC: colorectal cancer; DFS: disease-free survival; HR: hazard ratio; IL1RN: interleukin-1 receptor antagonist; IL23A: interleukin 23 alpha unit; MMR: DNA mismatch repair; OS: overall survival; OTU: Operational Taxonomic Units; TCGA: The Cancer Genome Altas; TME: tumor microenvironment; ZJU: Zhejiang university

Conflict of interest: The authors have declared that no competing interest exists.

Grant sponsor: National High Technology Research and Development Program of China; **Grant numbers:** 2012AA02A204, 2012AA02A50; **Grant sponsor:** National Human Genetic Resources Sharing Service Platform; **Grant number:** 2005DKA21300; **Grant sponsor:** The National Key R&D Program of China; **Grant number:** 2017YFC0908200; **Grant sponsor:** The National Natural Science Foundation of China; **Grant number:** U1804262
DOI: 10.1002/ijc.32775
History: Received 17 May 2019; Accepted 30 Oct 2019; Online 6 Nov 2019

Correspondence to: Weiting Ge, E-mail: geweiting@zju.edu.cn and Shu Zheng, E-mail: zhengshu@zju.edu.cn

What's new?

The heterogeneity of colorectal cancer makes it difficult to determine patients with a poor prognosis or in need of advanced therapy. Here, the authors constructed and validated a novel 5-gene (*SMAD4*, *MUC16*, *COL6A3*, *FLG*, and *LRP1B*) mutational prognostic signature to identify high-risk patients with Stage III colon cancer. Mutations of these five genes may lead to loss of intestinal barrier integrity, translocation of gut bacteria, and deregulation of interleukins and extracellular-related genes. Combining tumor genetic characteristics with dynamic tumor microenvironment changes may lead to more promising prognostic signatures, which could help better select cancer patients for systemic therapy after surgery.

Introduction

The heterogeneity of colorectal cancer (CRC) makes it difficult to determine which patients have a worse prognosis and which patients require further therapy beyond surgical resection.[1] TNM staging system is an important guide for physicians regarding treatment and prognosis. Early-stage disease is defined as cancers that have only locally invaded (Stage I–II) or that presentation with regional lymph-node metastases (Stage III). Adjuvant chemotherapy provides a survival benefit in patients with Stage III disease, and possibly in those with high-risk Stage II colon cancer.[2] However, the benefits of adjuvant chemotherapy are limited; some patients never have a relapse despite no treatment, whereas many patients have disease relapse despite therapy.[3,4] The current staging method is suboptimal due to the variation in outcomes that exist among patient in the same stage. Indeed, a better prognostic biomarker for outcome prediction and therapy assignment is urgently needed.

Single genetic characteristics, such as DNA mismatch repair (MMR) deficiency status, *RAS*-mutation or *BRAF*-mutation, have been proposed as prognostic biomarkers.[1,5] Somatic mutational profiling based molecular signatures have been developed to detecting patients at a high risk of recurrence.[6,7] Moreover, genetic events, gene-expression profile and the tumor microenvironment were integrated to enable four consensus molecular subtypes.[8] However, these molecular markers are difficult to integrate into the current staging system. The need for multiple detection methods also limits the clinical practical application of the above prognostic markers. Currently, the TNM staging system is currently the gold standard method used to predict prognosis and aid treatment decisions for CRC patients. There is a need to add prognostic and predictive value to the current staging system, which could be achieved with the use of validated biomarkers.

Thus, we identified the somatic mutations of CRC patients, constructed a novel stage-specific prognostic signature, and validated the signature with an independent sequencing data from The Cancer Genome Atlas (TCGA) data portal.[9] Furthermore, we examined the alterated gene/pathway and tumor microenvironment (TME) changes to investigate the cause of increased risk associated with the prognostic signature.

Materials and Methods

Sample collection and genomic DNA preparation

Tumor and matched normal mucosa-derived DNA was purified using a QIAamp DNA mini kit (QIAGEN, Hilden, Germany). After surgery, fresh tissue specimens were flash-frozen in liquid nitrogen and stored at −80°C until genomic DNA isolation. The pathologic diagnosis of each case was confirmed by reviewing hematoxylin and eosin (H&E)-stained slides, and the samples were excluded if they contained <40% tumor cells. All samples were obtained from patients diagnosed with primary CRC without chemotherapy prior to surgery. No patients developed or died from severe postoperative complications. After surgery, no Stage I patient received adjuvant chemotherapy. Only one Stage I rectal cancer patient received oral capecitabine after recurrence. Some Stage II and III patients were treated with standard 5-fluorouracil-based chemotherapy. All Stage IV patients received chemotherapy, including a 5-fluorouracil-based regimen and further EGFR- or VEGFR-targeted therapy as the second-line treatment. The stage was assessed using the 7th version of the American Joint Commission on Cancer guidelines. All patients signed a patient consent form and that our study was conducted in accordance with the Declaration of Helsinki. All procedures were approved by the Institutional Review Board (IRB) of the Second Affiliated Hospital, School of Medicine of Zhejiang University under protocol 2013-042.

DNA sequencing and detection of somatic mutations and indels

All samples were sequenced on the Illumina platform at Novogene Co. Ltd. The whole-genome sequencing was performed on an Illumina HiseqX. The exome sequencing was performed using an Agilent's SureSelect V5 exome enrichment kit (Agilent Technologies, Santa Clara, CA) on an Illumina HiSeq 2500. The panel sequencing was performed using a custom-designed panel that utilizes Agilent SureSelect technology to target the exonic region of 524 genes (Supporting Information Table S1) on an Illumina HiSeq 2000. The sequence reads were aligned to the human genome (GRCh37/hg19), and unique pairs were used for variant calling. Candidate variants and indels were detected using the Genome Analysis Toolkit.[10] Then, the somatic mutations and indels were identified using MuTect[11] and Strelka,[12] respectively. These variants were annotated with ANNOVAR.[13]

Identification of somatic mutations of colorectal cancer

We conducted two-phase sequencing to identify the somatic mutations in CRC patients recruited from the Second Affiliated Hospital, School of Medicine of Zhejiang University (ZJU). The tumor and normal mucosa tissues from 80 cases were subjected

to whole-exome capture sequencing (70 cases) or whole-genome sequencing (10 cases). The somatic mutations in the exomic region were subjected to further analysis. Combined with previously reported CRC driver genes (TCGA,[9] COSMIC[14] and NCCN guidelines[15]), the recurrently mutated genes in phase one were used to design a 524-gene panel. Then, the target regions of 258 additional cases were sequenced using this panel.

Clinical outcomes and testing of the prognostic mutational signature

The strategy used to derive and validate the prognostic mutational signatures is presented in Figure 1. Recurrently mutated genes in the nonhypermutated CRC samples were combined to form prognostic mutational signatures. Patients were separated into groups according to the gene mutation status of the signature. After calculated the different combinations, we define the patients without any gene mutation in the wild-type group and the patients with mutations in at least one of the genes in the mutant group. A Cox proportional hazard (PH) model was employed to evaluate the association between the signature and the clinical endpoints (overall survival [OS] and disease-free survival [DFS]). To rule out overfitting of the model, four separate statistical approaches were applied. First, we developed a k-gene mutational signature that was based on the Cox PH model of the training dataset and then applied the same signature from the Cox PH model to the test dataset. The k-gene signature was required statistically significant for the training and test dataset when subjected to both the hazard ratio (HR) and log-rank tests ($p < 0.05$). As the test dataset, somatic mutation profile and the clinicopathological information of the TCGA cohort were obtained from the TCGA project data portal (http://www.cbioportal.org) on July 2, 2017. Second, Harrell's concordance index (C-index) was used to quantify the predictive accuracies. To determine the minimum size of signature size that could discriminate the outcomes, we started the training and testing with one gene signature and stopped increasing the size of the gene signature once we obtained the maximal C-index. Third, the candidate signatures were subjected to univariate survival analysis in clinical and molecular subgroups to determine the optimal combination of gene signatures. Finally, multiple permutation testing of the selected signature was performed. The final follow-up of the patients from the Second Affiliated Hospital, School of Medicine of Zhejiang University (the ZJU cohort) occurred on October 1, 2016, while that of the patients from the TCGA cohort occurred on August 20, 2015. Only patients with survival data for more than 15-months were included in the prognostic-related analysis. The OS was measured in months from the date of surgery to the date of patient death. DFS was defined as months from the date of surgery to the date of first relapse.

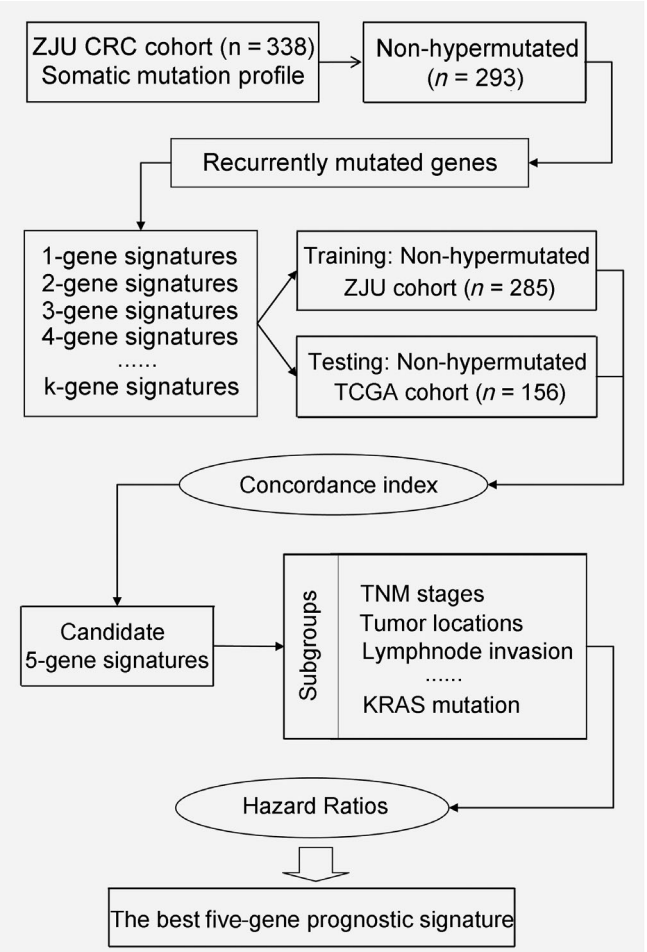

Figure 1. Strategy used to derive and validate the prognostic mutational signatures. Genes recurrently mutated in nonhypermutated colorectal cancer (CRC) were combined to form k-gene prognostic mutational signatures. A Cox proportional hazard model was employed to evaluate the association between the signature and the clinical endpoints. The filtered signature of the training cohort and testing cohort exhibited statistical significance in both the hazard ratio (HR) and log-rank tests ($p < 0.05$). The concordance index (C-index) was calculated to validate the predictive value. We began by training and testing one gene signature and stopped increasing the size of the gene signature (k) once we obtained the maximal C-index. The top five 5-gene signatures were subjected to a univariate survival analysis in clinical and molecular subgroups. The signature with significant HR (HR > 1 and $p < 0.05$) in most subgroups was selected as the best prognostic signature.

Bacterial 16S ribosomal RNA (16S rRNA) sequencing and the relative abundance evaluation of microbes

The total DNA was extracted from the tumor and matched normal mucosa tissues was extracted using the CTAB method. The bacterial 16S rRNA gene of distinct regions (16S V3–V4) was amplified using Phusion High-fidelity PCR mast mix (New England Biolabs, Ipswich, MA). The sequencing libraries were generated using an Ion plus fragment library kit (Thermo Scientific, Waltham, MA) and sequenced on an Ion S5 XL platform. The sequencing analysis was performed by Uparse software.[16] Sequences with ≥97% similarity were assigned to the same operational taxonomic units (OTU). The Silva Database was used

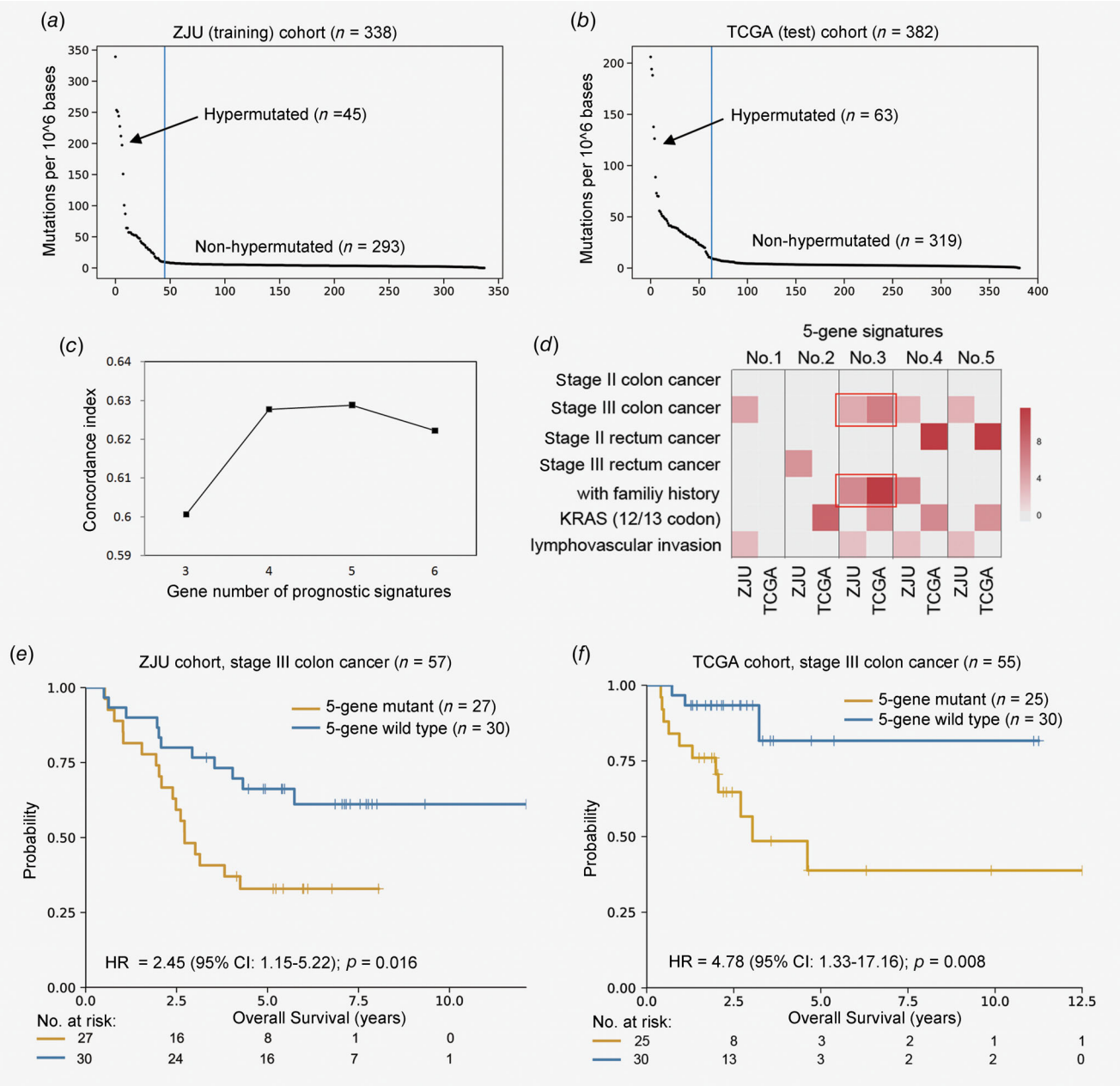

Figure 2. The five-gene mutational signature is associated with overall survival in patients with Stage III colon cancer. Mutation frequencies in each of the tumor samples from the ZJU cohort (*a*) and the TCGA cohort (*b*). The vertical line indicates the separation of hypermutated and nonhypermutated tumors (at 10 Mut/Mb). The maximal concordance index value was obtained by applying the five-gene mutational signature (*c*). Among the top five five-gene mutational signatures, signature No. 3 revealed two subgroups with significant HR (HR › 1 and p ‹ 0.05) both in the two cohorts (*d*). Kaplan–Meier estimates of overall survival were analyzed in Stage III colon cancer patients from the ZJU cohort (*e*) and the TCGA cohort (*f*) were analyzed. p values were calculated using the log-rank test. Abbreviations: CI, confidence interval; HR, hazard ratio. [Color figure can be viewed at wileyonlinelibrary.com]

based on the Mothur algorithm to annotate taxonomic information.[17] Bacterial community richness and diversity were evaluated by the Chao 1 estimator and the Shannon index, separately.

RNA sequencing and quantitative real-time PCR analysis

Tumor and matched normal mucosa-derived RNA was purified using an RNeasy mini kit (QIAGEN). The sequencing libraries were generated using a NEBNext Ultra RNA Library Prep Kit for Illumina (NEB and sequenced on an Illumina Hiseq platform). HTSeq v0.6.0 was used to count the reads numbers mapped to each gene. The differential expression analysis of two groups was performed using the DESeq2 R package (1.10.1). The cell type enrichment score of the tumor and matched normal mucosa tissue was calculated using xCell.[18] Real-time

quantitative (RT-qPCR) was performed using iTaq Universal SYBR Green Supermix (Biorad) on an Applied Biosystems 7500 Real-Time PCR machine. Expression data were normalized to GAPDH mRNA expression. Primer sequences are listed in Supporting Information Table S2 and were obtained from the PrimerBank (https://pga.mgh.harvard.edu/primerbank/).

Statistical analyses

The Kaplan–Meier survival curve analysis with a log-rank test was used to estimate the mutational signature in relation to OS and DFS. Fisher's exact test, Student's *t* test and the Mann–Whitney *U* test were used to determine the differences in the clinicopathological variables between the risk groups. A multivariate Cox regression analysis was performed to determine the contribution of the mutational signature to survival after adjusting for age, sex and stage. The Wald test was used in the multivariate Cox regression analysis. All statistical analyses were two-sided. A value of $p < 0.05$ was considered statistically significant. All analyses were performed using Python (3.6.0) and R (3.4.0).

Data availability

The whole-genome and exome capture sequencing data were deposited in the European Nucleotide Archive under study accession number EGAS00001001269.

Results

Basic information of patients and identification of somatic mutations

In total, we obtained somatic mutation profiles of the ZJU cohort including 338 CRC cases. The mutations in the targeted captured regions of 293 nonhypermutated CRC cases were subjected to further analysis (Fig. 2*a*). The threshold of hypermutation (>10 Mutations per Megabase) was determined according to a recent large-scale sequencing analysis.[19] As an independent testing dataset, 319 nonhypermutated samples from 382 CRC cases from the TCGA cohort were selected according to the same threshold (Fig. 2*b*). The threshold was visually confirmed by an uptick in the slope of the line (Figs. 2*a* and 2*b*). The demographics of all colorectal cancer patients in the two cohorts are presented in Supporting Information Table S3.

Identification of a novel five-gene signature for CRC patients with poor outcomes

To construct the prognostic mutational signature, 43 recurrently mutated genes with a ≥5% frequency of occurrence in colon or rectal cancer were included to allow for sufficient representation of the patients (Supporting Information Table S4). The patients with mutations in at least one gene of the signature constituted the mutant group. The patients without any gene mutation of the signature constituted the wild-type group. To rule out overfitting of the model, four separate statistical approaches which were mentioned in methods were applied. We developed k-gene mutational signatures that were based on the Cox PH model of the ZJU cohort ($n = 258$) and then applied the same signature from the Cox PH model to the TCGA cohort ($n = 236$). Five genes were determined as the optimal size of mutational signature when the maximal C-index reached (Fig. 2*c*). The top five five-gene signatures (Supporting Information Table S5) were subjected to a univariate survival analysis in clinical and molecular subgroups. As shown in Figure 2*d*, using signature No. 3, two subgroups exhibited significant HR (HR >1 and $p < 0.05$) in the two cohorts. Thus, we constructed the five-gene prognostic mutational signature as follows: *COL6A3*, *FLG*, *LRP1B*, *MUC16* and *SMAD4*. In addition, multiple permutation testing was performed to confirm the robustness of this model (Supporting Information Fig. S1).

Table 1. The five-gene signature is associated with a significantly increased risk of death in patients with Stage III colon cancer

Overall survival	Hazard ratio (95% CI)	*p*
ZJU cohort		
All stages colon cancer	2.68 (1.45–4.94)	0.001
Stage II colon cancer	2.26 (0.52–9.75)	0.264
Stage III colon cancer	2.45 (1.15–5.22)	0.016
TCGA cohort		
All stages colon cancer	2.91 (1.12–7.53)	0.021
Stage II colon cancer	1.20 (1.03–3.07)	0.964
Stage III colon cancer	4.78 (1.33–17.16)	0.008

Univariate Cox regression analysis was performed with the five-gene signatures to segregate colon cancer patients from the ZJU cohort and TCGA cohort into mutant and wild-type groups. *p* values were calculated by Wald test for each patient group related to disease-free survival.
Abbreviation: CI, confidence interval.

Table 2. The demographics of Stage III colon cancer patients

	ZJU ($n = 57$)	TCGA ($n = 55$)	*p*
Age, years, mean ± SD	60.65 ± 13.29	64.47 ± 13.32	0.131
Median follow-up, years (min–max)	6.77 (0.50–12.11)	2.47 (0.41–12.51)	0.005
Sex, female, *n* (%)	24 (42.1)	28 (50.9)	0.352
Stage			0.689
Stage IIIA, *n* (%)	1 (1.8)	3 (5.5)	
Stage IIIB, *n* (%)	41 (71.9)	31 (56.4)	
Stage IIIC, *n* (%)	15 (26.3)	17 (30.9)	
Location			0.082
Right sided colon, *n* (%)	28 (49.1)	36 (65.5)	
Left sided colon, *n* (%)	29 (50.9)	19 (34.5)	

Four patients in the TCGA cohort lacked information to determine the substaging for Stage III. Stage was assessed by the 7th version of the American Joint Commission on Cancer guidelines. The left-sided colon consists of the descending colon and sigmoid colon, with the remainder being classified into the right-sided colon. *p* values were calculated by *t*-test and Mann–Whitney test.

Identified Stage III colon cancer as the optimum subgroup for the prognostication of the five-gene signature

Patients with colon and rectal tumors or patients in different stages are managed differently.[20] To verify the prognostic power of the five-gene signature in both tumor types and avoid potential confounding effects, we investigated the correlations between the five-gene signature and recurrence/survival in patients of colon or rectal cancer with different stages with R0 resection. The patients with Stage III colon cancer with a mutant five-gene signature exhibited a significantly increased relative risk of death in both cohorts (ZJU: HR = 2.45, 95% confidence interval [CI] = 1.15–5.22, p = 0.016; TCGA: HR = 4.78, 95% CI = 1.33–17.16, p = 0.008; Table 1, Figs. 2*e* and 2*f*). However, the five-gene signature was not associated with OS in the patients with Stage II colon cancer (ZJU: p = 0.264; TCGA: p = 0.964; Table 1) or patients with rectal cancer (ZJU: p = 0.163; TCGA: p = 0.130). The demographics of patients with Stage III colon cancer are presented in Table 2. In this subgroup, there were 47.4% (27/57) of patients with mutations in at least one gene of the five-gene signature in training set and 45.5% (25/55) in the testing set. There was no statistically significant difference in the proportion of patients receiving adjuvant chemotherapy between the mutant and wild-type groups (ZJU: p = 0.117; TCGA: p = 0.090). We performed multivariate

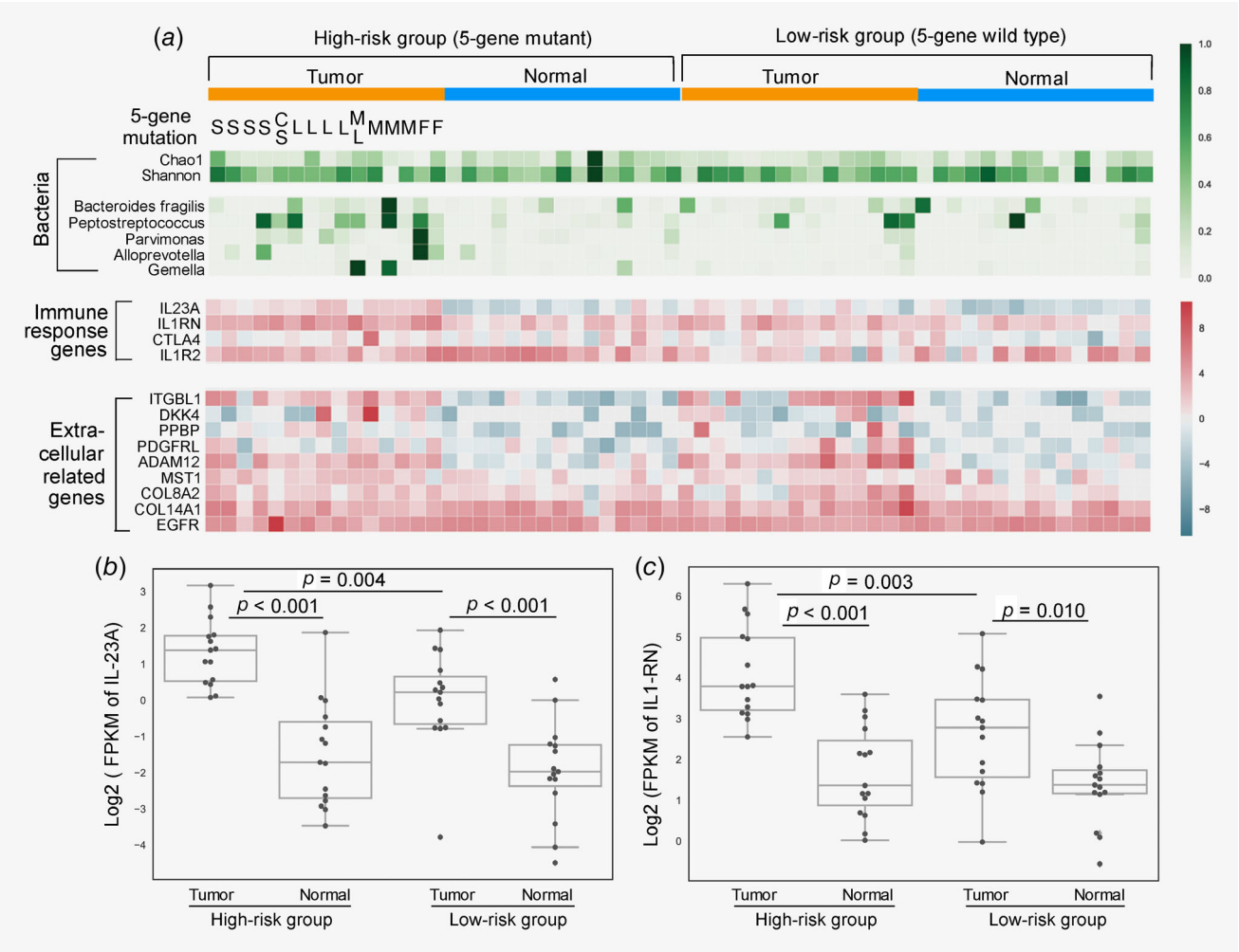

Figure 3. High-risk Stage III colon cancer patients are characterized by gut bacterial translocation and upregulation of IL23A and IL1RN of the tumor microenvironment. Tumor and matched normal mucosa of 30 cases of Stage III colon cancer were analyzed. Half of these cases with a mutant five-gene signature were classified as high-risk group, while the remaining cases without any mutation in the five-gene signature were classified as low-risk group. (*a*) Gene ontology enrichment analysis showed the extracellular related and immune response genes were significantly differential expressed between the risk groups. The overall richness (Chao 1) and diversity (Shannon) of bacterial community did not show differences between groups. *Bacteroides fragilis*, *Peptostreptococcus*, *Parvimonas*, *Alloprevotella and Gemella* are enriched in tumor samples with *COL6A3*, *LRP1B*, *MUC16* or *FLG* mutations. (*b*) IL23A was significantly upregulated in tumor samples of high-risk group compared to low-risk group. (*c*) IL1RN was significantly upregulated in tumor samples of high-risk group compared to low-risk group. Abbreviations: C, COL6A3 mutation; F, FLG mutation; FKPM, Fragments per Kilobase of transcript per Million fragments mapped; IL1RN, interleukin 1 receptor antagonist; IL-23A, interleukin 23 alpha; L, LRP1B mutation; M, MUC16 mutation; S, SMAD4 mutation. [Color figure can be viewed at wileyonlinelibrary.com]

analysis to determine independent predictors of overall survival in Stage III colon cancer patients. The five-gene signature was able to stratify Stage III colon cancer patients after adjusting age and sex (p = 0.015 and p = 0.006; Supporting Information Table S6). We observed similar results using DFS as the outcome measure in the ZJU cohort. The five-gene signature was associated with DFS in patients with Stage III colon cancer (HR = 2.39, 95% CI = 1.12–5.08, p = 0.020) but not in patients with Stage II colon cancer (p = 0.532; Supporting Information Table S7). Due to the lack of needed information, the correlation between the five-gene signature and DFS was not examined in the TCGA cohort.

High-risk Stage III colon cancer patients are characterized by gut bacterial translocation and upregulation of IL23A and IL1RN

To investigate the underlying causes of the increased risk, we performed RNA sequencing and bacterial 16S rRNA sequencing on 30 cases with Stage III colon cancer. Half of these cases with a mutant five-gene signature were classified as high-risk Stage III colon cancer, while the remaining cases as low-risk Stage III colon cancer (Fig. 3*a*). MSI-H, hypermutated or Stage IIIA colon cancer cases were not included in this analysis. There was no significant difference in sex (p = 0.845), age (p = 0.411), proportion of chemotherapy received (p = 0.729), of proportion of right-sided colon between groups (p = 0.686).

Forty-six downregulated and 39 upregulated genes were detected in patients of high-risk groups (Supporting Information Table S8). These differential expressed genes were significantly enriched in the extracellular related and immune response gene ontology (GO) terms (Supporting Information Table S9, Fig. 3*a*). *MUC16* and *FLG* are both known mechanical barrier genes.[21] *COL6A3* and *LRP1B* are involved in cell adhesion and tight junction disruption.[22] Significantly changes of extracellular related genes may be the consequence of loss function of these four genes. Previous study suggested that loss of barrier genes or tight junction may cause local loss of epithelial barrier function and translocation of gut commensal bacteria into the tumor microenvironment.[23] Using the same specimen with somatic mutations and expression profiles, we performed bacterial 16s rRNA sequencing. The overall richness and diversity of bacterial community did not show differences between the risk groups (the Chao 1 estimator: p = 0.657, the Shannon index: p = 0.749). Therefore, we selected cancer-related bacteria from previous studies.[24–26] As showed in Figure 3*a*, *Bacteroides fragilis*, *Peptostreptococcus, Parvimonas, Alloprevotella and Gemella* are enriched in tumor samples with *COL6A3*, *LRP1B*, *MUC16* or *FLG* mutations. The products of invaded bacteria activated tumor-associated myeloid cells and produced cytokine IL23. The p19 subunit of IL-23, *IL23A*, was significantly overexpressed in high-risk group (Fold change = 2.23, p = 0.004, Fig. 3*b*). We observed that *IL1RN* mRNA was also significantly upregulated compared to tumor samples of low-risk group (Fold change = 2.57, p = 0.003, Fig. 3*c*). The expression of the p40 subunit for IL-23, IL-12B, presents a low and similar manner in all samples. Expression levels of *IL23A* and *IL1RN* were verified by RT-qPCR analysis. In addition, we displayed chemotherapy resistance-related genes and *KRAS* 12/13 codons or *BRAF* V600E mutation status in Supporting Information Figure S2.

Discussion

In our study, we employed a clear strategy to establish a novel five-gene signature which associated with an increased risk of recurrence and death in patients with Stage III colon cancer. We also validated the prediction power of the signature with an independent data and ruled out the overfitting of the model by multiple statistical approaches. We examined not only the genetic change in tumor cells but also the modulatory alternation in TME factors including the gut bacterial translocation and the interleukin expression. Our results indicated that genetic changes of the five-gene signature may cause loss of intestinal barrier function and translocation of gut bacteria and affect the prognosis of Stage III colon cancer.

Studies investigating the CRC genome have revealed that very few mutations are shared between any two given primary CRCs. Whether a single genetic change accurately predicts the prognosis of patients, even using a driver gene as critical as KRAS, remains controversial.[27] Previous studies have constructed several prognostic signatures based on multiple genetic changes in a variety of cancers including CRC.[6,7,28] One of the advantages of these signatures is the ability to more broadly identify the high-risk patient through a combination of genetic events. Our five-gene signature also has this advantage to identify nearly half of high-risk patients with Stage III colon cancer.

Another advantage of our study is the determination of the applicable subgroup and the inapplicable subgroup of CRC patients. Our five-gene signature can identify patients with Stage III colon cancer at high risk of recurrence and death. However, this signature is not associated with the prognosis of patients with Stage II colon cancer or patients with rectal cancer. As shown in Figure 2*d*, both the positive and negative results were verified using an independent data. A staging and site-specific prognostic signature are more practical for current clinical applications.

In addition, although our study is retrospective, the results of the analysis of the association between five-gene signature and chemotherapy administration can support its application potential in guiding clinical treatment options. We observed that the patients with Stage III colon cancer with a mutant five-gene signature had equivalent OS regardless of whether they received adjuvant chemotherapy (p = 0.143, Supporting Information Fig. S3*a*). This finding suggests that these patients require a better adjuvant treatment option. Interestingly, the mutant five-gene signature could be a potential therapeutic target. For example, *MUC16* has been confirmed as a tumor neoantigen that can be targeted by T lymphocytes.[29] In contrast, patients with Stage III colon cancer with a wild-type five-gene signature should receive standard adjuvant chemotherapy based on either our results or current clinical

guidelines (HR = 0.26, 95% CI: 0.09–0.78, p = 0.010; Supporting Information Fig. S3*b*).

However, the limitations of our study are accompanied by advantages. A multigene signature has the advantage of covering more high-risk patients; however, the mechanism of the relationship between the risk signature and tumor progression is more difficult to elucidate. Our signature is based on the tumor mutation status of five genes (*SMAD4*, *MUC16*, *COL6A3*, *FLG* and *LRP1B*). *SMAD4* is an essential gene that regulates cell proliferation through the TGF-β signaling pathway.[30] *SMAD4* loss has been verified in identifying CRC patients at high risk of recurrence or death.[31] *SMAD4* loss alters bone morphogenic protein signaling to promote CRC cell metastasis.[30] *COL6A3* is a component of the extracellular matrix that remodels the extracellular matrix and contributes to cisplatin resistance in cancer cells.[32] *LRP1B* is a member of the LDL receptor family, is involved in focal adhesion, and inhibits cell proliferation and migration.[33] *MUC16*, which is also known as CA125, is a serum tumor biomarker and cell surface-associated mucin that is overexpressed in various cancers.[34] The *FLG* gene codes filaggrin, which is an intermediate-filament-associated protein that aggregates keratin intermediate filaments in the epidermis. The possible defect in the barrier function among *FLG* mutation carriers could result in a larger uptake through the skin or other tissues of foreign and harmful substances such as carcinogenic agents and less effective protection against bacteria and viruses (e.g., human papillomavirus, hepatitis B and C virus, Epstein–Barr virus, *Helicobacter pylori*) or suspected to be carcinogenic (e.g., *Salmonella typhi*, *Streptococcus bovis*, *Chlamydia pneumonia*).[35] *MUC16* and *FLG*, which are both known mechanical barrier genes, are related to pancreatic ductal adenocarcinoma and melanoma metastases.[21] *COL6A3* and *LRP1B* are involved in cancer metastasis through the regulation of cell adhesion.[32,33]

Previous studies have confirmed that the dysregulated localization of gut commensal bacteria plays a crucial role in CRC progression.[24,36,37] TME modulatory alterations may shape immune cell infiltration into tumor tissue and the infiltration status has been confirmed to be associated with recurrence and cancer-related death.[38,39] Our results revealed that mutations of *COL6A3*, *LRP1B*, *MUC16* and *FLG* genes may lead to loss of barrier function or tight junction, translocation of gut bacteria and deregulation of extracellular related and immune response genes. Significantly, the mutations were associated with upregulation of *IL23A* and *IL1RN*. *IL1RN* is an inhibitor of interleukin 1. For *IL23A*, a similar mechanism was previously described in *MUC2* deficient mice. Yang *et al.* observed that Muc2 deficient mice developing more tumors and Grivennikov *et al.* demonstrated that early transformation leads to a decreased Muc2 coverage and increased IL-23 dependent inflammation.[23,40] As showed in Supporting Information Figure S4, determining the unique mechanism explaining why the patients with a mutant five-gene signature had an increase risk is challenging.

As mentioned above, one of the advantages of our study is the inclusion of cases with different stages of colon and rectal cancer with clear applicability and inapplicability. However, in the case of Stage III colon cancer, the sample size was small. In addition to the small numbers of patients, another weakness of our study is the retrospective analysis of real-world data. As the validation dataset, the follow-up time of the TCGA data was significantly shorter than that of the ZJU data (Table 2 and Supporting Information Table S3). Future prospective studies are needed to determine whether the five-gene signature can be used clinically to guide decisions regarding adjuvant treatment for cancer patients.

In summary, the five-gene signature can identify patients with Stage III colon cancer who are at a high risk of recurrence and death. We established a risk-based stratification method based on the five-gene signature. Our study also suggested that increased risk was mainly caused by loss of barrier function and translocation of gut bacteria into the tumor microenvironment. Combining tumor genetic characteristics with dynamic tumor microenvironment changes may lead to more promising prognostic signatures, which could result in the more appropriate selection of cancer patients for systemic therapy after surgery.

Acknowledgements

We thank Dehao Wu, Jiaping Peng, Xiaoping Ding, Yi Zhang and Haomin Li for their technical support and American Journal Experts (AJE) for English language editing. This work was supported by funds from the National Key R&D Program of China (2017YFC0908200 to K.D.), the National Natural Science Foundation of China (U1804262 to L.W.), the National High Technology Research and Development Program of China (2012AA02A204 to W.G., 2012AA02A50 to L.C.) and the National Human Genetic Resources Sharing Service Platform (2005DKA21300 to W.G.).

References

1. Punt CJA, Koopman M, Vermeulen L. From tumour heterogeneity to advances in precision treatment of colorectal cancer. *Nat Rev Clin Oncol* 2017;14:235–46.
2. Bockelman C, Engelmann BE, Kaprio T, et al. Risk of recurrence in patients with colon cancer stage II and III: a systematic review and meta-analysis of recent literature. *Acta Oncol* 2015;54:5–16.
3. Grothey A, Sobrero AF, Shields AF, et al. Duration of adjuvant chemotherapy for stage III colon cancer. *N Engl J Med* 2018;378:1177–88.
4. Gao S, Tibiche C, Zou J, et al. Identification and construction of combinatory cancer Hallmark–based gene signature sets to predict recurrence and chemotherapy benefit in stage II colorectal cancer. *JAMA Oncol* 2016;2:37–45.
5. Walther A, Johnstone E, Swanton C, et al. Genetic prognostic and predictive markers in colorectal cancer. *Nat Rev Cancer* 2009;9:489–99.
6. Yu J, Wu WKK, Li X, et al. Novel recurrently mutated genes and a prognostic mutation signature in colorectal cancer. *Gut* 2015;64:636–45.
7. Sho S, Court CM, Winograd P, et al. A prognostic mutation panel for predicting cancer recurrence in stages II and III colorectal cancer. *J Surg Oncol* 2017;116:996–1004.
8. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. *Nat Med* 2015;21:1350–6.
9. Muzny DM, Bainbridge MN, Chang K, et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012;487:330–7.

10. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–303.
11. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 2013;31:213–9.
12. Kim S, Scheffler K, Halpern AL, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* 2018;15:591–4.
13. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38:e164.
14. Forbes SA, Beare D, Boutselakis H, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* 2017;45:D777–83.
15. Provenzale D, Gupta S, Ahnen DJ, et al. Genetic/familial high-risk assessment: colorectal version 1.2016, NCCN clinical practice guidelines in oncology. *J Natl Compr Canc Netw* 2016;14:1010–30.
16. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* 2013;10:996–8.
17. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;32:1792–7.
18. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol* 2017;18:220.
19. Campbell BB, Light N, Fabrizio D, et al. Comprehensive analysis of hypermutation in human cancer. *Cell* 2017;171:1042–56.
20. Tamas K, Walenkamp AM, de Vries EG, et al. Rectal and colon cancer: not just a different anatomic site. *Cancer Treat Rev* 2015;41:671–9.
21. Salerno EP, Bedognetti D, Mauldin IS, et al. Human melanomas and ovarian cancers overexpressing mechanical barrier molecule genes lack immune signatures and have increased patient mortality risk. *Oncoimmunology* 2016;5:e1240857.
22. Zhao Y, Li D, Zhao J, et al. The role of the low-density lipoprotein receptor-related protein 1 (LRP-1) in regulating blood-brain barrier integrity. *Rev Neurosci* 2016;27:623–34.
23. Grivennikov SI, Wang K, Mucida D, et al. Adenoma-linked barrier defects and microbial products drive IL-23/IL-17-mediated tumour growth. *Nature* 2012;491:254–8.
24. Kwong T, Wang X, Nakatsu G, et al. Association between Bacteremia from specific microbes and subsequent diagnosis of colorectal cancer. *Gastroenterology* 2018;155:383–90.
25. Wang X, Yang Y, Huycke MM. Microbiome-driven carcinogenesis in colorectal cancer: models and mechanisms. *Free Radical Biol Med* 2017;105:3–15.
26. Marchesi JR, Adams DH, Fava F, et al. The gut microbiota and host health: a new clinical frontier. *Gut* 2016;65:330–9.
27. Ogino S, Meyerhardt JA, Irahara N, et al. KRAS mutation in stage III colon cancer and clinical outcome following intergroup trial CALGB 89803. *Clin Cancer Res* 2009;15:7322–9.
28. Smith JJ, Deane NG, Wu F, et al. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology* 2010;138:958–68.
29. Balachandran VP, Luksza M, Zhao JN, et al. Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. *Nature* 2017;551:512–6.
30. Voorneveld PW, Kodach LL, Jacobs RJ, et al. Loss of SMAD4 alters BMP signaling to promote colorectal cancer cell metastasis via activation of rho and ROCK. *Gastroenterology* 2014;147:196–208.
31. Wasserman I, Lee LH, Ogino S, et al. SMAD4 loss in colorectal cancer patients correlates with recurrence, loss of immune infiltrate, and Chemoresistance. *Clin Cancer Res* 2019;25:1948–56.
32. Chen P, Cescon M, Bonaldo P. Collagen VI in cancer and its biological mechanisms. *Trends Mol Med* 2013;19:410–7.
33. Tabouret E, Labussière M, Alentorn A, et al. LRP1B deletion is associated with poor outcome for glioblastoma patients. *J Neurol Sci* 2015;358:440–3.
34. Streppel MM, Vincent A, Mukherjee R, et al. Mucin 16 (cancer antigen 125) expression in human tissues and cell lines and correlation with clinical outcome in adenocarcinomas of the pancreas, esophagus, stomach, and colon. *Hum Pathol* 2012;43:1755–63.
35. Skaaby T, Husemoen LLN, Thyssen JP, et al. Filaggrin loss-of-function mutations and incident cancer: a population-based study. *Br J Dermatol* 2014;171:1407–14.
36. Chin KF, Kallam R, O'Boyle C, et al. Bacterial translocation may influence the long-term survival in colorectal cancer patients. *Dis Colon Rectum* 2007;50:323–30.
37. Zhao L, Zhang X, Zuo T, et al. The composition of colonic commensal bacteria according to anatomical localization in colorectal cancer. *Engineering* 2017;3:90–7.
38. Wallace K, Lewin DN, Sun S, et al. Tumor-infiltrating lymphocytes and colorectal cancer survival in African American and Caucasian patients. *Cancer Epidemiol Biomarkers Prev* 2018;27:755–61.
39. Galon J, Costes A, Sanchez-Cabo F, et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* 2006;313:1960–4.
40. Popova NV, Yang K, Yang W, et al. Regulation of Hdac2 expression during intestinal tumorigenesis in $Muc2^{-/-}$ and $Muc2^{-/-}Apc^{1638/+}$ mice. *Cancer Res* 2005;65:880.