BMB
Reports

BMB Rep. 2026; 59(5): 273-282
www.bmbreports.org

Integrated analysis of RNA-sequencing data and clinical data for the molecular insights and applicability of immunotherapy in the Korean colorectal cancer patients

Jinseon Yoo[1,#], Jong Lyul Lee[2,#], Hyeran Shim[3,#], Soobok Joe[1], Jong-Hwan Kim[1], Jongbum Jeon[1], Jisu Kim[1], Chan Wook Kim[2], Seok-Byung Lim[2], In Ja Park[2], Yong Sik Yoon[2], Hoang Bao Khanh Chu[3], Jisun Kang[3], Sheehyun Cho[3], Hong Seok Lee[3], Young-Joon Kim[3,4,], Chang Sik Yu[2,*] & Seon-Young Kim[5,6,*]*

[1]Korea Bioinformation Center (KOBIC), Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon 34141, [2]Department of Surgery, Division of Colon and Rectal Surgery, University of Ulsan College of Medicine and Asan Medical Center, Seoul 05505, [3]Department of Biochemistry, College of Life Science and Biotechnology, Yonsei University, Seoul 03722, [4]LepiDyne Co., Ltd., Seoul 04779, [5]Personalized Genomic Medicine Research Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon 34141, [6]Department of Functional Genomics, University of Science and Technology (UST), Daejeon 34113, Korea

Colorectal cancer (CRC) is a major health concern and understanding its molecular characteristics is crucial for improving its diagnosis and treatment. Here, we present a comprehensive analysis utilizing RNA-sequencing (RNA-seq) data and clinical information from Korean patients with CRC. Differential gene expression analysis identified significant changes in gene expression between tumor and normal tissues. Gene Set Enrichment Analysis (GSEA) revealed dysregulated pathways associated with tumor progression. Furthermore, using CMScaller, we successfully stratified CRC tissues into distinct molecular subtypes. Upon reviewing the public consensus molecular subtype (CMS) signature, it was confirmed that it shares similar biological characteristics with the existing CRC. Additionally, biological characteristics of the group that could not be classified using CMScaller were found to resemble those of CMS2. Finally, distinguishing characteristics were observed between the tumor and normal groups when analyzed from an immunological perspective. Patients with CRC were checked for immunotherapy responsiveness, and those who clinically responded to immunotherapy were identified. Survival analysis confirmed that certain microsatellite stable (MSS) samples were responsive to immunotherapy and showed a relatively better prognosis. Furthermore, analysis of various immune cell types to identify genes involved in the response to immunotherapy revealed that RORC, NOS2, and KLRK1 are potential candidate genes. Our findings provide valuable insights into the molecular landscape of CRC in the Korean population and underscore the potential for integrating RNA-seq data with clinical information to improve cancer research and patient care. Immunotherapy was found to be effective in Korean patients with CRC. [BMB Reports 2026; 59(5): 273-282]

INTRODUCTION

As of 2022, colorectal cancer (CRC) ranked third in terms of the incidence of malignant tumors and second in terms of the cause of cancer-related deaths globally (1). According to the latest statistical data reported by the Ministry of Health and Welfare of Korea in 2024, CRC had the second highest crude and age-standardized incidence rates (2). The incidence rates are three to four times higher in developed countries than in developing or transitioning countries, and up to 2.5 million new cases are expected to be diagnosed worldwide by 2035 (3). However, in terms of cancer-related deaths, less variation is observed given the relatively higher case fatality rate in developing or transitioning countries. Although much progress has been made in the treatment of CRC, stage IV CRC still has a survival rate of less than 20%, which is much lower than that for stages I and II. Therefore, diagnosing CRC at an early stage and tailoring treatment for advanced stages offer crucial treatment opportunities. From this perspective, it is important to identify the molecular characteristics and apply the latest diagnostic tools and targeted treatments.

RNA-sequencing (RNA-seq) is a widely used analytical method based on gene expression level in the field of cancer research, and is commonly used to identify the characteristics of a cohort (4). The Cancer Genome Atlas Network (TCGA) consortium

*Corresponding authors. Young-Joon Kim, Tel: +82-2-2123-2628; Fax: +82-2-363-4083; E-mail: yjkim@yonsei.ac.kr; Chang Sik Yu, Tel: +82-2-3010-3494; Fax: +82-2-3010-6701; E-mail: csyu@amc.seoul.kr; Seon-Young Kim, Tel: +82-42-879-8500; Fax: +82-42-879-8519; E-mail: kimsy@kribb.re.kr
#These authors contributed equally to this work.

https://doi.org/10.5483/BMBRep.2024-0183

Received 18 November 2024, Revised 17 December 2024, Accepted 8 March 2025, Published online 31 May 2026

Keywords: Colorectal cancer, Gene expression, Immunotherapy, RNA-seq, Subtype

ISSN: 1976-670X (electronic edition)

performed a comprehensive analysis of CRC using RNA-seq and reported meaningful results (5). Advances in technology have made it possible to consider individual variability and move toward personalized medicine. Recently, the results of RNA-seq analysis in a Korean cohort were reported, and the RNA-seq data were analyzed according to the Consensus Molecular Subtype (CMS) classification (6). However, further analysis of the molecular and biological characteristics of the Korean CRC cohort is required in relation to diagnosis, treatment, and prognosis.

This study aimed to perform RNA-seq in a Korean cohort and analyze the results and clinical information. Additionally, we compared the molecular characteristics of CRC tumor with those of normal colon mucosa and resultingly, we aimed to deepen our understanding of the role of RNA-seq in the Korean CRC cohort and to confirm the applicability of immunotherapy.

RESULTS

Clinical characteristics of the study cohort and RNA-seq

To obtain the transcriptomic profile, RNA-seq data were generated from a cohort of 288 patients diagnosed with colorectal adenocarcinoma at the Asan Medical Center (AMC). The study cohort consisted of 249 tumor-only tissues, 70 tumor-normal matched tissues, and four normal-only tissues (Fig. 1A, Supplementary Table 1). The quality of the reads, as assessed using the Phred score, was predominantly > 30 (Supplementary Fig. 1A, Supplementary Table 2). Fastq files were aligned using STAR, and read counts were quantified using RSEM (Fig. 1B) (7, 8). Considering a previous report highlighting the strong influence of quality-based trimming on the estimation of gene and isoform expression levels (9), we performed an analysis without trimming using the default approach. Two samples with less than 70% uniquely aligned reads were excluded because they were deemed to have poor-quality data (Fig. 1C). The average mapped length distribution of reads across all samples was similar and the number of mapped reads exceeded 40 million, which indicated sufficient saturation for analysis (Supplementary Fig. 1B, Supplementary Fig. 1C, Supplementary Table 2). As a result, we preprocessed RNA-seq data from 323 tissues of 288 patients, and conducted quality control measures addressing both qualitative and quantitative aspects; this led to the exclusion of two samples.

This cohort comprised 193 and 128 retrospectively and prospectively collected tissues, respectively. The study cohort consisted of 126 female and 160 male patients, with a mean age of 60.5 ± 11.22 years. The follow-up period was 60.75 ± 30.96 months. The tumor was located in the right colon (cecal–splenic flexure of the transverse colon) in 86 patients (30.1%), left colon (splenic flexure of the transverse colon–distal sigmoid colon) in 91 patients (31.8%), and rectum in 109 patients (38.1%). The most frequently performed surgery was low anterior resection (111 patients, 38.8%), followed by right hemicolectomy (82 patients, 28.7%), and anterior resection (72 patients, 25.2%). The majority of the tumors were ulcerofungating: Bormann type II (225 patients, 78.7%), T3 (219 patients, 76.6%), N1b (87 patients, 30.4%), M0 (275 patients, 96.2%),

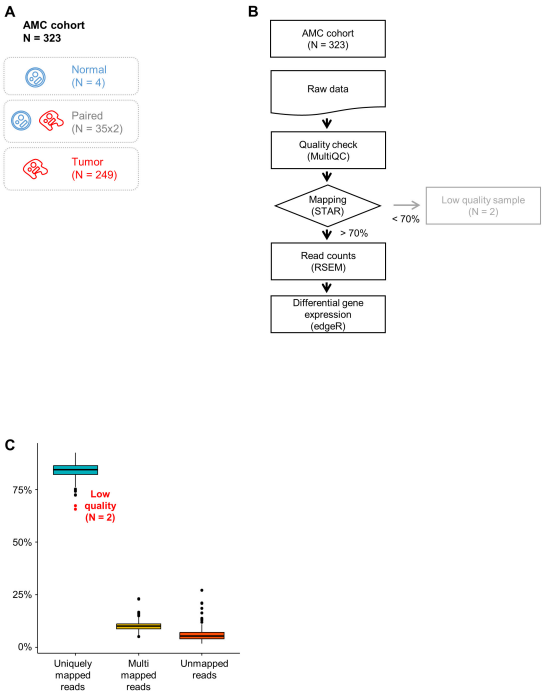

Fig. 1. Overview of AMC cohort and RNA-seq preprocessing. (A) The AMC cohort consists of 323 samples, including tumor-only, paired, and normal-only samples. The cohort comprises 249 tumor-only samples, 70 paired samples (35 tumor-normal pairs), and 4 normal-only samples. (B) A schematic flowchart illustrates the RNA-seq preprocessing pipeline. The pipeline includes the following steps: QC, Read Alignment, Sample Filtering, Normalization, and DEG Identification. (C) A boxplot summarizes the percentage of uniquely mapped reads, multi-mapped reads, and unmapped reads across all samples. The majority of samples used for downstream analysis show a high proportion of uniquely mapped reads, with the exception of two samples that had a uniquely mapped reads ratio below 70%.

stage IIIB (169 patients, 59.1%), and well- or moderately-differentiated (262 patients, 91.6%). Lymphovascular and perineural invasions were found in 169 (59.1%) and 105 patients (36.7%), respectively. Immunohistochemistry (IHC) results showed Microsatellite Instability (MSI)-high in 20 patients (7.0%), KRAS mutant in 24 patients (8.4%), NRAS mutant in 9 patients (3.1%), and BRAF mutant in 9 patients (3.1%). Two hundred forty-five patients (85.4%) received chemotherapy after surgery, and 31 patients (10.8%) received adjuvant radiotherapy. In the study cohort, 10 patients (3.5%) developed local recurrence, 91 (31.8%) developed systemic recurrence, and 6 (2.1%) developed combined local and systemic recurrences (Table 1). The distribution of the main variables used for the analysis, as well as those in the overall clinical data, was examined (Supplementary Fig. 2).

Analysis of differentially expressed genes (DEGs) between tumor and normal tissues in the AMC CRC cohort

The distribution of RNA-seq data from the AMC cohort was

Table 1. Clinical profile of the AMC CRC cohort

Variables	No. of patients (%) or values as mean ± standard deviation
	(n = 286)
Sex, female/ male	126/ 160 (44.1/ 55.9)
Age at operation, years	60.50 ± 11.22
Follow-up period, months	60.75 ± 30.96
Preoperative CEA level	
< 6.0 ng/dL	212 (74.1)
≥ 6.0 ng/dL	74 (25.9)
Tumor location	
Right (Cecum ~ SF of Transverse)	86 (30.1)
Left (SF of Transverse ~ Rectosigmoid)	91 (31.8)
Rectum	109 (38.1)
Operation name	
Right hemicolectomy	82 (28.7)
Left hemicolectomy	14 (4.9)
Anterior resection	72 (25.2)
Low anterior resection or ultra-low anterior resection	111 (38.8)
Abdominoperineal resection or Hartmann's operation	3 (1.0)
Total colectomy or Total proctocolectomy	4 (1.4)
Tumor size, cm	5.36 ± 1.84
Bormann type	
I	24 (8.4)
II	225 (78.7)
III	37 (12.9)
Harvested lymph node	
Number < 12	2 (0.7)
Number ≥ 12	284 (99.3)
T-category	
T1	1 (0.3)
T2	11 (3.8)
T3	219 (76.6)
T4a	50 (17.5)
T4b	5 (1.7)
N-category	
N0	46 (16.1)
N1a	72 (25.2)
N1b	87 (30.4)
N1c	11 (3.8)
N2a	39 (13.6)
N2b	31 (10.8)
M-category	
M0	275 (96.2)
M1a	7 (2.4)
M1b	2 (0.7)
M1c	2 (0.7)

assessed using a multidimensional scaling (MDS) plot which confirmed a distinct separation between tumor and normal tissues as well as a lack of batch effect between prospectively and retrospectively collected tissues (Fig. 2A). Subsequently, the DEGs between tumor and normal tissues were calculated (Fig. 2B, Supplementary Table 3). Gene Set Enrichment Analysis (GSEA) was performed using gene sets, such as KEGG pathways, biological processes, and GEO signatures, to compare tumor and normal tissues (Supplementary Table 4). Upregulated genes in tumor tissues were enriched in pathways associated

Table 1. Continued

Variables	No. of patients (%) or values as mean ± standard deviation
Pathologic stage	
I	2 (0.7)
IIA	39 (13.6)
IIB	3 (1.0)
IIC	1 (0.3)
IIIA	10 (3.5)
IIIB	169 (59.1)
IIIC	51 (17.8)
IVA	7 (2.4)
IVB	2 (0.7)
IVC	2 (0.7)
Differentiation	
WD/MD	262 (91.6)
PD/Mucinous	24 (8.4)
Lymphovascular invasion	
negative	117 (40.9)
positive	169 (59.1)
Perineural invasion	
negative	181 (63.3)
positive	105 (36.7)
Microsatellite instability	
MSS	238 (83.2)
MSI-L	12 (4.2)
MSI-H	20 (7.0)
Not done	16 (5.6)
KRAS	
Mutant	24 (8.4)
Wild	27 (9.4)
Not done	235 (82.2)
NRAS	
Mutant	9 (3.1)
Wild	32 (11.2)
Not done	245 (85.7)
BRAF	
Mutant	9 (3.1)
Wild	38 (13.3)
Not done	239 (83.6)
Adjuvant chemotherapy	
Yes	245 (85.7)
No	41 (14.3)
Adjuvant radiotherapy	15 (8.7)
Yes	31 (10.8)
No	255 (89.2)
Reccurrence	
No	179 (62.6)
Local recurrence	10 (3.5)
Systemic recurrence	91 (31.8)
Local and systemic recurrence	6 (2.1)

CEA, carcinoembryonic antigen; WD, well-differentiated; MD, moderately-differentiated; PD, poorly-differentiated; MSS, microsatellite instability stable; MSI-L, microsatellite instability low; MSI-H, microsatellite instability high; FL, 5-fluorouracil with leucovorin; FOLFOX, combination of 5-FU and oxaliplatin; FOLFIRI, combination of 5-FU and irinotecan.

with cell cycle, DNA replication, mismatch repair, and IL-17 expression, whereas downregulated genes were primarily involved in metabolism (Fig. 2C). IL-17 has been identified as a promoter of CRC tumorigenesis and a potential diagnostic marker of CRC (10). Defective DNA mismatch repair (dMMR) occurs in approximately 15% of CRCs patients and is associated with chemotherapy (11). Furthermore, when analyzed as a gene set using public data, colorectal adenocarcinomas were significantly enriched (Fig. 2D). The DEGs identified between tumor and normal tissues in our RNA-seq data exhibited transcriptomic characteristics comparable to those observed in other CRCs from public data.

CMS classification of RNA-seq data from AMC cohort

CMS classification was performed on RNA-seq data from the AMC cohort using the CMScaller (12). The subtypes are summarized using a heatmap and clinical data (Fig. 3A). Consistent with previous reports, MSI-H was most prevalent in CMS1, and the recurrence rate was higher in CMS4, which is known to have poor prognosis (Fig. 3B) (13). When survival analysis was performed between the CMS groups using clinical information, the CMS4 group showed the poorest prognosis after applying the Benjamini-Hochberg (BH) adjustment (Fig. 3C). DEGs were identified for each type, and the GSEA results showed that MSI and DNA repair were enriched in CMS1, MSS, and MYC in CMS2, differentiation related to the metabolic pathway in CMS3, and EMT in CMS4 (Fig. 3D, Supple-

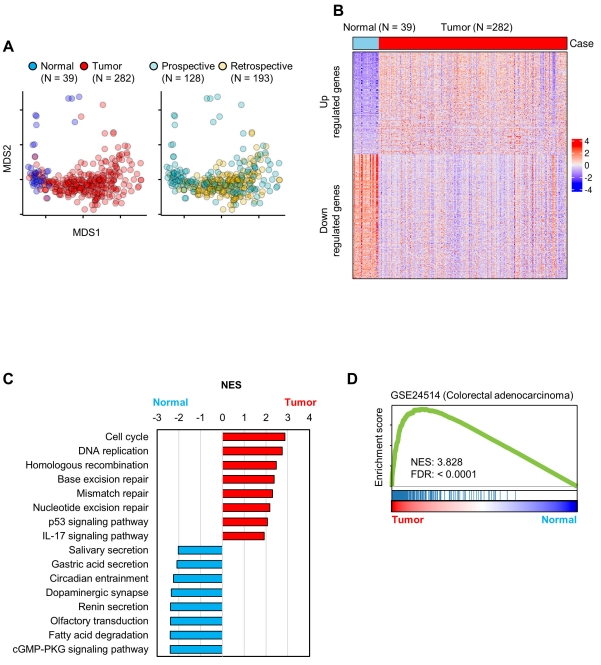

Fig. 2. Analysis of DEGs and gene sets in tumor and normal samples from the AMC cohort. (A) MDS plot showing the clustering of tumor and normal samples based on gene expression profiles. Tumor samples (red) and normal samples (blue) are distinctly separated. The prospective and retrospective samples are also color-coded to demonstrate their distribution. (B) Heatmap illustrating DEGs between tumor and normal samples. (C) KEGG pathway analysis of the DEGs. The bar chart shows the enriched pathways identified from the DEGs. The x-axis represents the normalized enrichment score (NES), while the y-axis lists the pathway names. (D) GSEA results highlighting RNA-seq GEO signatures. The NES and FDR values demonstrate the statistical significance of the results.

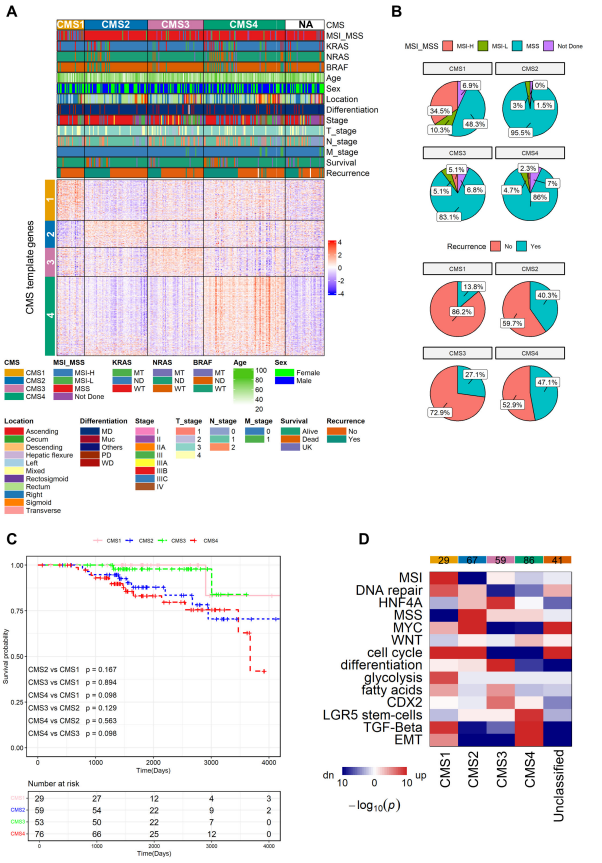

Fig. 3. CMS classification and analysis in the AMC cohort. (A) Heatmap of CMS classification for tumor samples (N = 282) from the AMC cohort, integrated with clinical and molecular data. Clinical parameters such as MSI status, stage, and other relevant information are annotated alongside the CMS classification. Missing data are indicated in white (NA values). (B) Pie charts illustrating the distribution of MSI types and recurrence status within the CMS groups. The CMS subgroups show the proportions of MSI-H and MSI-L/stable samples, as well as the presence or absence of recurrence. (C) Survival analysis across CMS groups. Kaplan-Meier (KM) survival curves demonstrate differences in overall survival among the CMS subgroups. The x-axis represents time (days), and the y-axis represents survival probability. The P-values shown are adjusted using the BH method. (D) Heatmap showing the enriched gene signatures and pathways associated with each group. The significance scores for different pathways are compared across each group, revealing distinct molecular signatures linked to each classification.

mentary Table 5), which is consistent with previously reported CMS results (14). Furthermore, gene expression-based analysis of the group not classified by CMScaller revealed biological characteristics similar to those of CMS2. This suggests that the unclassified group may not have been distinctly separated due to the lack of clear DEGs compared to CMS2 (Supplementary Fig. 3). Our analysis of the association between CMS and clinical information from the AMC cohort confirmed the findings of previous studies and helped verify the consistency of our data.

Immunological and purity analysis using tumor and normal RNA-seq data from the AMC cohort

First, we used the estimation of stromal and immune cells in malignant tumours using expression data (ESTIMATE) algorithm for immunological analysis and purity measurements in the AMC cohort, which is a tool designed to infer tumor purity and the presence of stromal and immune cells in tumor tissues using gene expression data. The results showed that the tumor group had lower immune, stromal, and ESTIMATE scores and higher tumor purity than the normal group ($P < 0.05$, Fig. 4A). Next, CIBERSORTx was used to measure and compare the immune cell infiltration fractions in the tumor and normal groups (Fig. 4B, C). Considering only the significant results, the tumor group had higher amounts of T cells CD4 naive, T cells CD4 memory activated, NK cells resting, Macrophages M0, Macrophages M1, Dendritic cells activated, Mast cells activated and Neutrophils. In contrast, other immune cell types were more prevalent in the normal group. Finally, when the expression levels of the three immune checkpoint inhibitors were examined in each group, CTLA4, PD1 (PDCD1), and CD80 were found to be high in the tumor group (Fig. 4D). These results show that there are differences in the immune microenvironment between the tumor and normal groups; for the tumor group, first-generation immunotherapy targeting CTLA4, as well as second-generation immunotherapy targeting PD-1, may be effective (15). Evidence suggests that increased CTLA4 and CD80 expression enhances their binding potential, increasing the applicability of CTLA4 as a target for first-generation immunotherapy strategies, while elevated PD-1 expression strengthens its interaction with PD-L1, contributing to T-cell exhaustion and making the PD-1/PD-L1 axis a key target for second-generation therapies designed to restore T-cell functionality and immune response.

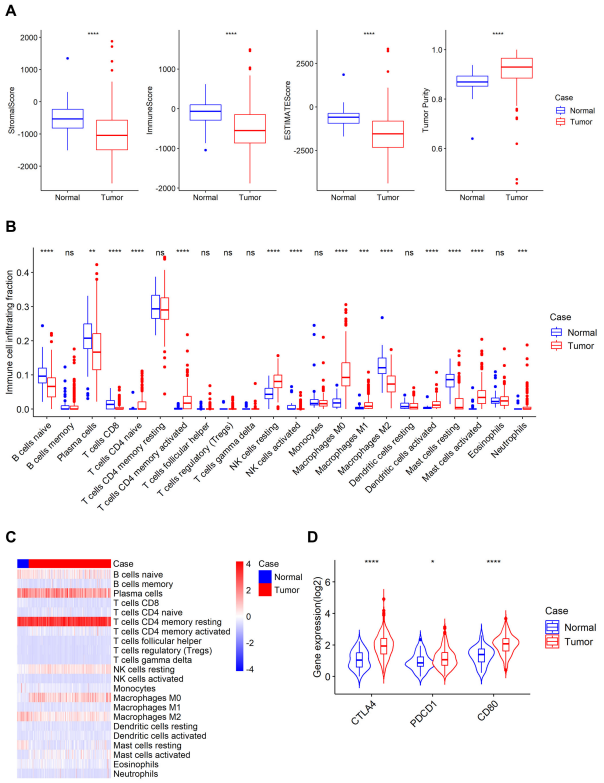

Fig. 4. Analysis of immunological aspects of the AMC cohort. (A) Boxplots showing stromal scores, immune scores, and tumor purity scores between tumor and normal groups of the AMC cohort using the ESTIMATE algorithm. Tumor groups exhibit significant differences in stromal, immune, and tumor purity scores compared to normal groups. (B) Boxplots comparing 22 immune cell type infiltration levels between tumor and normal groups using CIBERSORTx. Each boxplot represents the distribution of immune cell proportions across tumor and normal samples, highlighting significant differences in specific immune cell types. (C) Heatmap comparing the 22 immune cell type proportions between tumor and normal groups based on CIBERSORTx analysis. The heatmap illustrates the relative abundance of each immune cell type. (D) Violin plots showing the expression levels of three immune checkpoint inhibitors (e.g., PD-1, PD-L1, and CTLA-4) between tumor and normal groups. The plots indicate significant differences in expression levels, with statistical significance determined using the Wilcoxon test ($P < 0.05$). Notation: '*' represents statistically significant differences.

In silico analysis of immunotherapy response in the AMC cohort

This study additionally performed an analysis focusing on tumor samples because analysis of the tumor microenvironment (TME) helps identify predictors of immunotherapy response (16). The tumor immune dysfunction and exclusion (TIDE) algorithm was used to determine the response of tumor samples to cancer immunotherapy. TIDE is used to check for immune dysfunction and exclusion in tumor samples and predict the response to immunotherapy. The TIDE results showed that patients with AMC CRC could be divided into responders and nonresponders to immunotherapy (Fig. 5A). The characteristics of each group, classified as responders and nonresponders by TIDE, were examined from an immunological perspective (Supplementary Fig. 4). Additionally, the MSI status of CRC is a known biomarker for determining the applicability and responsiveness to immunotherapy, with the dMMR/MSI-H group reported to have a favorable prognosis in response to immunotherapy (17). In our survival analysis based on clinical information regarding the MSI status of CRC

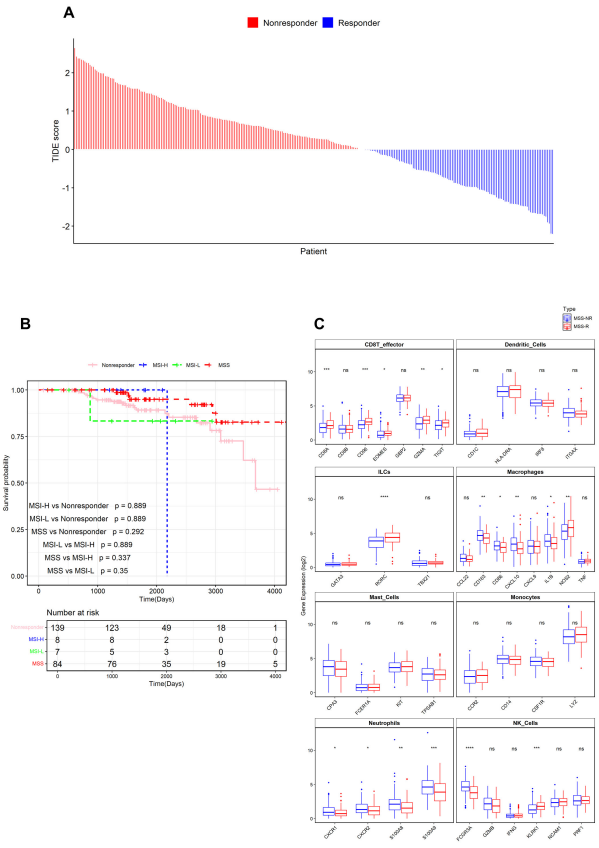

Fig. 5. *In silico* analysis of immunotherapy responsiveness in the AMC cohort. (A) Bar plot showing the classification of immunotherapy responsiveness in tumor samples from the AMC cohort based on the TIDE score. Tumor samples are categorized as responsive (blue) or non-responsive (red) to immunotherapy, with TIDE scores represented on the y-axis. (B) KM survival curves comparing overall survival between MSI subgroups (e.g., MSI-H and MSS) based on immunotherapy responsiveness. The x-axis represents time (days), and the y-axis shows survival probability. The P-values shown are adjusted using the BH method. (C) Boxplots comparing the expression levels of the CD8 T effector gene signature as well as genes associated with various innate cell types between MSS-R (responsive) and MSS-NR (non-responsive) groups. Statistical significance is determined, highlighting the differential expression of CD8 T effector-related genes and innate immune cell-associated genes.

tissues from AMC, the MSS-Responder group exhibited a better prognosis after applying the BH adjustment (Fig. 5B). Before adjustment, the significance level for the MSS vs. Nonresponder comparison was 0.049, which decreased to 0.292 after adjustment. Furthermore, when assessing the expression levels of various genes related to CD8 T effector cells and multiple innate cell types, we observed that the MSS-Responder group exhibited higher expression levels of several CD8 T effector genes as well as innate cell type-related genes such as RORC, NOS2, and KLRK1 compared to the MSS-Nonresponder group. Additionally, signatures utilized in the TIDE algorithm, including MSI Expr Sig, Merck18, MDSC, and CAF, also showed potential for predicting responders, further supporting the identification of immune checkpoint inhibitor responder groups among AMC CRC patients. These findings suggest that these genes and signatures could serve as potential candidate markers for identifying responders. Thus, these results highlight the clinical potential of applying immune checkpoint therapies to patients within this group (Fig. 5C). These results align with those of previous reports and indicate the potential applicability of immunotherapy and immunotherapy-based combination treatments for a specific subset of CRC tissues from AMC (18-20).

DISCUSSION

We analyzed RNA-seq data from 321 CRC tissues, thereby making them accessible for future research purposes. The provided dataset consisted of read counts and gene expression matrices along with the corresponding clinical information. Furthermore, we employed the CMScaller to assign CMS information to each tumor sample. Filtering steps were applied based on the mapping ratio and the quantity of mapped reads, which could be adjusted according to user preferences. Both read counts and normalized gene expression data were available in the dataset. In this study, we identified the biological characteristics and compared the tumor and normal groups. Differential expression and gene set analyses were performed based on gene expression levels, and it was confirmed that pathways related to DNA repair, p53, and IL-17 signaling were enriched in the tumor group. We used the reported CRC-related data to confirm that it was similar to existing CRC characteristics, which was analyzed using TCGA-COADREAD public data. CMS was confirmed and analyzed using clinical information and *in silico* methods, incorporating TCGA-COADREAD data for validation (Supplementary Fig. 5). These results suggest that globally accepted treatments can be applied to Korean patients with CRC. Differences in the microenvironment composition between the tumor and normal groups were confirmed through analysis of the immune microenvironment, and the possibility of applying immunotherapy to tumors was identified by examining the expression levels of specific markers. Finally, we investigated the possibility that some patients with MSS-CRC may respond to immunotherapy. Despite promising findings, several limitations must be noted. First, while *in-silico* analyses offer valuable insights, they inherently lack the biological validation that *in vitro* or *in vivo* studies can provide. Thus, further experimental studies are necessary to confirm the potential immunotherapy responsiveness of these patient groups. Notably, previous studies have validated the accuracy and reproducibility of CIBERSORT in analyzing tumor-infiltrating immune cell (TIIC) subsets using immunohistochemistry (IHC) data and multiple independent datasets, including TCGA CRC and GSE39582. In our study, CIBERSORTx was applied to analyze immune-infiltrating cells in AMC Korean CRC samples,

leveraging its proven reliability across diverse data sources (21). Lastly, although the findings contribute foundational data for MSS-CRC treatment approaches, the complexity of immune response mechanisms in MSS-CRC warrants more extensive studies, possibly involving multi-omics approaches, to comprehensively understand predictive markers and therapeutic implications. In conclusion, we performed a transcriptome analysis on 321 Korean CRC and normal tissues from AMC and confirmed that the results were similar to the previously reported transcriptome characteristics of CRC. Additionally, specific CRC patient groups that were expected to show immunotherapeutic effects were investigated. Candidate genes associated with the responsiveness of these groups were identified, highlighting the clinical significance of this analysis in suggesting potential therapeutic applications. These results can serve as a basis for future treatment of this patient group.

MATERIALS AND METHODS

Patients and tissue samples

Our study, using CRC and normal colonic mucosa samples, involved retrospective and prospective samples from the AMC, Seoul, Korea, between 2011 and 2019. The retrospective samples included 194 fresh frozen (FF) CRC samples and those biospecimen and data used in this study was provided by Asan Bio-Resource Center, Korea Biobank Network (2017-14 [152]). The prospective samples included 129 FF samples consisting of 38 normal colonic mucosa samples and 91 CRC samples. Normal mucosal samples were obtained at least 5 cm from CRC samples. Patients giving prospective sample provided written informed consent to participate in this study, and the study protocol was approved by the institutional review board of AMC (registration number: 2017-1350), and conducted in accordance with the Declaration of Helsinki. Patients were enrolled when the CRCs were clinically stage II, III, or IV and the tumor was large enough that there was no problem in diagnosis, even if the tumor was excised by approximately 1 cm. Patients were excluded if they underwent preoperative chemoradiotherapy or neoadjuvant chemotherapy or if they had hereditary CRC or other malignancies within 5 years of surgery. The following variables that might affect prognosis were recorded: sex; age at surgery; tumor location; surgery type; preoperative carcinoembryonic antigen; tumor size; Bormann type; T-,N-, and M-stage; differentiation; number of retrieved lymph nodes (based on 12); lymphovascular invasion; perineural invasion; MSI; IHC of KRAS, NRAS, and BRAF; adjuvant chemotherapy or radiotherapy; survival status; recurrence status; and survival outcomes (5-year overall and disease-free survival).

RNA extraction and sequencing

The tumor and adjacent normal FF tissues obtained from patients with CRC were dissected into 20-40 mg sections and homogenized 3-4 times for 15 s at a frequency of 30 Hz using a Tissue Lyser II (Qiagen, Hilden, Germany). RNA was extracted using the RNeasy Mini Kit (Qiagen, Hilden, Germany) according to a standard animal tissue RNA extraction protocol. Total RNA concentration was determined using Quant-IT RiboGreen (R11490; Invitrogen, Waltham, MA, USA). To assess RNA integrity, the samples were run on a TapeStation RNA ScreenTape system (5067-5576, Agilent, Santa Clara, CA, USA). Only high-quality RNA preparations (RNA integrity > 7.0) were used for the library construction. Subsequently, the total RNA was subjected to rRNA depletion using the Ribo-Zero Gold rRNA Removal Kit (MRZG12324, Illumina, San Diego, CA, USA), followed by mixing with 2 μl of 100-fold diluted ERCC Mix2 solution from the ERCC RNA Spike-In Mix kit (4456740, Ambion, Austin, TX, USA). An RNA-seq library was prepared using a TruSeq RNA Sample Prep Kit (Illumina). The libraries were subsequently subjected to paired-end sequencing (2×100 bp read length) on an Illumina HiSeq2000 platform (Illumina) at Macrogen Inc. (Gangnam-gu, Seoul, Republic of Korea).

RNA-seq analysis

The quality of FASTQ files was analyzed using MultiQC (v.1.15) (22). The human reference genome was obtained from the NCBI genome database (https://www.ncbi.nlm.nih.gov/genome/), and genome indexing was performed using STAR (v.2.7.8a). Sequenced reads were mapped to the human genome (hg38) using STAR, and gene expression levels were quantified using the count module in RSEM (v.1.3.3). The edgeR (v.3.32.2) (23) package was used to select DEGs from the RNA-seq count data between conditions (fold change > 2, FDR < 0.05). Meanwhile, the trimmed mean of M-values normalization (TMM)-normalized CPM (counts per million) value of each gene was added to 1 and log2-transformed for further analysis. A heatmap was generated using R (v.4.0.2, https://www.r-project.org/) ComplexHeatmap package (v.2.6.2) (24).

GSEA

GSEA was performed to compare tumor and normal samples using the GSEApy tool (v.0.10.1, https://github.com/zqfang/GSEApy) (25) with categories of KEGG pathways, biological processes, WikiPathways, Jensen DISEASES, Disease Perturbations from GEO, and RNA-seq GEO signatures in the Enrichr library (https://maayanlab.cloud/Enrichr/) (26). A cut-off threshold of $P < 0.05$ was used to obtain the significantly enriched gene sets.

Subtyping AMC tumor samples using CMScaller

Using the RNA-seq data of the tumor samples ($N = 282$) from the AMC patient group, subtype classification based on CMS was performed using CMScaller. Quantile normalization and log2 transformation were performed on the RSEM values as inputs using the RNAseq = TRUE parameter of the CMScaller module. The CMS template genes ($N = 458$) were derived from the templates in the CMS file of the CMScaller module, and a heatmap was generated by representing the log2CPM values for each template gene after z-normalization.

Quantification of immune cell infiltrating fraction, immune/stromal scores, and tumor purity

CIBERSORTx was used to calculate 22 immune cell-infiltrating fractions in the AMC cohort (27). Afterwards, only samples with a P-value < 0.05 were used for analysis. The ESTIMATE algorithm was used to calculate the immune, stromal, and ESTIMATE scores, and the formula provided in the paper was used to calculate the purity (28).

Responsiveness prediction analysis for immunotherapy

The TIDE algorithm (29) was used to predict the responsiveness to immunotherapy. A low TIDE score indicates high responsiveness to cancer immunotherapy, and a high TIDE score indicates a low response by escaping immunity.

Statistical analysis

For comparisons between two groups, the Wilcoxon rank-sum test was performed. Survival analysis was adjusted using the BH correction to control for false discovery rates (FDR). To account for multiple testing and reduce the likelihood of false-positive findings, the BH correction was applied to all survival analyses by ranking the raw P-values in ascending order and adjusting them based on the total number of comparisons. This ensures a balance between minimizing Type I errors and maintaining statistical power. All statistical analyses were conducted using R software version 4.0.2.

ACKNOWLEDGEMENTS

This research was supported by the KRIBB Initiative of the Korea Research Council of Fundamental Science and Technology and the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Ministry of Science & ICT (grant numbers NRF-2017M3A9A7050614, 2017M3A9A7050610, and 2022R1F1A1074317 to J.L. Lee).

We would like to thank Editage (www.editage.co.kr) for editing and reviewing this manuscript for English language.

CONFLICTS OF INTEREST

The authors have no conflicting interests.

DATA AVAILABILITY

RNA-seq data were deposited in the Korean Nucleotide Archive (KoNA) under the accession number KAP230607. Preprocessed read counts and normalized gene expression data were deposited in Zenodo (https://zenodo.org/records/8170863).

REFERENCES

1. Bray F, Laversanne M, Sung H et al (2024) Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 74, 229-263
2. National Cancer Center (2023) Annual report of cancer statistics in Korea in 2021. Available from: https://www.mohw.go.kr/board.es?mid=a10107010100&bid=0043&act=view&list_no=1481909&tag=&cg_code=&list_depth=1
3. Dekker E, Tanis PJ, Vleugels JLA, Kasi PM and Wallace MB (2019) Colorectal cancer. Lancet 394, 1467-1480
4. Del Vecchio F, Mastroiaco V, Di Marco A et al (2017) Next-generation sequencing: recent applications to the analysis of colorectal cancer. J Transl Med 15, 246
5. Muzny DM, Bainbridge MN, Chang K et al (2012) Comprehensive molecular characterization of human colon and rectal cancer. Nature 487, 330-337
6. Lee J, Kim JH, Chu HBK et al (2024) Comprehensive RNA-sequencing analysis of colorectal cancer in a Korean cohort. Mol Cells 47, 100033
7. Dobin A, Davis CA, Schlesinger F et al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21
8. Li B and Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323
9. Williams CR, Baccarella A, Parrish JZ and Kim CC (2016) Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC Bioinformatics 17, 103
10. Razi S, Baradaran Noveiry B, Keshavarz-Fathi M and Rezaei N (2019) IL-17 and colorectal cancer: from carcinogenesis to treatment. Cytokine 116, 7-12
11. Sinicrope FA (2010) DNA mismatch repair and adjuvant chemotherapy in sporadic colon cancer. Nat Rev Clin Oncol 7, 174-177
12. Eide PW, Bruun J, Lothe RA and Sveen A (2017) CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. Sci Rep 7, 16618
13. Guinney J, Dienstmann R, Wang X et al (2015) The consensus molecular subtypes of colorectal cancer. Nat Med 21, 1350-1356
14. Thanki K, Nicholls ME, Gajjar A et al Consensus molecular subtypes of colorectal cancer and their clinical implications. Int Biol Biomed J 3, 105-111
15. Sun Q, Hong Z, Zhang C, Wang L, Han Z and Ma D (2023) Immune checkpoint therapy for solid tumours: clinical dilemmas and future trends. Signal Transduct Target Ther 8, 320
16. Kavun A, Veselovsky E, Lebedeva A et al (2023) Microsatellite instability: a review of molecular epidemiology and implications for immune checkpoint inhibitor therapy. Cancers 15, 2288
17. Weng J, Li S, Zhu Z et al (2022) Exploring immunotherapy in colorectal cancer. J Hematol Oncol 15, 95
18. Thibaudin M, Limagne E, Hampe L, Ballot E, Truntzer C and Ghiringhelli F (2022) Targeting PD-L1 and TIGIT could restore intratumoral CD8 T cell function in human colorectal cancer. Cancer Immunol Immunother 71, 2549-2563
19. Kikuchi T, Mimura K, Okayama H et al (2019) A subset of patients with MSS/MSI-low-colorectal cancer showed increased CD8(+) TILs together with up-regulated IFN-γ. Oncol Lett 18, 5977-5985
20. Cai L, Chen A and Tang D (2024) A new strategy for immunotherapy of microsatellite-stable (MSS)-type advanced colo-

rectal cancer: multi-pathway combination therapy with PD-1/PD-L1 inhibitors. Immunology 173, 209-226
21. Zhou R, Zhang J, Wang L, Wang J, Yang X and Wang W (2018) Profiles of immune infiltration in colorectal cancer and their clinical significance: a gene expression-based study. Cancer Med 7, 4496-4508
22. Ewels P, Magnusson M, Lundin S and Kaller M (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinform 32, 3047-3048
23. Robinson MD, McCarthy DJ and Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140
24. Gu Z, Eils R and Schlesner M (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32, 2847-2849
25. Subramanian A, Tamayo P, Mootha VK et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102, 15545-15550
26. Kuleshov MV, Jones MR, Rouillard AD et al (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44, W90-W97
27. Newman AM, Steen CB, Liu CL et al (2019) Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol 37, 773-782
28. Yoshihara K, Shahmoradgoli M, Martínez E et al (2013) Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Comm 4, 2612
29. Jiang P, Gu S, Pan D et al (2018) Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. Nat Med 24, 1550-1558