Genomic Alterations Characterization in Colorectal Cancer Identifies a Prognostic and Metastasis Biomarker: FAM83A|IDO1

zaoqu liu
Zhengzhou University First Affiliated Hospital Department of Interventional Radiology
yuyuan zhang
Zhengzhou University First Affiliated Hospital Department of Interventional Radiology
qin dang
Zhengzhou University First Affiliated Hospital
kunpeng wu
Zhengzhou University First Affiliated Hospital Department of Interventional Radiology
Dechao Jiao
Zhengzhou University First Affiliated Hospital Department of Interventional Radiology
zhen li
Zhengzhou University First Affiliated Hospital Department of Interventional Radiology
Zhenqiang Sun
Zhengzhou University First Affiliated Hospital
xinwei han (fcchanxw@zzu.edu.cn)
Zhengzhou University First Affiliated Hospital https://orcid.org/0000-0003-4407-4864

Research

Keywords: colorectal cancer, genomic alteration, mutation signature, molecular subtype, prognosis, metastasis

Posted Date: November 19th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-108062/v1

Abstract

Background

Genomic alterations constitute crucial elements of colorectal cancer (CRC). Accumulating evidences have elucidated their clinical significance in predicting outcomes and therapeutic efficacy. However, a comprehensive understanding of CRC genomic alterations from a global perspective is lacking.

Methods

A total of 2778 patients in 15 public datasets were enrolled. Tissues and clinical information of 30 patients were also collected. Consensus clustering was performed for samples classification based on mutation signatures.

Results

We identified two distinct mutation signature clusters (MSC) featured by massive mutations and dominant somatic copy number alterations (SCNA) respectively. MSC-1 was associated with defective DNA mismatch repair, exhibiting more frequent mutations such as ATM, BRAF, and SMAD4. The mutational co-occurrences of BRAF-HMCN and DNAH17-MDN1 as well as the methylation silence event of MLH-1 were only found in MSC-1. MSC-2 was linked to the carcinogenic process of age and tobacco chewing habit, exhibiting dominant SCNA such as MYC (8q24.21) and PTEN (10q23.31) deletion as well as CCND3 (6p21.1) and ERBB2 (17q12) amplification. MSC-1 displayed higher immunogenicity and immune infiltration. MSC-2 had better prognosis and significant stromal activation. Based on the two subtypes, we identified and validated the expression relationship of FAM83A and IDO1 as a robust biomarker for prognosis and distant metastasis of CRC in 15 independent cohorts and qRT-PCR assay.

Conclusions

We identified two subtypes with heterogeneous molecular alterations and functional status, which might advance precise treatment and clinical management in CRC. A robust biomarker for predicting prognosis and distant metastasis of CRC was identified and validated.

Background

Colorectal cancer (CRC) remains the fourth most prevalent cancer and the second lethal cancer globally[1]. Although the considerable improvements in surgical techniques and chemoradiotherapy have extended overall survival for many patients, the mortality of CRC remains persistently high[2]. Recently, immunotherapy has made tremendous progress and achieved clinical success in CRC, but only a small proportion of patients benefit from this [3]. Hence, it is imperative to improve individualized treatment and clinical management in CRC.

For decades, the TNM and Dukes classification have been valuable for assessing the prognosis and treatment method of CRC patients[4]. However, accumulating evidences indicate that CRC patients with the same stage present diverse biological behaviors and clinical outcomes considering highly heterogeneity[5]. Thus, these conventional criterions fail to meet the needs of precision treatment in CRC. With the process of molecular biology, the CRC Subtyping Consortium proposed four consensus molecular subtypes (CMSs) in 2015[6]. The CMS classification can better help guide clinical treatment and evaluate prognosis. For example, CMS4 featured by epithelial-mesenchymal transition (EMT) and primary resistant to anti-EGFR therapy, has a poor prognosis relative to other subtypes; whereas CMS3 linked to metabolic reprogramming and altered cellular metabolism, displays favorable survival [6, 7]. Of note, the CMS classification only consider a fraction of genomic variations such as BRAF, TP53, KRAS mutations, HNF4A amplification, and homozygous deletion of PTEN, but there is a wide range of genomic variation in CRC, which can't fully explain the molecular heterogeneity of CRC and might ignore a large number of potential therapeutic targets and genomic drivers. Thus, it's necessary to systematically explore the heterogeneity of CRC based on the global genomic variations and further provide references for optimizing targeted treatment of CRC patients.

Currently, 30 mutation signatures have been summarized by previous research[8], which can be attributed to specific sources of DNA lesions, such as defective DNA repair and exogenous or endogenous mutagen exposures. Regrettably, the mutation signatures of CRC have not been dissected in detail until now. In addition, CRC possesses massive genomic variations [7], some of which play a vital role in predicting prognosis and guiding treatment. A previous study demonstrated that colon cancer patients harboring the same BRAF(V600E) oncogenic lesion generally displayed a low median survival[9]. A randomized, phase III trial indicated that patients with RAS wild type were sensitive to anti-EGFR therapy, conversely, patients with KRAS mutations display the resistance to anti-EGFR therapy [10]. Besides, CRC patients with LMNA-NTRK1 gene fusion were sensitive to TRKA kinase inhibitor Entrectinib [11]. In addition to these genomic variations that have been proven to be predominantly associated with prognosis and targeted therapies in CRC, there are still plenty of genomic variations that might have clinical significance waiting to be discovered.

In this research, we have systematically extracted 8 mutation signatures in CRC. Based on these mutation signatures, we performed consensus clustering, hoping to recognize heterogeneous molecular subtypes and better understand the genomic characteristics of CRC. Ultimately, we identified two distinct mutation signature clusters (MSC) featured by massive mutation load and dominant somatic copy number alterations (SCNA) respectively. MSC-1 was linked to mutation signatures 6, 15, and 20, indicating defective DNA mismatch repair; MSC-2 was linked to mutation signatures 1 and 29, related to the carcinogenic process of age and tobacco chewing habit. Notably, the two subtypes were also characterized by distinct genomic drivers (including mutation, SCNA, and methylation drivers), prognosis, functional status and immune microenvironment, as well as MSI status. In addition, based on the two subtypes, we identified and validated the expression relationship of FAM83A and IDO1 as a promising biomarker for prognosis and distant metastasis of CRC patients in 15 independent cohorts and qRT-PCR

assay. These results deepened the understanding of the heterogeneity of CRC, and facilitated the individualized treatment and clinical management of patients with CRC.

Methods

Data source and process

The mutation data (Varscan 2), somatic copy number alteration (SCNA) data, HumanMethylation450 array, and RNA-seq count data of CRC were obtained from TCGA portal. We also retrieved a total of 14 expression microarrays containing GSE17536, GSE17537, GSE103479, GSE29621, GSE38832, GSE39084, GSE39852, GSE71187, GSE72970, GSE87211, GSE27854, GSE21510, GSE18105, and GSE71222 from GEO database. For 11 microarrays from Affymetrix Human Genome U133 Plus 2.0 Array, the “CEL” raw data were obtained and further processed via robust multiarray averaging algorithm implemented in the affy R package. For 3 other microarrays, we directly recruited the normalized matrix files. The corresponding clinical information was also downloaded and the details were listed in Table S1. Ultimately, a total of 2778 patients were collected, of which 2294 patients had the survival information and 1144 patients had the metastasis information.

Deciphering mutational signatures in CRC

Masked somatic mutational profiles of 535 CRC patients were obtained from TCGA datasets. The trinucleotideMatrix function of the maftools package[12] was performed to extract immediate 5' and 3' bases flanking the mutated site and then obtained a 96 × 535 mutation subtype frequency matrix. Subsequently, we applied NMF package to extract mutation signature, the optimal rank was determined by the cophenetic coefficient and the residuals sum of squares (RSS). De novo mutational signatures were then compared to curated signatures in COSMIC[13] using cosine similarity[14] (https://cancer.sanger.ac.uk/cosmic/signatures_v2). The APOBEC enrichment analysis described by Roberts et al.[15] was further performed by Maftools package.

Consensus clustering

Based on the extracted mutation signatures, consensus clustering was utilized to determine the membership of possible clusters within the CRC patients using the ConsensusClusterPlus R package[16]. We set as 80% item resampling, 100% gene resampling, a maximum evaluated k of nine. The optimal number of clusters was determined by cumulative distribution function (CDF) and proportion of ambiguous clustering (PAC)[17]. In addition, Nbclust [18] package which provided 26 indices for determining the number of clusters, was also used to assess the best clustering scheme.

Mutation and SCNA analysis

The MutSigCV 1.4 software[19] was employed to identify the significantly mutated genes (SMGs) for two MSC subtypes of CRC. Genes with q-value <0.05 and mutation frequency >10% were defined as mutation drivers. The Gistic2.0 software was employed to identify significantly altered segments. The fragments

with q <0.05 and alteration frequency >25% were considered as SCNA drivers. The load of loss or gain was quantified as the total number of all genes with SCNA at the focal and arm levels. The mutation and SCNA in mismatch repair genes including MLH1, MLH3, MSH2, MSH3, MSH4, MSH5, MSH6, PMS1, and PMS2 were also explored in two MSC subtypes.

Identification of methylation driven genes

We downloaded the HumanMethylation450 array for TCGA-CRC cohort, and applied the IlluminaHumanMethylation450kanno.ilmn12.hg19 package for annotations. In order to identify the methylation driven genes (MDGs) in CRC, we employed two method to dissect the methylation profile. One method was MethylMix, based on the beta distribution, which was designed to recognize gene expression that were significantly related to methylation events[20]. The other method was derived from Wheeler et al. study, identifying epigenetic silencing genes according to the absolute expression difference between the methylation and unmethylation groups[21]. The MDGs were ultimately determined by the intersection of the two methods. In addition, if one MDG had the dominantly difference in both the mRNA expression and DNA methylation profile between the two MSC subtypes ($p < 0.05$), we then labeled it as subtype-specific MDG (ssMDG).

Functional annotation and immune related indicators assessment

We performed gene set enrichment analysis (GSEA) between the two MSC subtypes, and the biological function with FDR <0.05 was significant. The 50 Hallmark pathways were also retrieved from Molecular Signature Database (MSigDB v7.1). Based on the Hallmark gene sets, we utilized the GSVA algorithm to transform the gene expression matrix into pathway enrichment score matrix. The limma R package was applied to further reveal the discrepancy pathways between the two MSC subtypes, and the threshold was set to FDR <0.05 as well as $|\log2FC| > 0.2$. The abundance of eight immune cells and two nonimmune cells populations was assessed via MCP-counter R package. In addition, we also calculated or recruited 17 immunogenicity indicators from previous research encompassing nonsilent mutation rate, MSI score, SNV neoantigens, Indel neoantigens, Cancer testis antigens (CTA) score, Aneuploidy score (AS), Intratumor heterogeneity (ITH), number or fraction of segments alteration, homologous repair deficiency (HRD), loss of heterozygosity (LOH), B cell receptor (BCR) or T cell receptor (TCR) diversity, and cytolytic activity (CYT) [22-24]. The antigen processing and presenting machinery score (APS) used to measure antigen presentation capacity was further calculated according to a previous report [25]. The details of immune related indicators please refer to Table S2. Moreover, multi-omics regulations of 75 immunomodulator molecules were further analyzed (Table S3), including somatic mutation, SCNA, and DNA methylation [22]. The FDR was obtained from Benjamini-Hochberg multiple correction.

Identify the reliable gene pair markers for prognosis and distant metastasis

We aimed at identifying the mRNA expression relationship of two genes with prognosis and distant metastasis significance to facilitate clinical management. To ensure the robust and stability of our results, the 11 independent CRC cohorts with complete prognostic information encompassing TCGA-CRC,

GSE17536, GSE17537, GSE103479, GSE29621, GSE38832, GSE39084, GSE39852, GSE71187, GSE72970 as well as GSE87211 were retrieved to develop the promising prognosis markers, and the 7 independent CRC cohorts with distant metastasis information including TCGA-CRC, GSE39084, GSE29621, GSE27854, GSE21510, GSE18105 as well as GSE71222 were utilized to further explore the metastasis predictive power of identified prognostic markers (Table S1). The pipeline was as follows: (1) According to the criterion: |log2FC| >1 and FDR <0.05, the edgeR package was applied to screen the differentially expression genes (DEGs) between the MSC subtypes in the TCGA-CRC cohort. (2) Based on the extracted subtype-specific DEGs, we converted the mRNA expression matrix into the two gene expression relationships matrix. For one gene pair A|B, if the expression of A was greater than that of B, then labeled “A>B”, conversely, “B>A”. If the expression of A was equal to B, discarded the sample. If the proportion of “A>B” or “B>A” was greater than 90% in corresponding cohort, discarded the gene pair. (3) Univariate Cox regression analysis was implemented to screen the gene pairs with significant prognostic value (FDR <0.05) in each cohort. If one gene pair with FDR <0.05 in more than five cohorts, then defined it as the consensus prognosis gene pair signature (CPGPS). (4) For each CPGPS, we further explored its metastasis predictive power in 7 independent cohorts with metastasis information.

Human CRC specimens

The human cancer tissues used in this study were approved by Ethnics Committee of The First Affiliated Hospital of Zhengzhou University in December 19, 2019, and the TRN is 2019-KW-423. A total of 30 paired CRC tissues and matched adjacent nontumor tissues were obtained from patients after receiving surgical resection at The First Affiliated Hospital of Zhengzhou University. None of patients received any preoperative chemotherapy or radiotherapy. Written informed consent was obtained from all patients. The inclusion criteria were as follows: no preoperative chemotherapy, radiotherapy, or targeted therapy; no other types of tumors; no autoimmune diseases. The specimens obtained during surgery were immediately snap frozen in liquid nitrogen and stored at -80°C until RNA extraction. Clinical staging of the specimens was based on NCCN (2019) guidelines. The details of patients please refer to Table S4.

RNA preparation and quantitative real-time PCR

Total RNA was isolated from CRC tissues, and paired adjacent nontumor tissues with RNAiso Plus reagent (Takara, Dalian, China) according to the manufacturer’s instructions. RNA quality was evaluated using a NanoDrop One C (Waltham, MA, USA), and RNA integrity was assessed using agarose gel electrophoresis. An aliquot of 1 μg of total RNA was reverse-transcribed into complementary DNA (cDNA) according to the manufacturer’s protocol using a High-capacity cDNA Reverse Transcription kit (TaKaRa BIO, Japan). Quantitative real-time PCR (qRT-PCR) was performed using SYBR Assay I Low ROX (Eurogentec, USA) and SYBR® Green PCR Master Mix (Yeason, Shanghai, China) to detect the expression. The data was normalized to the expression of GAPDH. The sequences of the primers were as follows:

GAPDH forward (5'- to 3'-): GGAGCGAGATCCCTCCAAAAT

GAPDH reverse (5'- to 3'-): GGCTGTTGTCATACTTCTCATGG

FAM83A forward (5'- to 3'-): CAGATCTCTGACAGTCACCTCAAG

FAM83A reverse (5'- to 3'-): CTGCCTGACTTGGCACAGTA

IDO1 forward (5'- to 3'-): ATATGCCACCAGCTCACAGG

IDO1 reverse (5'- to 3'-): AGCTTTCACACAGGCGTCAT

Statistical analysis

Correlations between two continuous variables were assessed via Spearman's correlation coefficients. The Fisher's exact test or Pearson's chi-squared test was applied to compare categorical variables. Continuous variables were compared between two groups through the Wilcoxon Rank-Sum test or T test. The Wilcoxon Signed Rank test was utilized to compare the genes expression difference between the paired CRC tissues and matched adjacent nontumor tissues in the qRT-PCR assay. The Kaplan-Meier and Cox regression analysis was performed by survival R package. All P values were two-side, with $p < 0.05$ as statistically significant. The whole data processing, statistical analysis, and plotting were conducted in R 3.6.4 software.

Results

Extraction of mutation signatures in CRC

A total of 192905 mutations were detected in 535 samples with a median of 91, including single nucleotide variants (SNVs) and small insertions and deletions (Indels). SNV was the main mutational type, in which C>T displayed the highest frequency followed by C>A and T>C. Among the top 10 most frequently mutated genes, APC possessed the highest number of delete mutations (236) and TTN possessed the highest number of missense mutations (646) (Fig S1). To gain insights into the potential mutation generation processes operative in patients with CRC, we decomposed the mutation signatures via NMF algorithm (Fig S2A). Subsequently, eight mutation signatures were extracted from the CRC genomic data and annotated them against the (COSMIC) signature nomenclature based on cosine similarity (Fig S2B). Therefore, the extracted mutation signatures were ultimately called as cosmic signature 1, 6, 10, 15, 20, 28, 29, and 30 (Fig 1A). The clocklike signature 1 is thought to be connected with age-related accumulation of spontaneous deamination of 5-methylcytosine. Signatures 6, 15, and 20 are all associated with defective DNA mismatch repair. Signature 10 featured by altered activity of the error-prone polymerase POLE is often found in six cancer types including CRC. Signature 29 exhibits transcriptional strand bias for C>A mutations due to tobacco chewing habit. Signatures 28 and 30 have been observed in a subset of stomach and breast cancers remaining unknown etiology.

Generation of the mutation signature relevant subtypes

Based on the mutation signatures deciphered in the CRC genome, consensus clustering analysis revealed the two subtypes was the optimal choice (Fig 1B). The CDF curve, PAC value, and Nbclust results further

confirmed the stable and reliable of the cluster results (Fig S2C-SE). We annotated the two subtypes as mutation signature cluster (MSC) 1 and 2 respectively. The Kaplan-Meier survival analysis suggested MSC-1 was significantly associated with favorable prognosis (Log-rank OS: p =0.005; DFS: p =0.070) (Fig 1C-D). Of note, MSC-1 (n=226) had dominant signature 6, 15, and 20, linked to defective DNA mismatch repair. MSC-2 (n=309) was characterized by signature 1 and 29, which was associated with age and tobacco chewing habit (Fig 1E; Fig S2F). We also observed the APOBEC signature enrichment score was significantly higher (p =0.003) in MSC1, which indicated MSC-1 had a higher occurrence of C>T transition in TpCpW motifs (Fig 1F). Previous study demonstrated the mutation of APOBEC family might contribute to high tumor mutation burden (TMB) [26]. Therefore, we further explored the mutation in the APOBEC family and found the rare mutation in the CRC patients. Of note, MSC-1 had the more mutation proportion relative to MSC-2, in line with the APOBEC enrichment score (p =0.003) (Fig 1G), which might give rise to the high mutation rate of MSC-1.

Somatic mutation landscape of two subtypes

The tumor mutation burden (TMB) in MSC-1 was significantly higher than MSC-2 (p <0.001) (Fig 3SA). The higher TMB may tend to occur in patience with defective DNA mismatch repair relative mutation signatures[7]. We further determined 28 mutation driven genes with MutSigCV q-value <0.05 and mutation frequency >10% in CRC (Table S5; Fig 2A). Out of these 28 genes, 18 genes have been reported in at least one CRC associated research, such as APC, TP53, KRAS, SYNE1, PIKSCA, and FBXW7 et.al. Besides, ten novel drivers were identified including DNAH11, USHA2, HMCN1, HYDIN, MDN1, DST, VPS13B, DNAH8, EYS, and NBEA. We also dissected the prognostic role of these genes. The mutation of EYS prolonged DFS, and the mutation of USH2A suggested an unfavorable OS (Fig 2B-C). In two MSC subtypes, 22 out of 28 drivers exhibited significant mutation differences (Fig 2D). Consistent with the high TMB in MSC-1, the mutation frequency of most divers was also superior in MSC-1, such as ATM, SOX-9, and PRKDC. Of note, APC and KRAS, the early mutation event of colon adenoma–carcinoma process[27], dominantly mutated in MSC-2, which implied that familial adenomatous polyposis (FAP) may be one of main causes of MSC-2. Further analyses revealed a predominant mutation co-occurrence relationship between KRAS and SYNE1, TP53 and SYNE1, as well as APC and USH2A et.al (Fig S3B). Interestingly, we found some specific co-occurrence phenomenon such as BRAF-HMCN and DNAH17-MDN1 that appeared only in MSC-1, which suggested the specific co-occurrences of BRAF-HMCN and DNAH17-MDN1 could be promisingly employed to distinguish different subtypes (BRAF-HMCN: p <0.001; DNAH17-MDN1: p <0.001) (Fig 2E-F). In addition, for the first time, we revealed the prognostic value of some co-occurrence relationship, the co-occurrence of APC-TP53 demonstrated a favorable DFS (Fig S3C), and the co-occurrence of APC-KRAS, KRAS-TP53, and KRAS-SYNE1 was significantly associated with worse DFS (Fig 2G-I). Furthermore, the literature had confirmed that CRC patients with a defected mismatched repair (MMR) system could lead to hypermutation and microsatellite instability (MSI)[28]. Hence, we investigated the mutation status of nine known MMR genes, and the results exhibited MSC-1 had most mutations of MRR genes (Fig S3D), and the cases with MMR genes mutation relatively high in MSC-1 (26% vs. 7%; p <0.001) (Fig 2J), which was in line with its specific mutation signatures such as signature 6, 15, and 20.

SCNA investigation of two subtypes

In arm level, both the gain and loss load were significantly higher in the MSC-2 relative to MSC-1 ($p < 0.05$). Although there was no statistical significance in the focal level load between the two subtypes, the slightly superior trends were also shown in MSC-2 (Fig S4A). Different from MSC-1 which was characterized by higher mutation load, MSC-2 might be dominant in the alteration of copy number. According to the GISTIC algorithm, we eventually identified 39 driven segments encompassing 14 amplification segments and 25 deletion segments (Table S6-S7; Fig S4B). We further compared the alteration frequency of 39 segments between the two subtypes, and found MSC-1 had generally low frequency compared to MSC-2, in accordance with the CNA load (Fig 3A). We also found a multitude of oncogenes and tumor suppressor genes in these driven segments, which might play an essential role in the tumorigenesis and progression of CRC, such as MYC (8q24.21), CCND3 (6p21.1), ERBB2 (17q12), PTEN (10q23.31), SMAD4 (18q21.2), and APC (5q22.2) et.al (Fig 3B). Although MSC-2 had generally frequent SCNA events of these genes, there were still high fraction amplification or deletion occurred in MSC-1, such as MYC, FTK3, MCC as well as NOTCH and TGF-beta pathway associated genes. Interestingly, we found oncogenes with only amplification and tumor suppressor genes with only deletion. Thus, the gene expression difference between gain and no-gain or loss and no-loss was further explored, and we found oncogenes with gain was more prone to overexpress such as ERBB2, MYC, and MLST8, and the expression of tumor suppressor genes with loss was predominantly lower than no-loss group such as APC, SMAD4, and PTEN ($p < 0.001$) (Fig 3C; Fig S4C). These results suggested CNA status played a master regulator role in the aberrant expression of oncogenes and tumor suppressor genes in CRC. Further survival analysis demonstrated the prognosis significance of these genes (Fig S4D-E). For the first time reported, the gain of MLST8 and MAP2K2 could prolong OS (Fig 3C; Fig S4F), the gain of CCND3 indicated the worse DFS (Fig 3D), as well as the loss of CTNN6, DKK1, APC, MCC and SMAD4 was also associated with unfavorable DFS (Fig S4F). Moreover, we also investigated the CNA of MMR genes, and found the fraction of patients with the MMR genes deletion was higher in MSC-2 relative to MSC-1 (62% vs. 53%; $p = 0.042$) (Fig 3F-G). It was nonnegligible that some MMR genes displayed high level of loss frequency such as MLH3, MSH4, MSH3, and MLH1, so that might diminish the expression of MMR genes and give rise to MSI of CRC.

Methylation driven genes

To identify methylation driven genes (MDGs) in CRC, the MethyMix package and the Wheeler criterion were employed. The MethyMix algorithm screened 608 genes that their expression was significantly related to methylation events, and the Wheeler criterion filtered out 147 epigenetic silencing genes (Table S8). Eventually, we determined a total of 69 MDGs by the intersection of the two methods. Further univariate Cox regression uncovered the prognosis significance of these MSGs (Table S9). The high methylation of TBX1, GREB1L, and CNNM1 was significantly associated with the unfavorable OS (Fig 4A-C). In addition, we defined the subtype-specific MDG (ssMDG), and 13 ssMDGs were significantly different in their expression and methylation between the two MSC subtypes (Fig S5A-S5B). In terms of these ssMDGs, we observed significant negatively correlation between the expression and methylation

level (Fig 4D). MSC-2 dominated in the hypermethylation of AQP5 and ZNF304 compared to MSC-1. To our knowledge, AQP5 was a potential epigenetic driver of tumor development[29]. The other 11 ssMDGs were specific for MSC-1, such as ADAM32, SLC35D3, and TMEM150C. Of note, the MLH1 was also a specific ssMDG of MSC-1. As illustrated, the methylation level of MLH-1 was dominantly higher relative to MSC-2, and the expression level was the opposite (Fig S5A-S5B). Previous report demonstrated the hypermethylation of MLH-1 was a potential mechanism contributing to the MSI of CRC[30]. We thus divided CRC patients into the methylation cases and unmethylation cases based on the threshold of beta =0.3, and found all methylation cases occurred in MSC-1 (22% vs. 0%; $p < 0.001$) (Fig 4E), which explained its specific MSI-associated mutation signatures to some extent.

Functional status, immune cell infiltration and immunogenicity assessment

We performed the biological process and KEGG pathway enrichment analysis through the GSEA approach. The MSC-1 subtype was tightly associated with immune related pathways such as adaptive immune response, antigen processing and presentation, response to interferon-gamma, and Th1 and Th2 cell differentiation (Fig 5A and 5C). The MSC-2 subtype was significantly enriched in reactive stroma related pathways such as epidermis or mesenchymal morphogenesis, mesenchymal cell proliferation, transform growth factor beta (TGF-β) signaling pathway, and Wnt signaling pathway (Fig 5B and 5D). Further GSVA Hallmark pathways assessment suggested the similar result to the above, also elucidated the MSC-1 was dominant in the immune activation such as canonical T cell excitation pathway: interferon-gamma, and the MSC-2 was master in the stromal activation such as TGF-β process (Fig 5E). In addition, we also evaluated the subpopulations difference of eight immune cells and two nonimmune cells between the two subtypes (Fig 5F). Consistently, the immune killing cells such as T cells, CD8+ T cells, cytotoxic lymphocytes, and nature killer cells were superior in MSC-1, and MSC-2 was characterized by higher fibroblasts. The leukocyte and stomal fraction data retrieved from Thorsson et al. study also demonstrated the dominant role of the two subtypes in immune activation and stromal activation, respectively (Fig 5G-H).

Furthermore, 17 indicators were applied to decode the immunogenicity features of the two subtypes (Table S2; Fig 5I). In line with the specific mutation signatures such as 6, 10, and 15 in MSC-1, the nonsilent mutation rate and MSI score were higher in MSC-1 (Fig 5J-K), meanwhile, SNV and Indels neoantigens were also more prone to occur in MSC-1 relative to MSC-2 (Fig 5L-M), but there was no significance in term of CTA score (Fig S6A). Conversely, the CNV-relevant indicators were slightly high in MSC-2 such as AS, ITH, number or fraction of segments alteration, HRD, and LOH, although most of them did not reach statistical significance (Fig 5N and Fig S6B-S6G). These results implied the immunogenicity of two subtypes might be derived from different genome alterations. In addition, the BCR/TCR diversity and CYT that may reflect a robust antitumor response and cytolytic activity were also higher in MSC-1 (Fig S6H-S6K; Fig 5O). Overall, although there was heterogeneity between the two subtypes in different aspects of immunogenicity, MSC-1 still displayed the stronger immunogenicity compared to MSC-2, which might arise from the predominant mutation pattern. The stronger immunogenicity further conferred the superior immune activation in MSC-1.

The expression and regulation of immune checkpoint molecules

We next explored the expression and regulation difference of 75 immune checkpoint molecules (ICMs) between the two MSC subtypes in multi-omics dimensions (Table S3). Obviously, the expression of ICMs was generally high in MSC-1 (Fig 6A; Fig S7A-S7B). For example, the MHC molecules displayed relatively low expression in the MHC-2 (Fig 6B). We further calculated the antigen processing and presenting machinery score (APS) via ssGSEA algorithm, and observed the MSC-2 also presented a lower APS (Fig 6C). That suggested antigen presentation capacity might be impaired. In line with the immune activation status, MSC-1 demonstrated higher expression of stimulatory ICBs such as CCL5, CD40, and ITGB2 (Fig S7A). Meanwhile, the inhibitory ICBs such as IDO1, PDCD1, CTLA4, and CD274 also predominantly expressed in MSC-1, which implied the overexpression of inhibitory ICMs might be the immune escape mechanism of MSC-1 (Fig S7B).

Furthermore, we integrated the mutation, SCNA, and methylation profile to decipher the regulation of ICMs. Notably, although the mutation of ICMs displayed generally rare frequency (Fig 6A), it still presented some effects on the expression of ICMs, for example, the mutation of HLA-B and ITGB2 displayed significant lower expression only in MSC-1, but a slightly high was observed in MSC-2 (Fig 6D; Fig S7C). On the contrary, the SCNA of ICMs demonstrated the relatively prevalent frequency (Fig 6A). CD40 had the highest amplification frequency, but there was no significant expression difference between gain and no-gain groups (Fig S7D). Consistent with the deletion status, the expression of CD276, ICOSLG, TNFRSF9, TNFRSF14, and TNRSF18 was relatively low in loss group compared to no-loss group (Fig 6E-G and Fig S7E-S7F). In addition, the hypermethylation of ICMs also played a critical regulation role in a number of ICMs such as HLA-B, CXCL10, and CD40, we observed their expression was significantly negative correlation with the methylation profile (HLA-B: $r = -0.51$; CXCL10: $r = -0.43$; CD40: $r = -0.46$; all $p < 0.001$) (Fig 6H-J).

Identify the reliable gene pair markers for prognosis and distant metastasis

A total of 108 differentially expression genes (DEGs) were screened between the two MSC subtypes (Fig 7A; Table S11). We further transformed the gene expression matrix into the two gene expression relationships matrix. Based on the pipeline to screen consensus prognosis gene pair signature (CPGPS), we eventually determined three gene pairs with dominant prognostic significance in at least five cohorts, which were FAM83A|IDO1, FABP4|KLK12, and FABP4|GBP5 (Fig 7B-C; Fig S7A). Of note, the gene pairs with a single relationship ratio >90% of cases in corresponding cohort would be expurgated. Ultimately, the FAM83A|IDO1 was removed in the GSE103479 and GSE87211 cohorts, the FABP4|KLK12 was absented in the GSE103479, GSE87211, GSE18105, GSE21510, GSE27854, and GSE71222 cohorts, and the FABP4|GBP5 was deleted in the TCGA-CRC, GSE103479, GSE72970, GSE87211, GSE18105, GSE21510, GSE27854, and GSE71222 cohorts. The expression relationships of FAM83A and IDO1 was significantly associated prognosis in 7/9 of cohorts (Fig 7B and Fig 7D-L), and it was a poor prognostic factor when FAM83A > IDO1 in terms of the mRNA expression level. Although there was no significance in GSE17537 and GSE72970, the FAM83A|high group still indicated the adverse prognosis, which might be

due to their relatively small sample sizes (Fig 7F and 7L). The gene pair FABP4|KLK12 was also a prognosis marker which exhibited significance in 6/9 of cohorts. It turned out FABP4 > KLK12 was predominantly associated with unfavorable prognosis (Fig 7C and Fig 7M-U). In addition, the patients with FABP4 > GBP5 were more prone to have a poor prognosis in 5/7 of cohorts (Fig S1A-S1H). Further multivariate Cox analysis revealed FAM83A|IDO1 was an independent prognosis factor in most cohorts (7/9) (Table S11). Conversely, the two gene pairs FABP4|KLK12 and FABP4|GBP5 did not perform well here.

We then dissected the predictive role in the CRC metastasis of the three gene pairs. Interestingly, the distant metastatic features of CRC were significantly distinct between FAM83A|high and IDO1|high groups in all cohorts (TCGA-CRC: 29% vs. 15%, p =0.007; GSE29621: 46% vs. 17%, p =0.031; GSE39084: 44% vs. 16%, p =0.028; GSE18105: 49% vs. 23%, p =0.021; GSE21510: 47% vs. 25%, p =0.026; GSE27854: 45% vs. 18%, p =0.006; GSE71222: 24% vs. 8%, p =0.012) (Fig 8A-G). Due to the overrepresented single relationship of gene pair, the FABP4|KLK12 and FABP4|GBP5 were retained in three and two cohorts with metastasis information, respectively. Although the statistical significance was not reached in most cohorts, the proportion of patients with metastasis varied between FABP4|high and KLK12|high groups or FABP4|high and GBP5|high groups. For example, the FABP4|high group had more metastasis ratio compared to KLK12|high group (TCGA-CRC: 24% vs. 12%, p =0.001; GSE29621: 57% vs. 25%, p = 0.173; GSE39084: 31% vs. 31%, p =1.000) (Fig 8H-J), and the dominant fraction of metastasis cases occurred in the FABP4|high group relative to GBP5|high group (GSE29621: 44% vs 23%, p =0.199; GSE39084: 44% vs. 27%, p =0.278) (Fig 8K-L). Therefore, the predictive performance of FABP4|KLK12 and FABP4|GBP5 was much weaker than that of FAM83A|IDO1. Taken together, the expression relationships of FAM83A and IDO1 was a very promising biomarker for prognosis and distant metastasis of CRC patients.

Verified the role of FAM83A|IDO1 in prognosis and metastasis using qRT-PCR

QRT-PCR assay was performed in 30 paired CRC tissues and matched adjacent nontumor tissues (Table S4). We observed FAM83A was overexpressed in tumors relative to adjacent nontumor tissues, and the expression of IDO1 was the opposite (p <0.001) (Fig 9A-B). The role of FAM83A|IDO1 in prognosis and metastasis was further explored in the qRT-PCR assay. The clinical outcome details (including survival status and metastasis status) of 30 CRC patients were shown in Fig 9C. There was no correlation between the expression of FAM83A and IDO1. In line with the previous results, when the expression of FAM83A was higher than IDO1, patients had worse OS and DFS (Log-rank p <0.001) (Fig 9D-E), as well as stronger tendency of distant metastasis (71% vs. 13%, p =0.007) (Fig 9F).

Discussion

Elegant efforts have demonstrated multifarious genomic alterations were critical to the prognosis and target therapy of CRC[31-35]. We sought to better delineate the molecular diversity of CRC via mutation signatures that reflect different mutational processes, such as spontaneous deamination of 5-methylcytosine (signature 1), defective DNA mismatch repair (signature 6, 15, and 20), recurrent POLE

somatic mutations (signature 10), and tobacco chewing habit (signature 29). Based on these signatures, we identified two heterogeneous subtypes, MSC-1 and MSC-2. The two subtypes exhibited tremendous differences of the genomic alterations encompassing mutation, SCNA, and DNA methylation. The distinct TME status and immune escape mechanisms reinforced their molecular variability. We also observed significant clinical difference between the subtypes in terms of OS and DFS. In addition, to facilitate clinical application, we constructed the gene pair pipeline to develop a prognosis and distant metastasis biomarker: FAM83A|IDO1, and further validated in 15 independent datasets and qRT-PCR assay.

MSC-1, a mutation dominant subtype, characterized by signature 6, 15, and 20, was linked to defective DNA mismatch repair. In line with this, MSC-1 harbored massive mutation drivers such as ATM, BRAF, and HMCN1, which played vital roles in the tumorigenesis and development of many cancers[36]. Previous studies indicated that ATM was involved in cell cycle regulation and DNA damage recognition and repair, which might increase cell resistance to cisplatin[37]. Approximately 10% of patients with metastasis CRC possessed BRAF v600 mutation, which was relative to a poor prognosis[38]. Interestingly, the specific co-occurrence phenomenon including BRAF-HMCN and DNAH17-MDN1 appeared only in MSC-1, which suggested these specific co-occurrences could be promisingly employed to distinguish different subtypes. Besides, a multitude of methylation drivers, such as ADAM32, MLH-1, and CTTNBP2, were significantly epigenetically silenced in MSC-1. Interestingly, the methylation silence event of MLH-1 only appeared in MSC-1, which had been reported to be attributed to oncogenesis in CRC by involving in the serrated neoplasia pathway[39]. Combined with BRAF mutation, we then speculated serrated neoplasia pathway might be an important process of tumorigenesis in MSC-1.

MSC-2, a CNA dominant subtype, characterized by signature 1 and 29, was relative to the mutagenesis of spontaneous deamination of 5-methylcytosine and tobacco chewing habit. MSC-2 displayed the loss of MYC (8q24.21), SMAD4 (18q21.2), and PTEN (10q23.31) as well as the gain of CCND3 (6p21.1) and ERBB2 (17q12). Oncogene ERBB2 has been assessed as amplification or overexpression in multiple cancers including colon cancers[40, 41]. As reported, ERBB2 amplification is an emerging therapeutic target and may also be a negative predictor for response to anti-EGFR therapy in CRC[42]. Another promising candidate is SMAD4, a tumor suppressor that is the central node in TGFβ signaling[43]. Researches have demonstrated that the loss of SMAD4 is associated with worse prognosis and predisposition to chemoresistance, such as 5-fluorouracil, leucovorin, and irinotecan[44]. Of note, although MSC-2 demonstrated less TMB relative to MSC-1, mutation drivers APC and KRAS predominantly exhibited high mutation frequency in MSC-2, which occurred early in the progression from colorectal adenoma to malignant carcinoma [39]. As reported, approximately 85% of CRC are thought to evolve from conventional adenomas accompanied with the mutations of APC, SMAD4, TP53, KRAS, and PI3KCA resulting in Wnt-β-catenin and TGF-β pathway activation; this process is referred to as the adenoma-to-carcinoma sequence. The above analysis inspired us that conventional adenoma-to-carcinoma sequence may be an important process of oncogenesis in MSC-2.

In the present study, we also assessed the difference of immune cells, stromal cells infiltration, and immune escape mechanisms between the two subtypes. Consistent with high mutation load, numerous

innate and adaptive immune cells infiltration characterized the TME in MSC-1, linked to the immune inflammation status. The TME status was also validated by the activation of immune-related pathways, including adaptive immune response, antigen processing and presentation of peptides antigens, and response to interferon-gamma. In CRC, immune checkpoint inhibitors have been proved effective in heavily mutated tumors with MMR defective or high levels of MSI[3], which imply that patients in MSC-1 may benefit more from immunotherapy. Although accompanied with both MSI and immune activation, MSC-1 exhibited unfavorable OS and RFS. High level of immunosuppressive molecules in TME may trigger the immune resistance and escape mechanisms in MSC-1. Compared with MSC-1, MSC-2 was characterized by higher fibroblasts and the lack of adaptive immune cells, accompanied with the stromal-associated pathways activation such as epidermis or mesenchymal morphogenesis, mesenchymal cell proliferation, TGF-β signaling pathway, and Wnt signaling pathway. Combined with the weaker immunogenicity, the insufficient immune cell infiltration in MSC-2 attributed to the extinctive immune escape, which given us a clue that patients MSC-2 subtype might exhibit unfavorable response to immunotherapy. Therefore, comprehensive analysis in molecular and immune microenvironment variability might contribute to the optimize treatment and clinical management of CRC patients.

In addition, we comprehensively revealed plenty of prognosis relevant genomic events. In this study, we observed the mutation of EYS, as well as the gain of MLST8 and MAP2K2 can prolong OS, while the mutation of USH2, the loss of DKK1, APC, MCC, and SMAD4, as well as the methylation of TBX1 were linked to unfavorable prognosis. In addition, the prognostic value of some co-occurrence relationship was revealed for the first time, the co-occurrence of APC-TP53 demonstrated a favorable DFS, and the co-occurrence of APC-KRAS, KRAS-TP53, and KRAS-SYNE1 was significantly associated with poor DFS. Importantly, to facilitate clinical application, we identified three gene pairs with prognosis significance, such as FAM83A|IDO1, FABP4|KLK12, and FABP4|GBP5. FAM83A|IDO1 had the better performance for predicting prognosis in 11 public datasets and our own cohort, and it was an independent prognosis factor for CRC. Meanwhile, FAM83A|IDO1 also exhibited excellent performance in assessing the distant metastasis status in 7 public datasets and our own cohort. Patients with FAM83A|high had a higher risk of metastasis than patients with IDO1|high. Traditionally, the batch effects of different platforms and the definition of cut-off values severely limit the clinical translation and application of previous biomarkers. In this study, we only focused on the mathematical relationship between mRNA expression of two genes, which completely ignores the batch effect among different platforms and does not need to define the cut-off value, it is just a binary size relationship. Therefore, the mRNA expression relationship of FAM83A and IDO1 was a promising biomarker for prognosis and metastasis in clinical application.

Our study also has a few limitations. Firstly, this study provided multi-dimensional references for genomic alterations in CRC, but lacked microscopic experimental verification. Secondly, owing to the lack of data, our study only considered the interpatient heterogeneity and did not take into account the intratumoral heterogeneity. Thirdly, the mRNA expression relationship of FAM83A and IDO1 focused on the mathematical relationship of two genes, but the biological relationship was unknown.

Conclusion

We described a novel molecular classification of CRC in two clusters, suggesting the intertumoral molecular variability. The two subtypes displayed distinct genomic drivers, prognosis, functional status, immune microenvironment, and MSI status, which might advance precise treatment and clinical management in CRC. Promisingly, we also identified and validated a robust and promising biomarker for predicting prognosis and metastasis of CRC patients.

Abbreviations

CRC: colorectal cancer; MSC: mutation signature cluster; SCNA: somatic copy number alterations; EMT: epithelial-mesenchymal transition; MMR: mismatched repair; MSI: microsatellite instability; ICMs: immune checkpoint molecules; qRT-PCR: quantitative real-time PCR; TMB: tumor mutation burden; FAP: familial adenomatous polyposis; GSEA: gene set enrichment analysis; CTA: cancer testis antigens; AS: aneuploidy score; ITH: intratumor heterogeneity; HRD: homologous repair deficiency; LOH: loss of heterozygosity; BCR: B cell receptor; TCR: T cell receptor; CYT: cytolytic activity; APS: antigen processing and presenting machinery score; SMGs: significantly mutated genes.

Declarations

Ethics approval and consent to participate

The human cancer tissues used in this study were approved by Ethnics Committee of The First Affiliated Hospital of Zhengzhou University in December 19, 2019, and the TRN is 2019-KW-423.

Consent for publication

Not applicable.

Availability of data and materials

Not applicable.

Competing interests

The authors declare no conflict of interest.

Data accessibility

Not applicable.

Funding

Project supported by the National Natural Science Foundation of China (Grant No. U1904143).

Authors' contributions

LZQ made conceptualization; LZQ, DQ and SZQ involved in methodology; SZQ, JDC, LZ and HXW provided the resources; LZQ and DQ analyzed the data; LZQ, ZYY and WKP prepared original draft, reviewed and edited the manuscript; WKP supervised the study.

Acknowledgements

Not applicable.

References

1. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA Cancer J Clin, 2018. **68**(6): p. 394-424.
2. Casado-Saenz, E., et al., *SEOM clinical guidelines for the treatment of advanced colorectal cancer 2013.* Clin Transl Oncol, 2013. **15**(12): p. 996-1003.
3. Le, D.T., et al., *Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade.* Science, 2017. **357**(6349): p. 409-413.
4. Piñeros, M., et al., *Essential TNM: a registry tool to reduce gaps in cancer staging information.* Lancet Oncol, 2019. **20**(2): p. e103-e111.
5. Zhang, B., et al., *Proteogenomic characterization of human colon and rectal cancer.* Nature, 2014. **513**(7518): p. 382-7.
6. Guinney, J., et al., *The consensus molecular subtypes of colorectal cancer.* Nat Med, 2015. **21**(11): p. 1350-6.
7. Dienstmann, R., et al., *Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer.* Nat Rev Cancer, 2017. **17**(2): p. 79-92.
8. Alexandrov, L.B., et al., *Clock-like mutational processes in human somatic cells.* Nat Genet, 2015. **47**(12): p. 1402-7.
9. Prahallad, A., et al., *Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR.* Nature, 2012. **483**(7387): p. 100-3.
10. Douillard, J.Y., et al., *Randomized, phase III trial of panitumumab with infusional fluorouracil, leucovorin, and oxaliplatin (FOLFOX4) versus FOLFOX4 alone as first-line treatment in patients with previously untreated metastatic colorectal cancer: the PRIME study.* J Clin Oncol, 2010. **28**(31): p. 4697-705.
11. Sartore-Bianchi, A., et al., *Sensitivity to Entrectinib Associated With a Novel LMNA-NTRK1 Gene Fusion in Metastatic Colorectal Cancer.* J Natl Cancer Inst, 2016. **108**(1).
12. Mayakonda, A., et al., *Maftools: efficient and comprehensive analysis of somatic variants in cancer.* Genome Res, 2018. **28**(11): p. 1747-1756.

13. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415-21.
14. Kandoth, C., et al., *Mutational landscape and significance across 12 major cancer types.* 2013. **502**(7471): p. 333-339.
15. Roberts, S.A., et al., *An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers.* Nat Genet, 2013. **45**(9): p. 970-6.
16. Wilkerson, M.D. and D.N. Hayes, *ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking.* Bioinformatics, 2010. **26**(12): p. 1572-3.
17. Şenbabaoğlu, Y., G. Michailidis, and J.Z.J.S.r. Li, *Critical limitations of consensus clustering in class discovery.* 2014. **4**(1): p. 1-13.
18. Malika, C., et al., *NbClust: an R package for determining the relevant number of clusters in a data Set.* 2014. **61**: p. 1-36.
19. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-218.
20. Cedoz, P.L., et al., *MethylMix 2.0: an R package for identifying DNA methylation genes.* Bioinformatics, 2018. **34**(17): p. 3044-3046.
21. Charoentong, P., et al., *Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade.* Cell Rep, 2017. **18**(1): p. 248-262.
22. Thorsson, V., et al., *The Immune Landscape of Cancer.* Immunity, 2018. **48**(4): p. 812-830 e14.
23. Aran, D., M. Sirota, and A.J. Butte, *Systematic pan-cancer analysis of tumour purity.* Nat Commun, 2015. **6**: p. 8971.
24. Rooney, M.S., et al., *Molecular and genetic properties of tumors associated with local immune cytolytic activity.* Cell, 2015. **160**(1-2): p. 48-61.
25. Wang, S., et al., *Antigen presentation and tumor immunogenicity in cancer immunotherapy response prediction.* Elife, 2019. **8**.
26. Middlebrooks, C.D., et al., *Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors.* Nat Genet, 2016. **48**(11): p. 1330-1338.
27. Vogelstein, B., et al., *Cancer genome landscapes.* Science, 2013. **339**(6127): p. 1546-58.
28. *Comprehensive molecular characterization of human colon and rectal cancer.* Nature, 2012. **487**(7407): p. 330-7.
29. Kiely, M., et al., *Age-related DNA methylation in paired normal and tumour breast tissue in Chinese breast cancer patients.* Epigenetics, 2020: p. 1-15.
30. Vilar, E. and S.B. Gruber, *Microsatellite instability in colorectal cancer-the stable evidence.* Nat Rev Clin Oncol, 2010. **7**(3): p. 153-62.
31. Van Cutsem, E., et al., *Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer.* N Engl J Med, 2009. **360**(14): p. 1408-17.

32. Liu, H., et al., *Copy number variations primed lncRNAs deregulation contribute to poor prognosis in colorectal cancer.* Aging (Albany NY), 2019. **11**(16): p. 6089-6108.
33. Ganesh, K., et al., *Immunotherapy in colorectal cancer: rationale, challenges and potential.* Nat Rev Gastroenterol Hepatol, 2019. **16**(6): p. 361-375.
34. Feng, Q., et al., *A specific KRAS codon 13 mutation is an independent predictor for colorectal cancer metachronous distant metastases.* Am J Cancer Res, 2015. **5**(2): p. 674-88.
35. Douillard, J.Y., et al., *Panitumumab-FOLFOX4 treatment and RAS mutations in colorectal cancer.* N Engl J Med, 2013. **369**(11): p. 1023-34.
36. Martínez-Jiménez, F., et al., *A compendium of mutational cancer driver genes.* Nat Rev Cancer, 2020. **20**(10): p. 555-572.
37. Manic, G., et al., *Trial Watch: Targeting ATM-CHK2 and ATR-CHK1 pathways for anticancer therapy.* Mol Cell Oncol, 2015. **2**(4): p. e1012976.
38. Kopetz, S., et al., *Encorafenib, Binimetinib, and Cetuximab in BRAF V600E-Mutated Colorectal Cancer.* N Engl J Med, 2019. **381**(17): p. 1632-1643.
39. Strum, W.B., *Colorectal Adenomas.* N Engl J Med, 2016. **374**(11): p. 1065-75.
40. Ursini-Siegel, J., et al., *Insights from transgenic mouse models of ERBB2-induced breast cancer.* Nat Rev Cancer, 2007. **7**(5): p. 389-97.
41. Wang, D.S., et al., *Liquid biopsies to track trastuzumab resistance in metastatic HER2-positive gastric cancer.* Gut, 2019. **68**(7): p. 1152-1161.
42. Bertotti, A., et al., *A molecularly annotated platform of patient-derived xenografts ("xenopatients") identifies HER2 as an effective therapeutic target in cetuximab-resistant colorectal cancer.* Cancer Discov, 2011. **1**(6): p. 508-23.
43. Heldin, C.H., K. Miyazono, and P. ten Dijke, *TGF-beta signalling from cell membrane to nucleus through SMAD proteins.* Nature, 1997. **390**(6659): p. 465-71.
44. Wasserman, I., et al., *SMAD4 Loss in Colorectal Cancer Patients Correlates with Recurrence, Loss of Immune Infiltrate, and Chemoresistance.* Clin Cancer Res, 2019. **25**(6): p. 1948-1956.

Figures

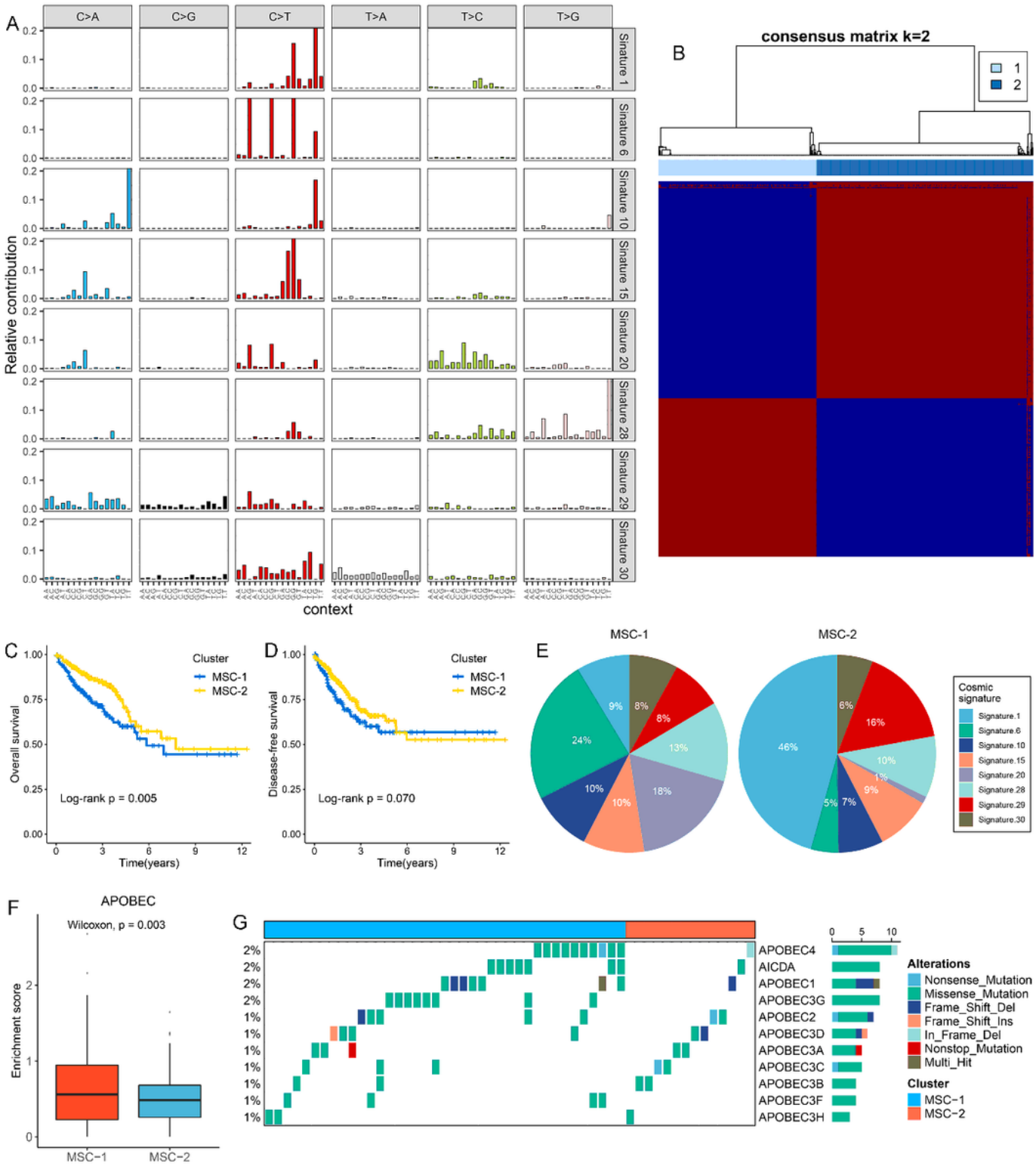

Figure 1

The extraction of mutation signatures and generation of the mutation signature relevant subtypes in CRC. A. Eight mutation signatures were deciphered (mutation signature 1, 6, 10, 15, 20, 28, 29, and 30) based on NMF algorithm and COSMIC signatures. B. The consensus score matrix of all samples when k = 2. A higher consensus score between two samples indicates they are more likely to be grouped into the same cluster in different iterations. C-D. Kaplan–Meier analysis for OS (C) and DFS (D) between MSC-1 and

MSC-2. E. Pie charts show the relative proportion of eight categories of mutation patterns in MSC-1 and MSC-2, respectively. F. The difference of APOBEC enrichment score between MSC-1 and MSC-2. G. Mutational oncoplot of 11 APOBEC associated genes in two subtypes.

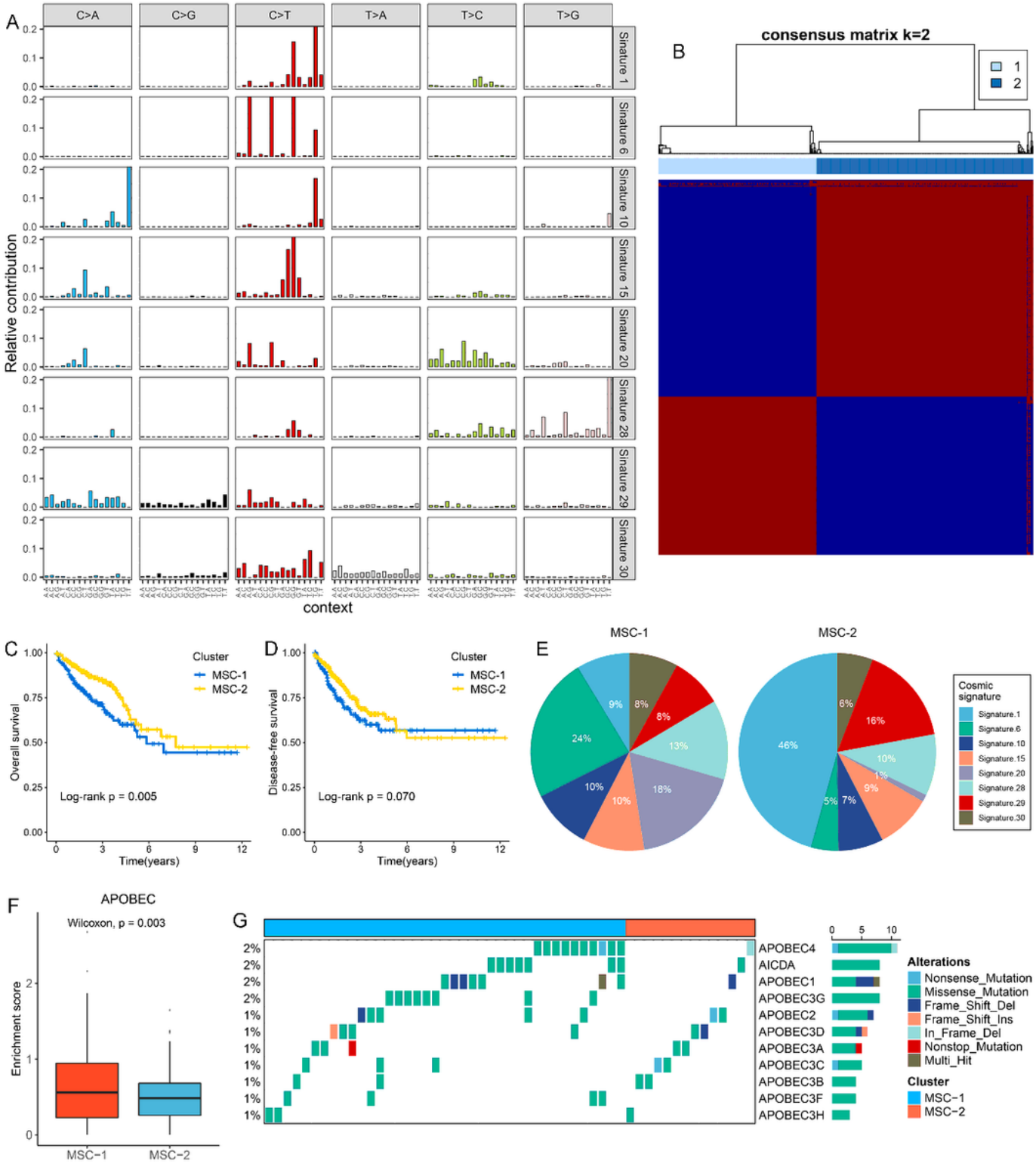

Figure 1

The extraction of mutation signatures and generation of the mutation signature relevant subtypes in CRC. A. Eight mutation signatures were deciphered (mutation signature 1, 6, 10, 15, 20, 28, 29, and 30) based

on NMF algorithm and COSMIC signatures. B. The consensus score matrix of all samples when k = 2. A higher consensus score between two samples indicates they are more likely to be grouped into the same cluster in different iterations. C-D. Kaplan–Meier analysis for OS (C) and DFS (D) between MSC-1 and MSC-2. E. Pie charts show the relative proportion of eight categories of mutation patterns in MSC-1 and MSC-2, respectively. F. The difference of APOBEC enrichment score between MSC-1 and MSC-2. G. Mutational oncoplot of 11 APOBEC associated genes in two subtypes.

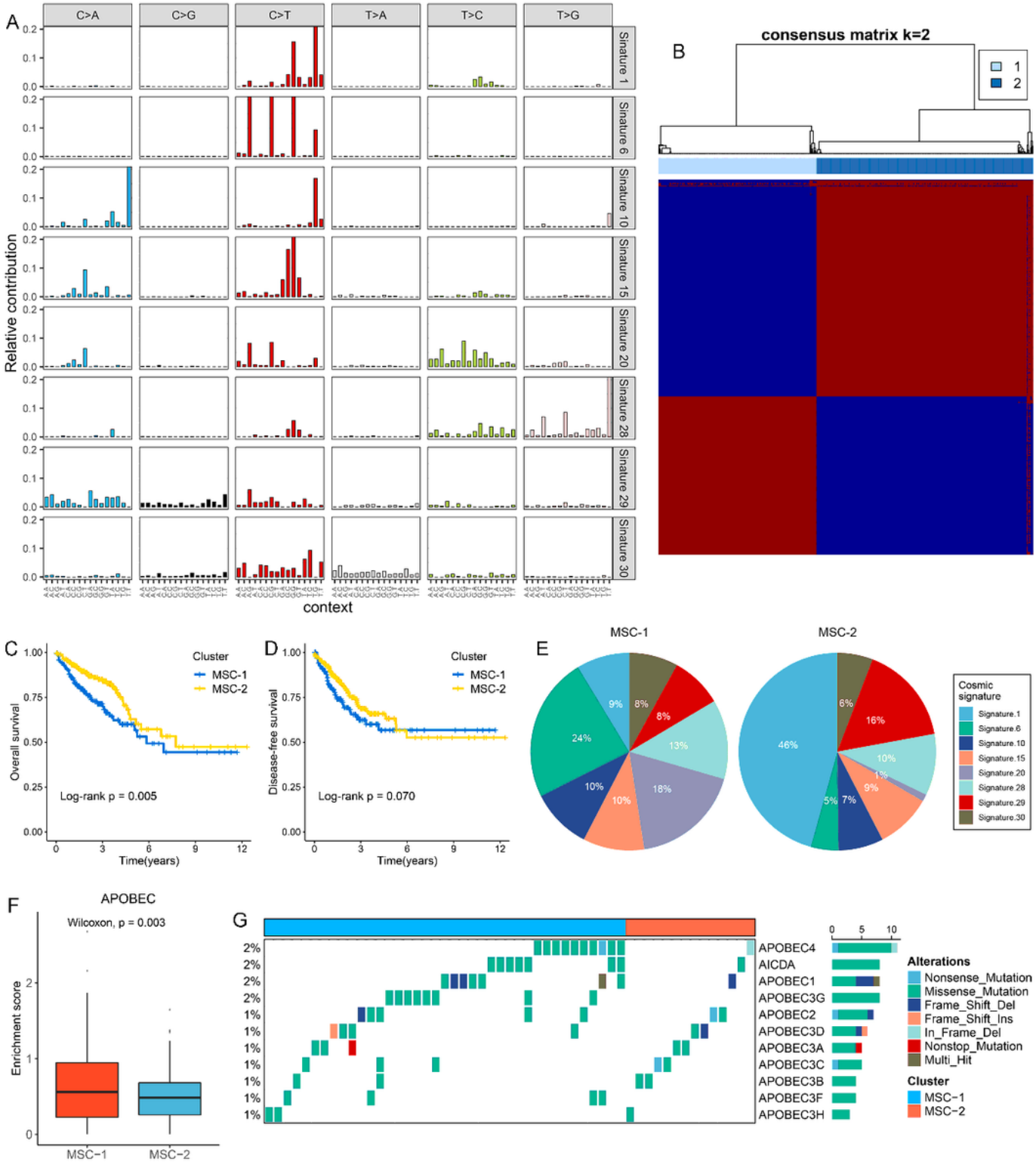

Figure 1

The extraction of mutation signatures and generation of the mutation signature relevant subtypes in CRC. A. Eight mutation signatures were deciphered (mutation signature 1, 6, 10, 15, 20, 28, 29, and 30) based on NMF algorithm and COSMIC signatures. B. The consensus score matrix of all samples when k = 2. A higher consensus score between two samples indicates they are more likely to be grouped into the same cluster in different iterations. C-D. Kaplan–Meier analysis for OS (C) and DFS (D) between MSC-1 and MSC-2. E. Pie charts show the relative proportion of eight categories of mutation patterns in MSC-1 and MSC-2, respectively. F. The difference of APOBEC enrichment score between MSC-1 and MSC-2. G. Mutational oncoplot of 11 APOBEC associated genes in two subtypes.

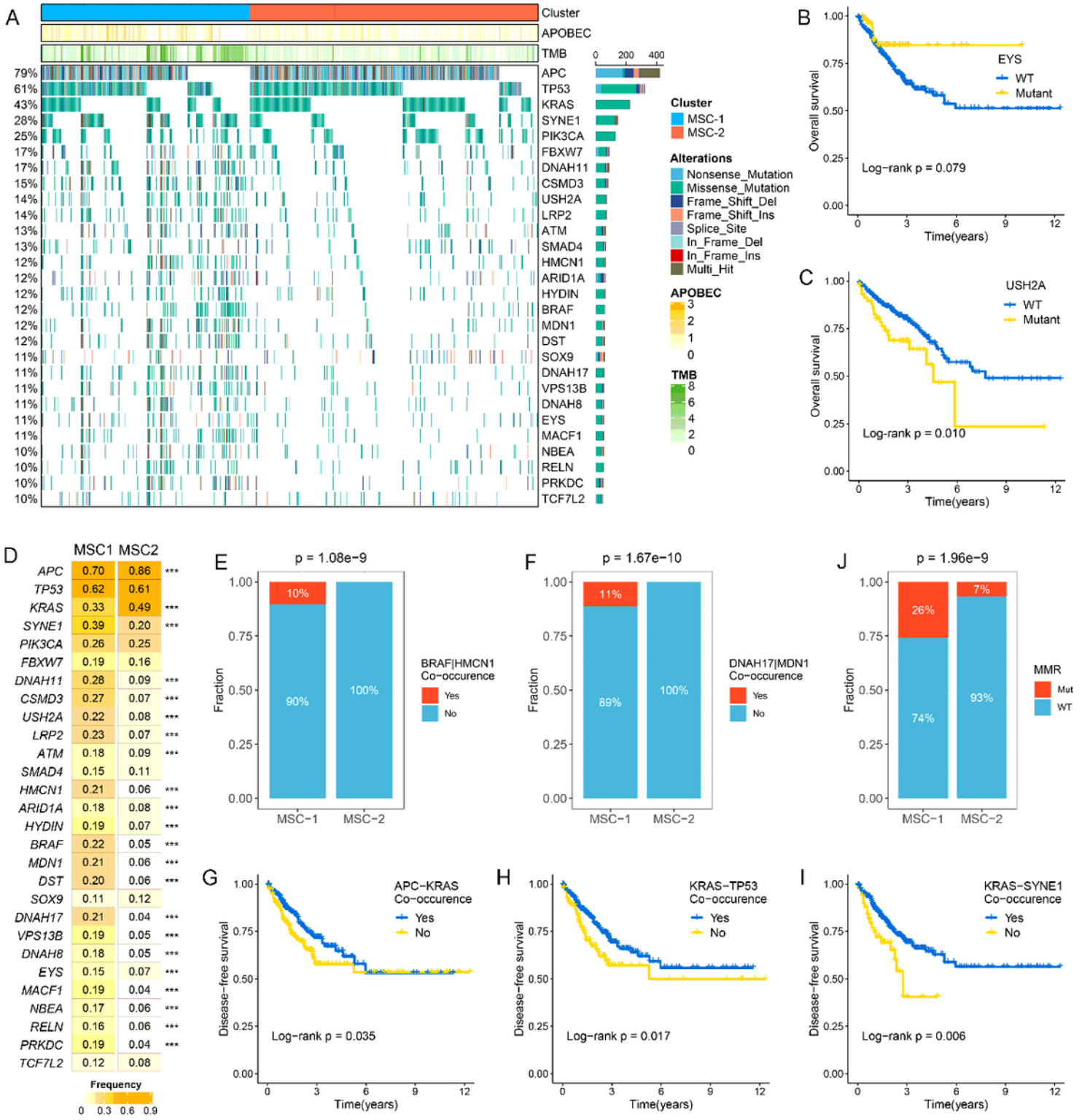

Figure 2

The mutation driven genes in CRC. A. Mutational oncoplot of mutation driven genes in MSC-1 and MSC-2. Rows are genes and columns are tumor samples. B-C. Kaplan–Meier survival analysis of EYS (B) and USH2A mutations (C). D. The mutation frequency of mutational drivers in two subtypes, ***P < 0.001. E-F. The relative proportion of BRAF-HMCN (E) and DNAH17-MDN1 (F) co-occurrences in two subtypes. J. The relative proportion of patients with the MMR genes mutations in two subtypes. G-I. Kaplan–Meier survival analysis of APC–KRAS (G), KRAS–TP53 (H), and KRAS–SYNE1 (I) co-occurrence.

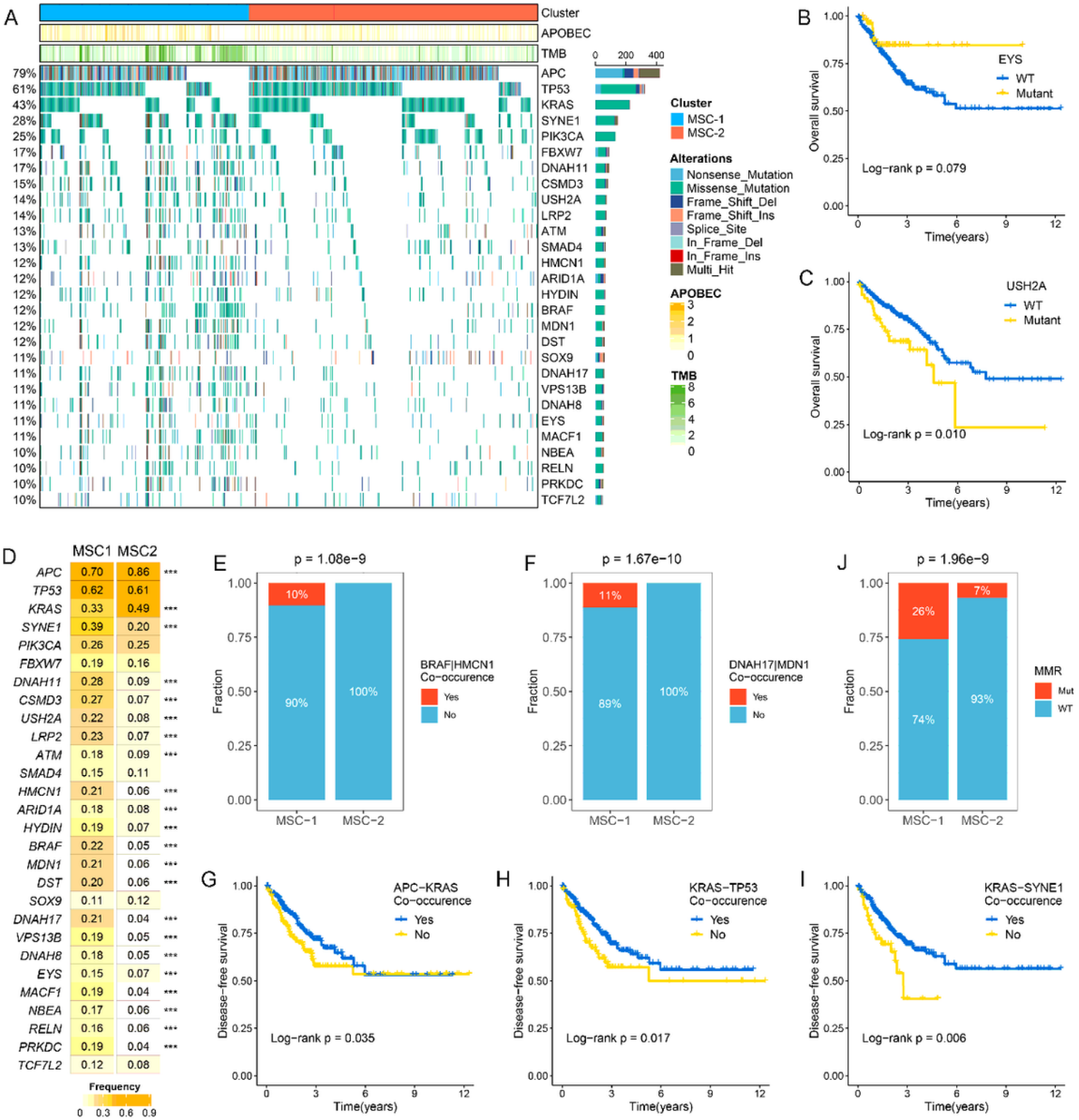

Figure 2

The mutation driven genes in CRC. A. Mutational oncoplot of mutation driven genes in MSC-1 and MSC-2. Rows are genes and columns are tumor samples. B-C. Kaplan–Meier survival analysis of EYS (B) and USH2A mutations (C). D. The mutation frequency of mutational drivers in two subtypes, ***P < 0.001. E-F. The relative proportion of BRAF-HMCN (E) and DNAH17-MDN1 (F) co-occurrences in two subtypes. J. The relative proportion of patients with the MMR genes mutations in two subtypes. G-I. Kaplan–Meier survival analysis of APC–KRAS (G), KRAS–TP53 (H), and KRAS–SYNE1 (I) co-occurrence.

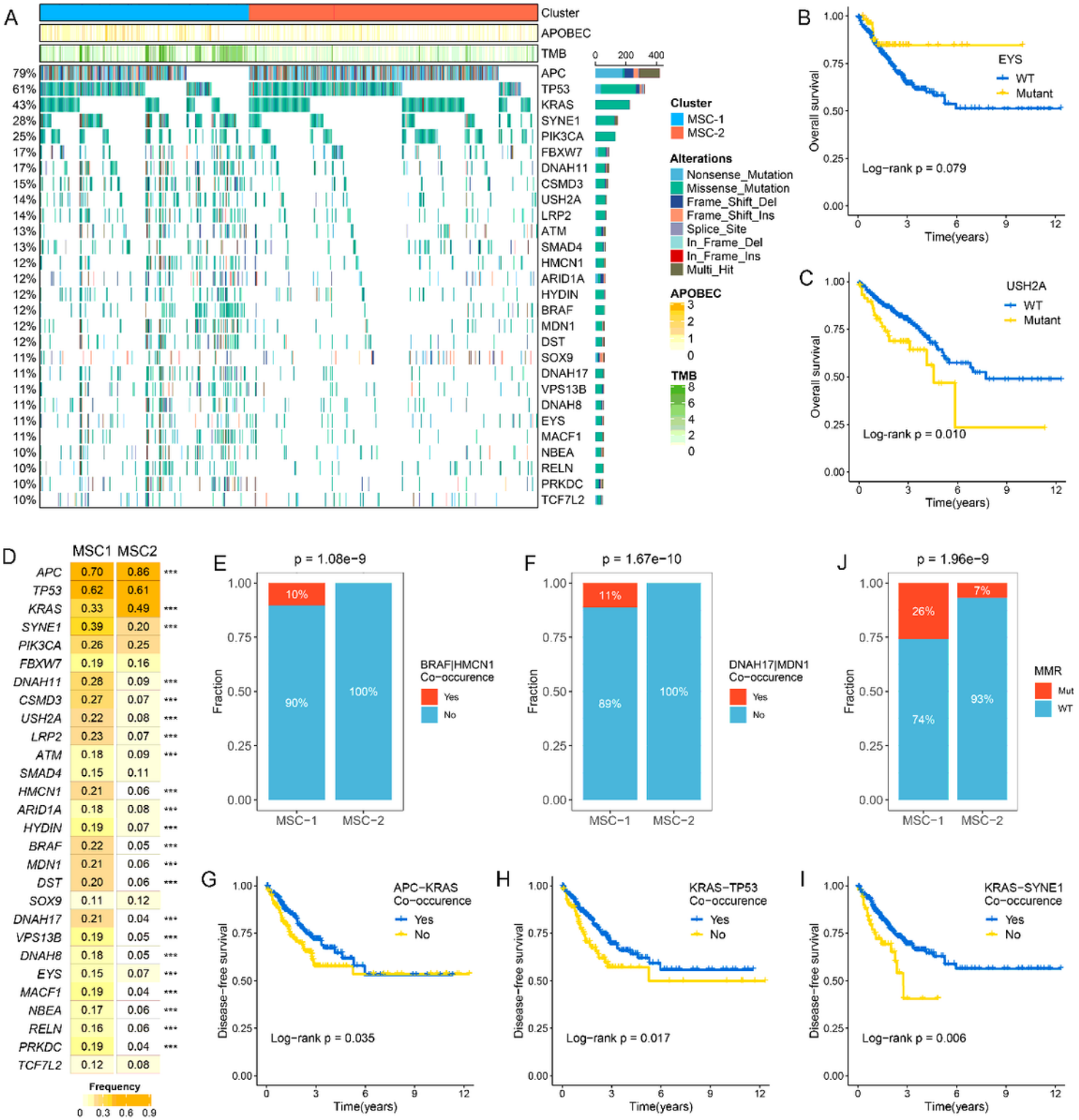

Figure 2

The mutation driven genes in CRC. A. Mutational oncoplot of mutation driven genes in MSC-1 and MSC-2. Rows are genes and columns are tumor samples. B-C. Kaplan–Meier survival analysis of EYS (B) and USH2A mutations (C). D. The mutation frequency of mutational drivers in two subtypes, ***P < 0.001. E-F. The relative proportion of BRAF-HMCN (E) and DNAH17-MDN1 (F) co-occurrences in two subtypes. J. The relative proportion of patients with the MMR genes mutations in two subtypes. G-I. Kaplan–Meier survival analysis of APC–KRAS (G), KRAS–TP53 (H), and KRAS–SYNE1 (I) co-occurrence.

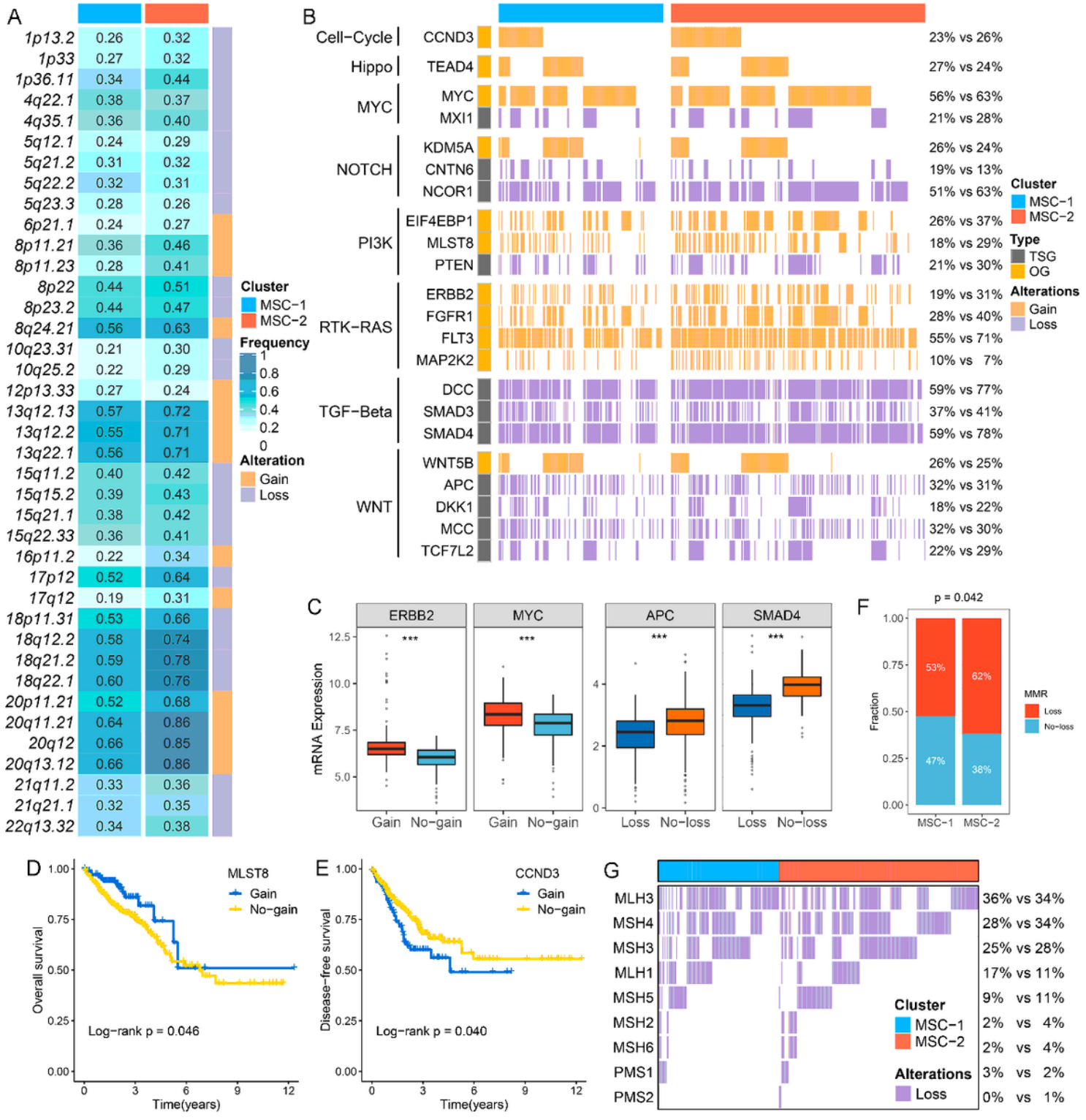

Figure 3

The driven segments identified from GISTIC algorithm in CRC. A. The amplification (orange) and deletion (purple) frequency of 39 driven segments in two subtypes. B. The distribution of CNA relevant oncogenes and tumor suppressive genes in two subtypes. C. The expression difference of ERRB2 and MYC between the gain and no-gain groups, as well as APC and SMAD4 between the loss and no-loss groups. ***, $P < 0.001$. D-E. Kaplan–Meier survival analysis of MLST (D) and CCND3 (E) gain. F. The relative proportion of patients with the MMR genes deletions in two subtypes. G. Oncoplot for the deletion of nine MMR-related genes in two subtypes.

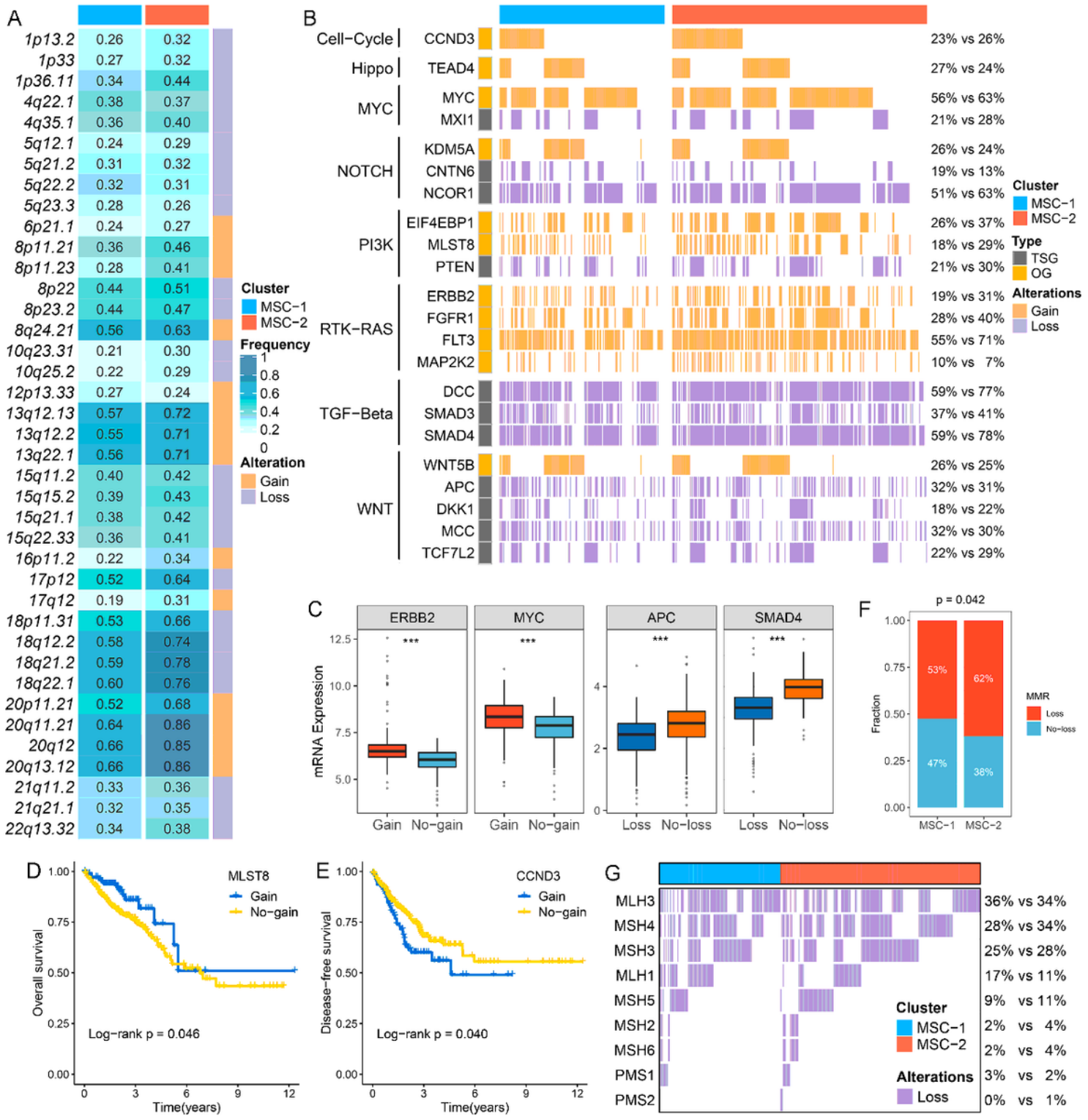

Figure 3

The driven segments identified from GISTIC algorithm in CRC. A. The amplification (orange) and deletion (purple) frequency of 39 driven segments in two subtypes. B. The distribution of CNA relevant oncogenes and tumor suppressive genes in two subtypes. C. The expression difference of ERRB2 and MYC between the gain and no-gain groups, as well as APC and SMAD4 between the loss and no-loss groups. ***, P < 0.001. D-E. Kaplan–Meier survival analysis of MLST (D) and CCND3 (E) gain. F. The relative proportion of

patients with the MMR genes deletions in two subtypes. G. Oncoplot for the deletion of nine MMR-related genes in two subtypes.

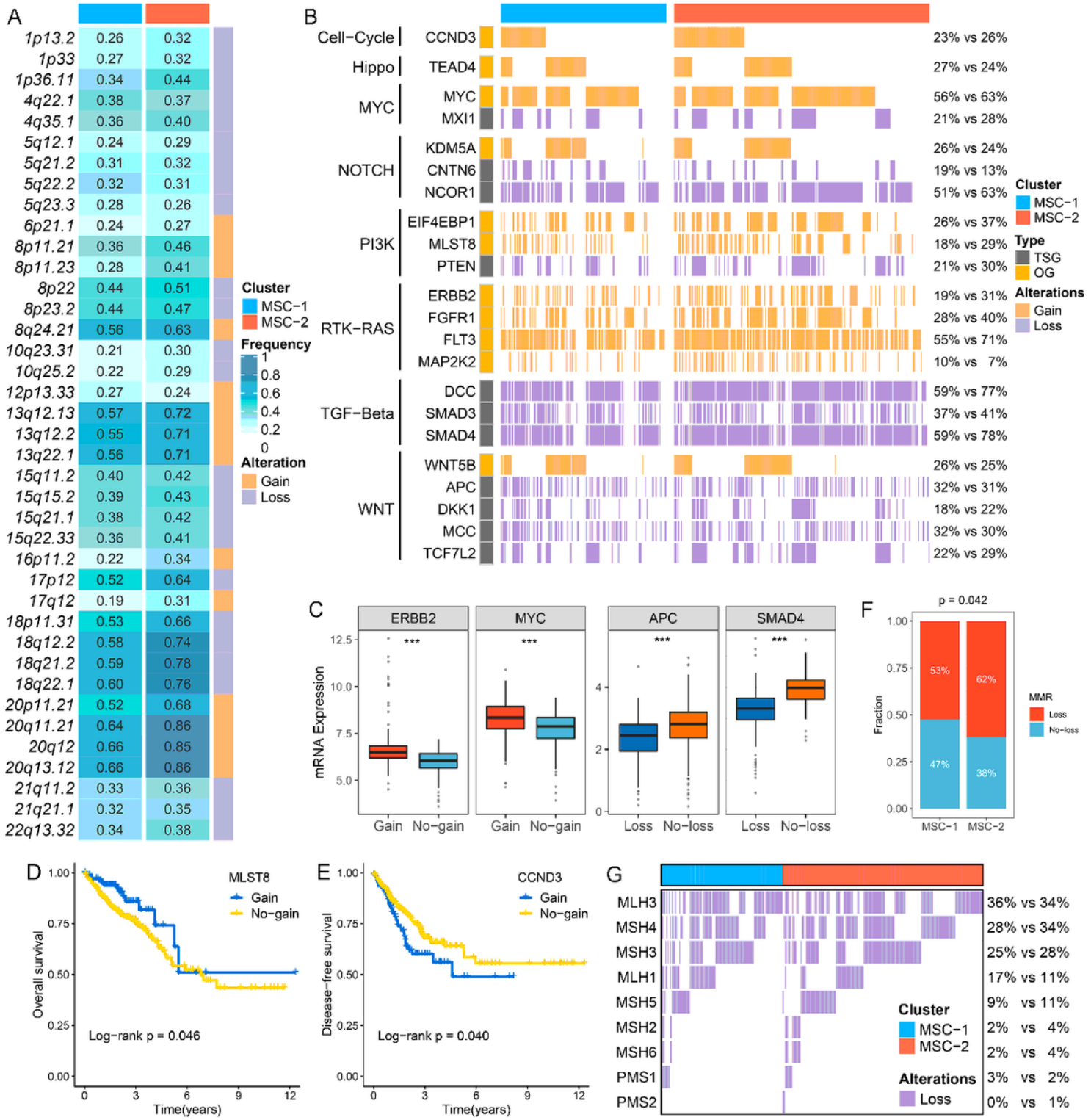

Figure 3

The driven segments identified from GISTIC algorithm in CRC. A. The amplification (orange) and deletion (purple) frequency of 39 driven segments in two subtypes. B. The distribution of CNA relevant oncogenes and tumor suppressive genes in two subtypes. C. The expression difference of ERRB2 and MYC between

the gain and no-gain groups, as well as APC and SMAD4 between the loss and no-loss groups. ***, P < 0.001. D-E. Kaplan–Meier survival analysis of MLST (D) and CCND3 (E) gain. F. The relative proportion of patients with the MMR genes deletions in two subtypes. G. Oncoplot for the deletion of nine MMR-related genes in two subtypes.

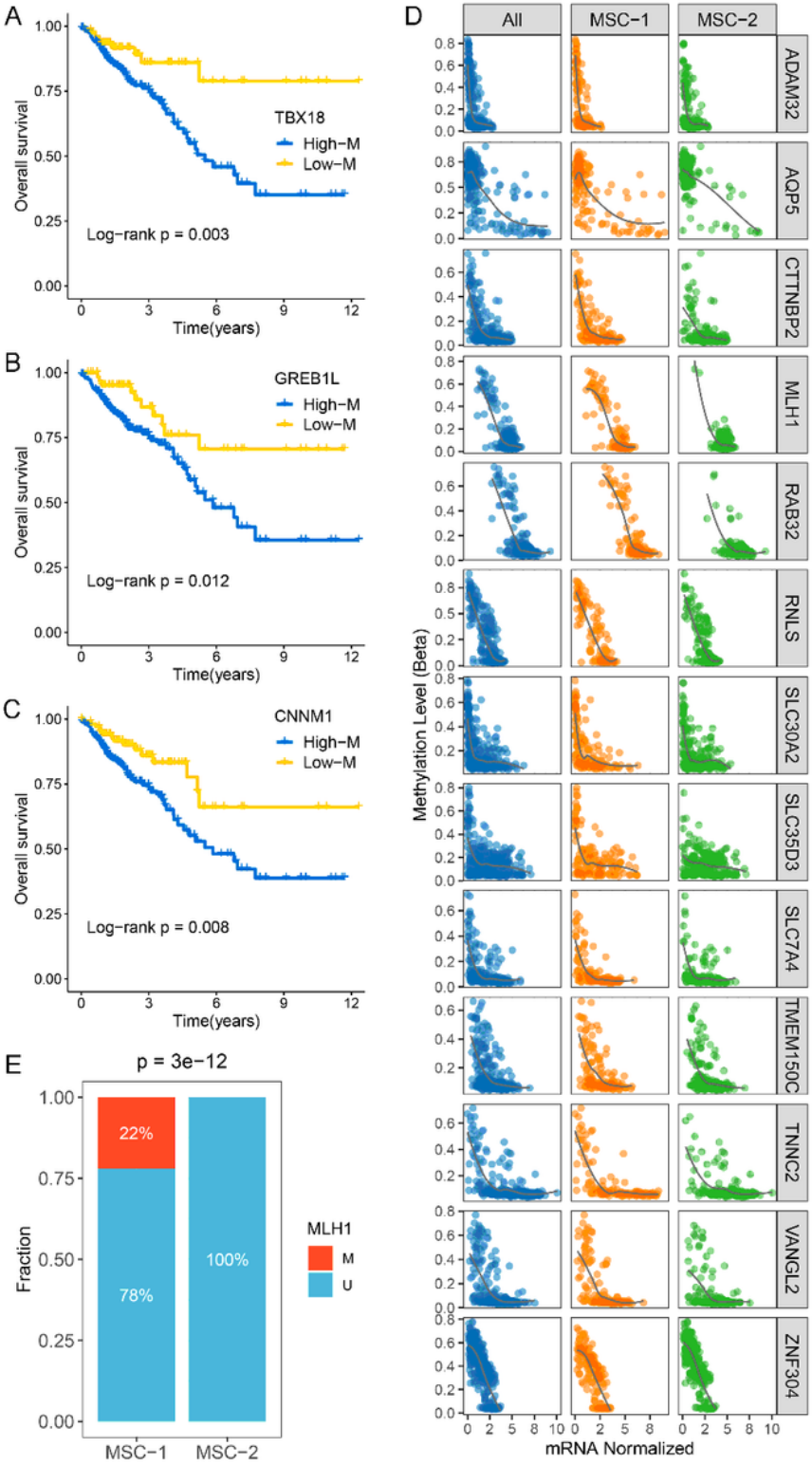

Figure 4

The methylation driven genes in CRC. A-C. Kaplan–Meier survival analysis of TBX18 (A), GREB1L, (B) and CNNM1 (C) methylation. D. The correlation analysis between the methylation and mRNA expression levels of13 ssMDGs. E. The relative proportion of patients with the MMR genes methylation events between two subtypes.

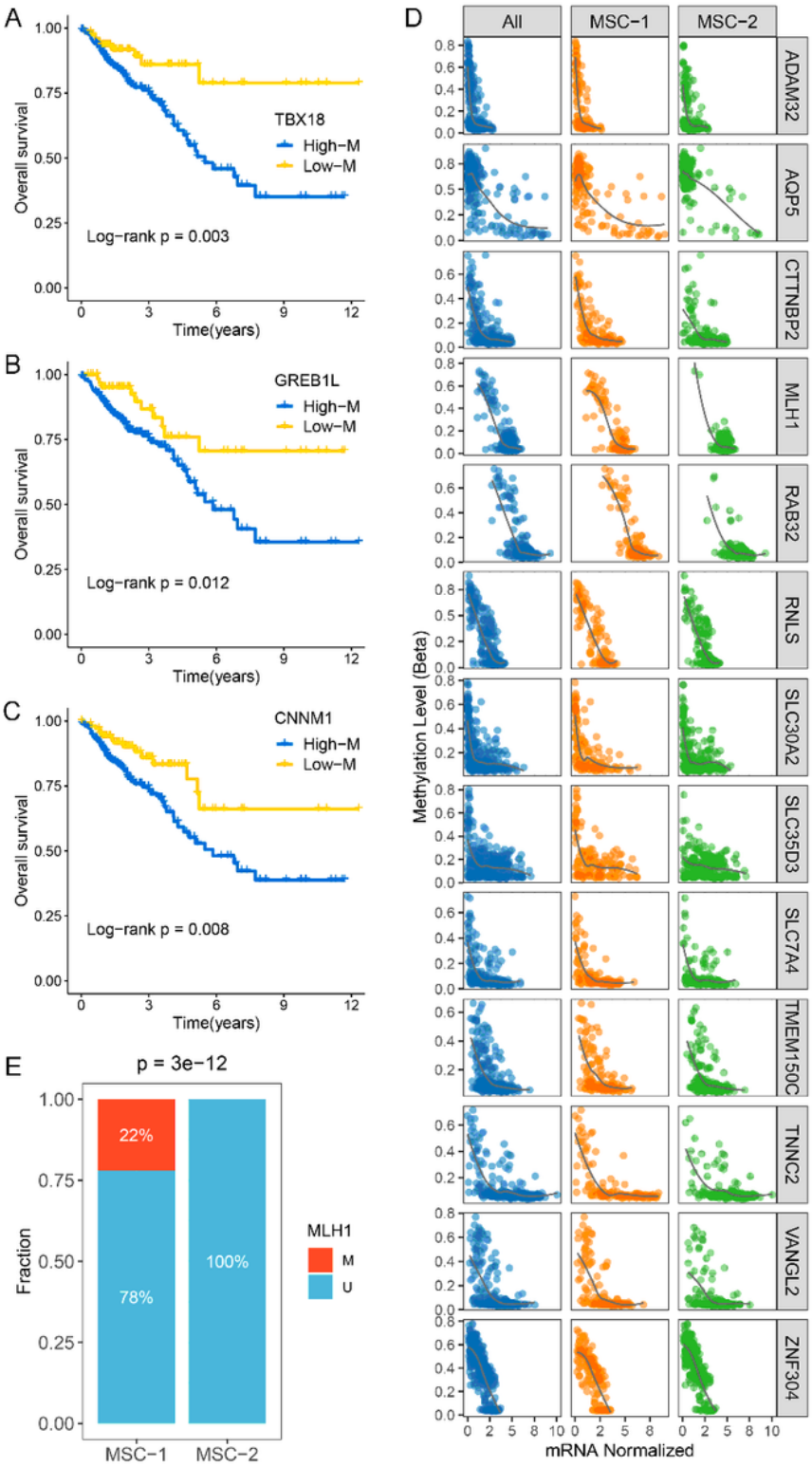

Figure 4

The methylation driven genes in CRC. A-C. Kaplan–Meier survival analysis of TBX18 (A), GREB1L, (B) and CNNM1 (C) methylation. D. The correlation analysis between the methylation and mRNA expression levels of13 ssMDGs. E. The relative proportion of patients with the MMR genes methylation events between two subtypes.

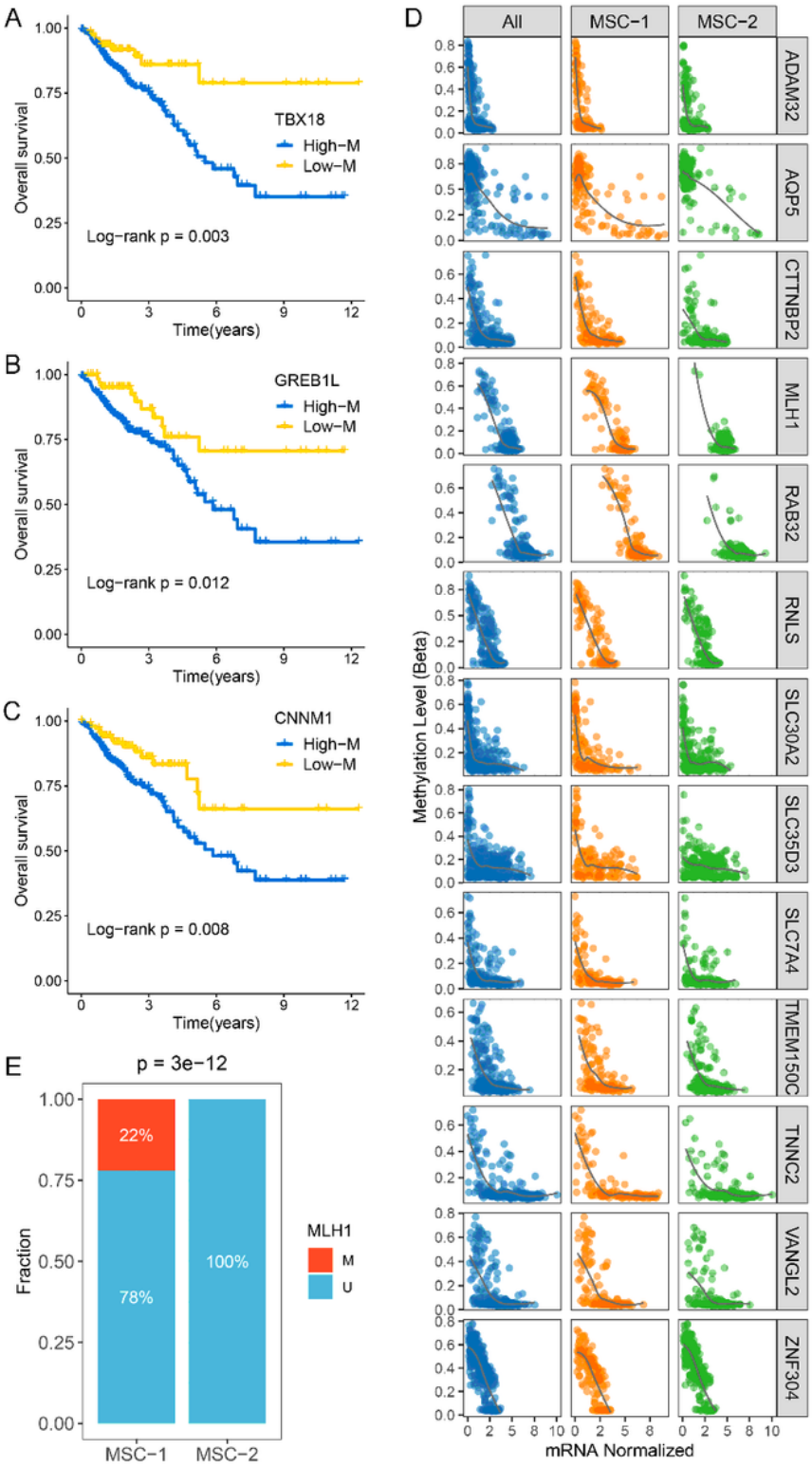

Figure 4

The methylation driven genes in CRC. A-C. Kaplan–Meier survival analysis of TBX18 (A), GREB1L, (B) and CNNM1 (C) methylation. D. The correlation analysis between the methylation and mRNA expression levels of13 ssMDGs. E. The relative proportion of patients with the MMR genes methylation events between two subtypes.

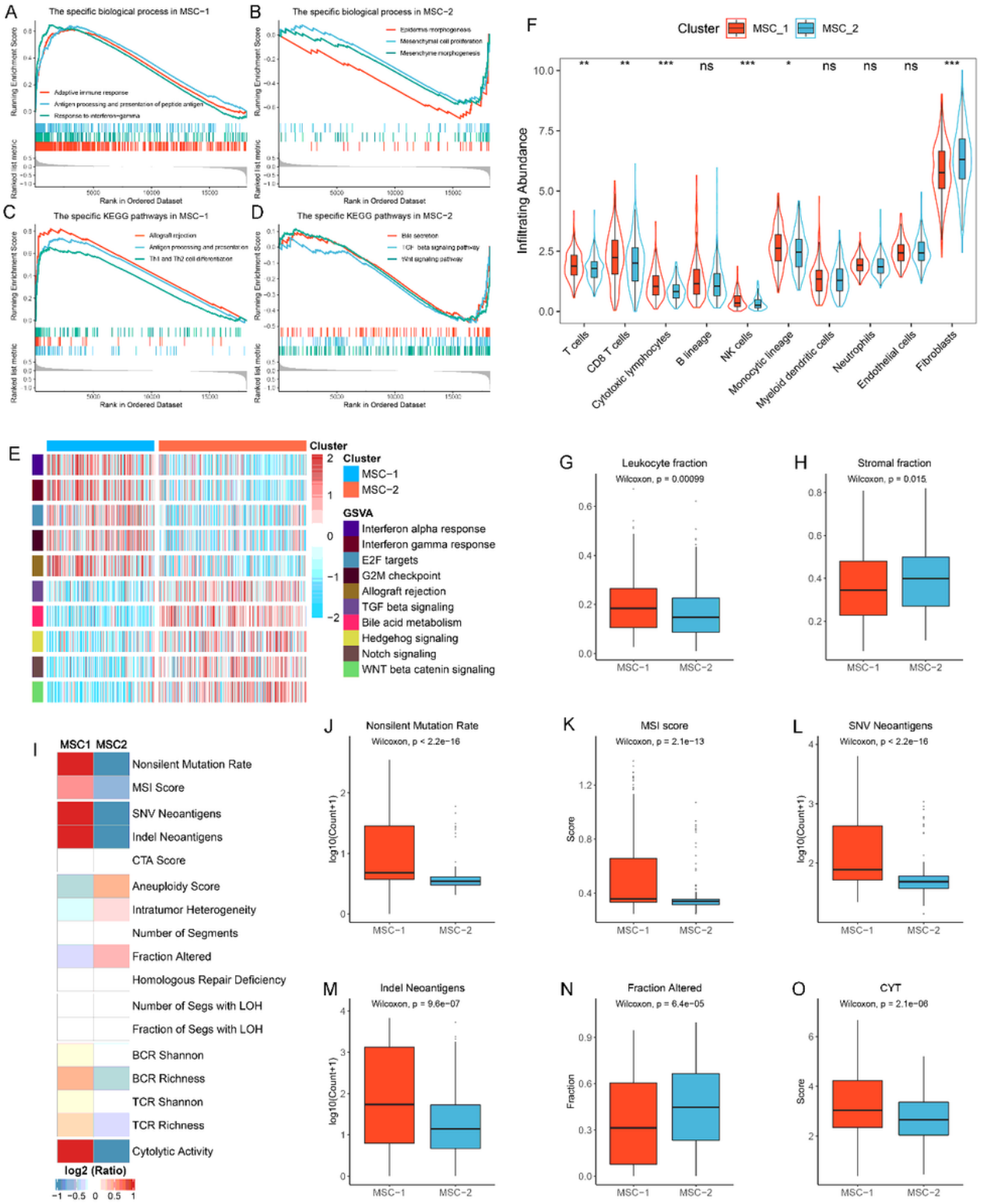

Figure 5

Functional status, immune cell infiltration and immunogenicity assessment. A-B. The biological process significantly enriched in MSC-1 (A) and MSC-2 (B). C-D. The KEGG pathways significantly enriched in MSC-1 (C) and MSC-2 (D). E. The specific Hallmark pathways in MSC-1 and MSC-2. F. The infiltration abundance of eight immune cells and two nonimmune cells populations in MSC-1 and MSC-2. ns, $P > 0.05$; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. G-H. The distribution of leukocyte (G) and stomal (H) fraction in MSC-1 and MSC-2. I. The comparison of 17 immunogenicity associated indicators between two subtypes, the cell represented by the mean value of corresponding cluster divided by the overall mean value. J-O. The distribution of nonsilent mutation rate (J), MSI score (K), SNV neoantigens (L), Indel neoantigens (M), fraction of segments alteration (N), and cytolytic activity (CYT) (O) in MSC-1 and MSC-2.

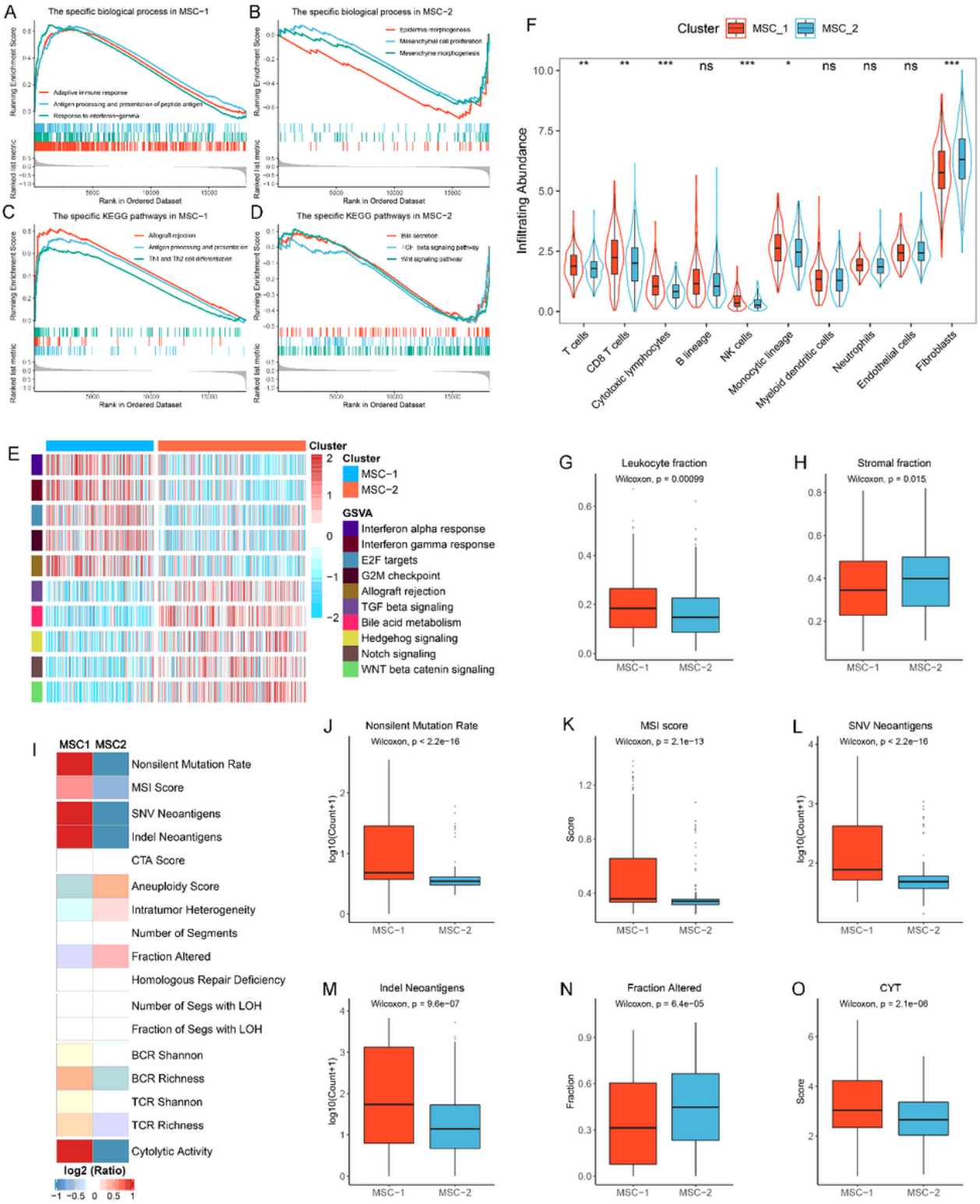

Figure 5

Functional status, immune cell infiltration and immunogenicity assessment. A-B. The biological process significantly enriched in MSC-1 (A) and MSC-2 (B). C-D. The KEGG pathways significantly enriched in MSC-1 (C) and MSC-2 (D). E. The specific Hallmark pathways in MSC-1 and MSC-2. F. The infiltration abundance of eight immune cells and two nonimmune cells populations in MSC-1 and MSC-2. ns, P > 0.05; *, P < 0.05; **, P < 0.01; ***, P < 0.001. G-H. The distribution of leukocyte (G) and stomal (H) fraction

in MSC-1 and MSC-2. I. The comparison of 17 immunogenicity associated indicators between two subtypes, the cell represented by the mean value of corresponding cluster divided by the overall mean value. J-O. The distribution of nonsilent mutation rate (J), MSI score (K), SNV neoantigens (L), Indel neoantigens (M), fraction of segments alteration (N), and cytolytic activity (CYT) (O) in MSC-1 and MSC-2.

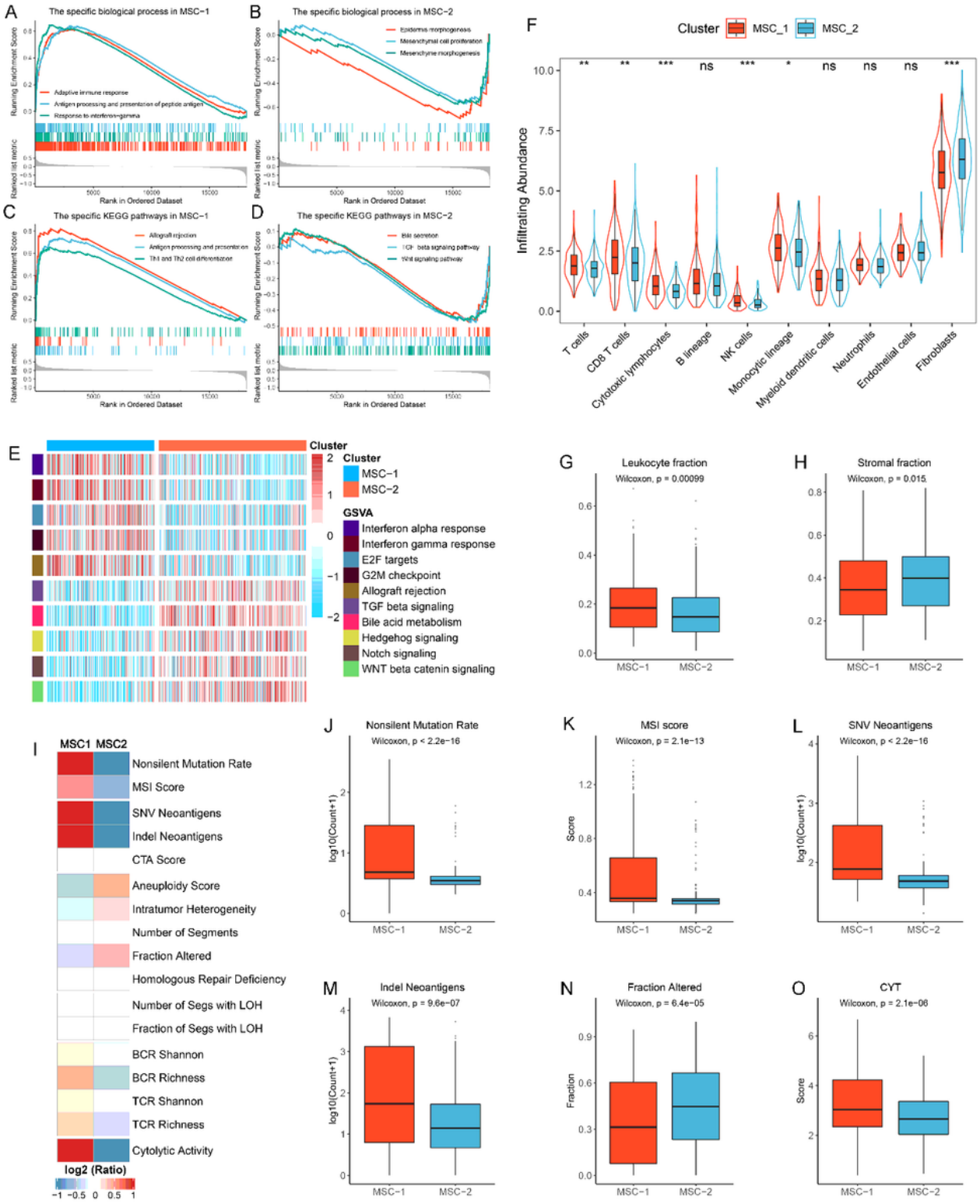

Figure 5

Functional status, immune cell infiltration and immunogenicity assessment. A-B. The biological process significantly enriched in MSC-1 (A) and MSC-2 (B). C-D. The KEGG pathways significantly enriched in MSC-1 (C) and MSC-2 (D). E. The specific Hallmark pathways in MSC-1 and MSC-2. F. The infiltration abundance of eight immune cells and two nonimmune cells populations in MSC-1 and MSC-2. ns, $P > 0.05$; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. G-H. The distribution of leukocyte (G) and stomal (H) fraction in MSC-1 and MSC-2. I. The comparison of 17 immunogenicity associated indicators between two subtypes, the cell represented by the mean value of corresponding cluster divided by the overall mean value. J-O. The distribution of nonsilent mutation rate (J), MSI score (K), SNV neoantigens (L), Indel neoantigens (M), fraction of segments alteration (N), and cytolytic activity (CYT) (O) in MSC-1 and MSC-2.

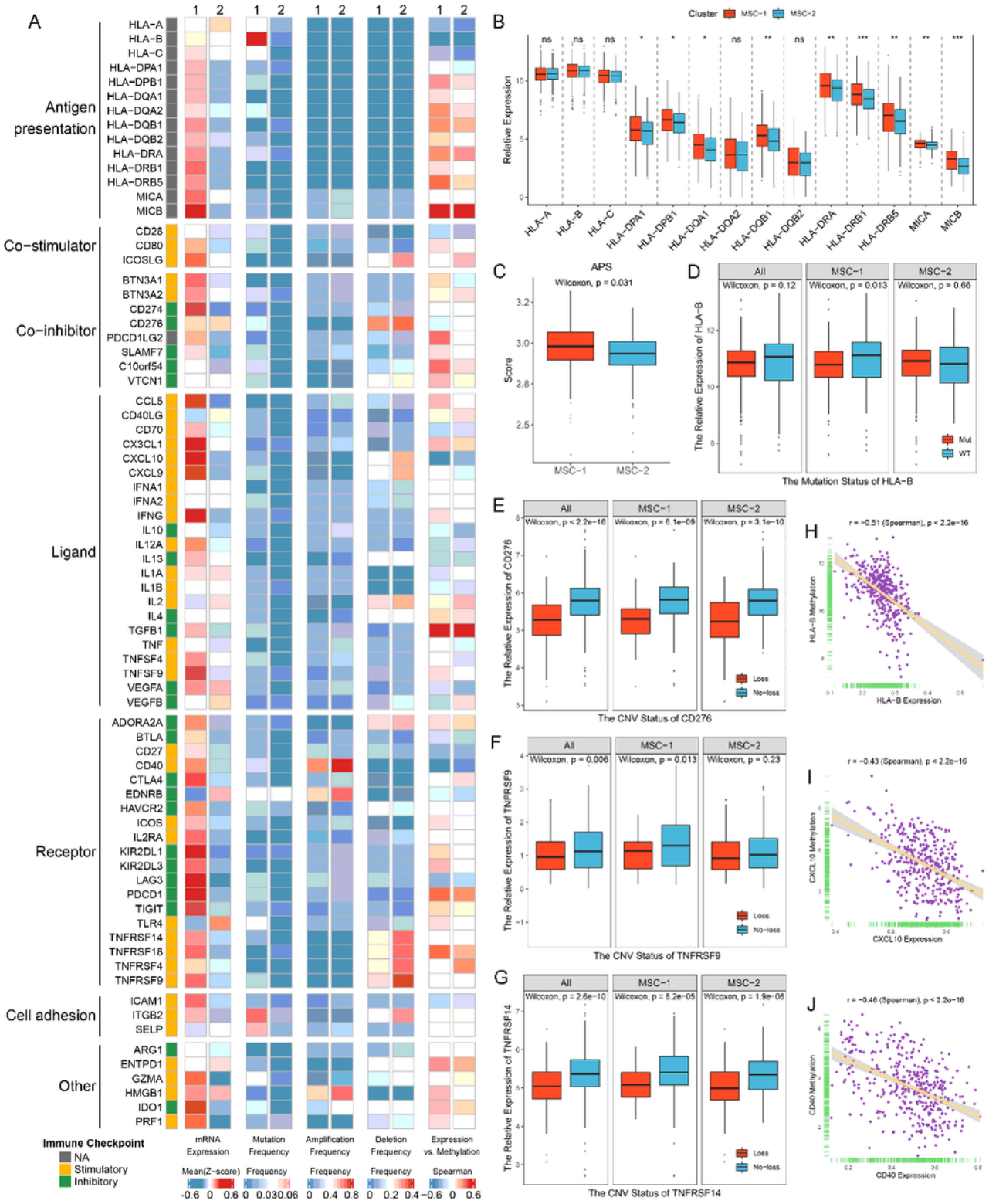

Figure 6

Multi-omics analysis of 75 immunomodulators in the TCGA-CRC cohort. A. From left to right: mRNA expression (z-score), mutation frequency, amplification frequency, deletion frequency, and expression versus methylation (gene expression correlation with DNA-methylation beta value) of 75 immunomodulators in MSC-1 and MSC-2. B. The expression difference of MHC molecules between two subtypes. ns, P > 0.05; *, P < 0.05; **, P < 0.01; ***, P < 0.001. C. The distribution of APS score in MSC-1

and MSC-2. D-G. The expression difference of HLA-B (D) between the mutant and wild groups, as well as CD276 (E), TNFRSF9 (F), and TNFRSF14 (G) between loss and no-loss groups. H-J. The correlation analysis between the methylation and mRNA expression levels of HLA-B (H), CXCL10 (I), and CD40 (J).

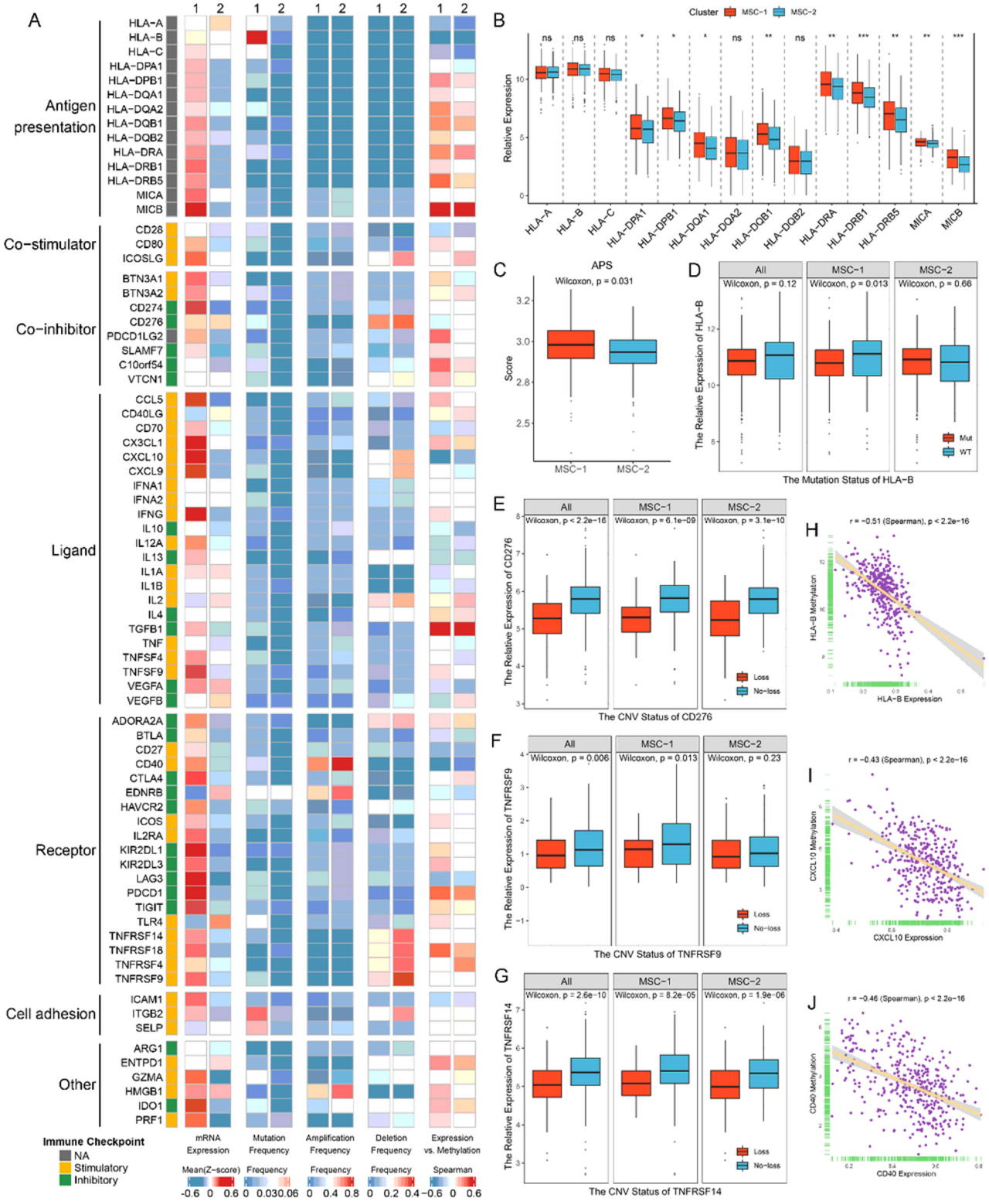

Figure 6

Multi-omics analysis of 75 immunomodulators in the TCGA-CRC cohort. A. From left to right: mRNA expression (z-score), mutation frequency, amplification frequency, deletion frequency, and expression

versus methylation (gene expression correlation with DNA-methylation beta value) of 75 immunomodulators in MSC-1 and MSC-2. B. The expression difference of MHC molecules between two subtypes. ns, P > 0.05; *, P < 0.05; **, P < 0.01; ***, P < 0.001. C. The distribution of APS score in MSC-1 and MSC-2. D-G. The expression difference of HLA-B (D) between the mutant and wild groups, as well as CD276 (E), TNFRSF9 (F), and TNFRSF14 (G) between loss and no-loss groups. H-J. The correlation analysis between the methylation and mRNA expression levels of HLA-B (H), CXCL10 (I), and CD40 (J).

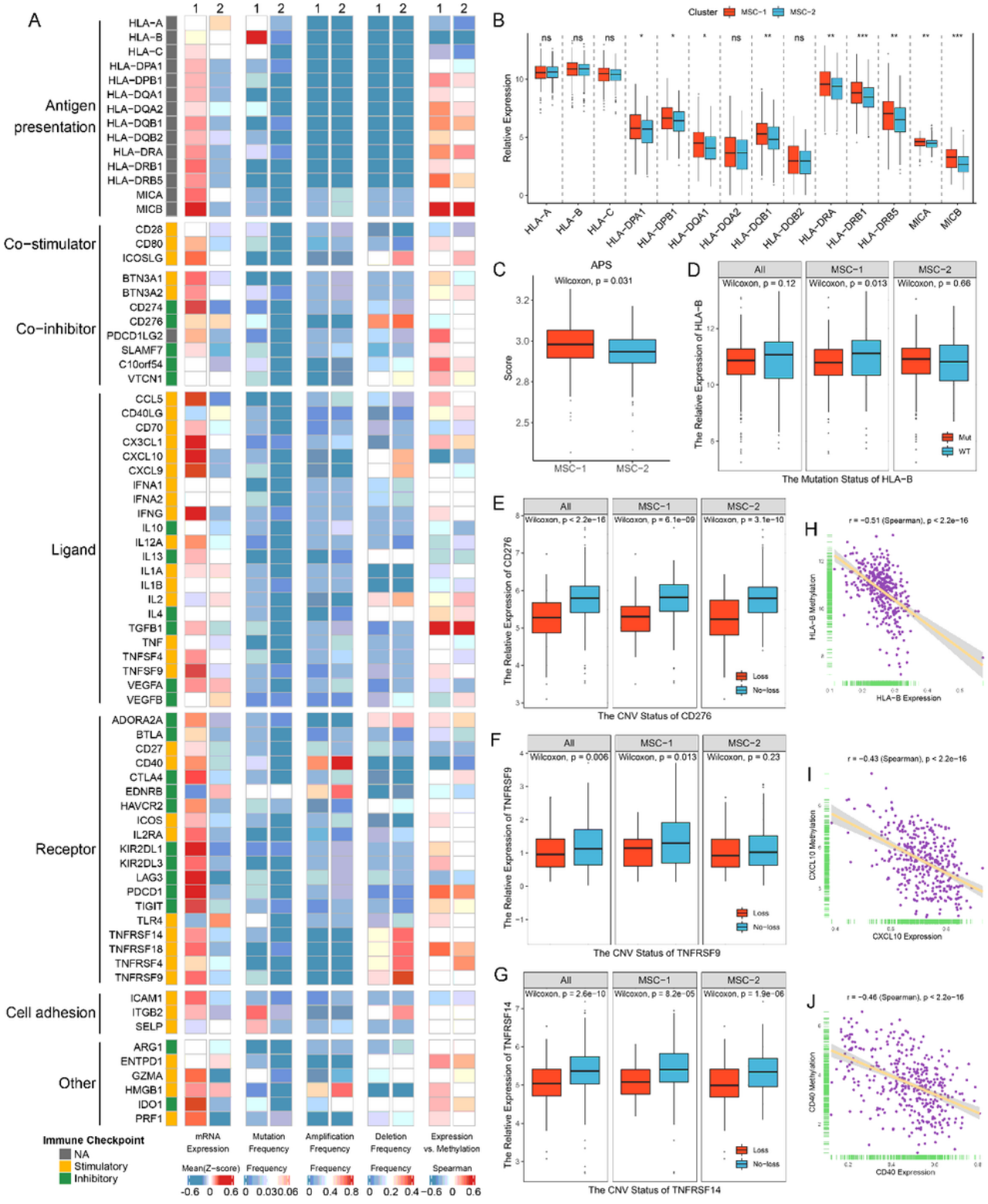

Figure 6

Multi-omics analysis of 75 immunomodulators in the TCGA-CRC cohort. A. From left to right: mRNA expression (z-score), mutation frequency, amplification frequency, deletion frequency, and expression versus methylation (gene expression correlation with DNA-methylation beta value) of 75 immunomodulators in MSC-1 and MSC-2. B. The expression difference of MHC molecules between two subtypes. ns, $P > 0.05$; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. C. The distribution of APS score in MSC-1 and MSC-2. D-G. The expression difference of HLA-B (D) between the mutant and wild groups, as well as CD276 (E), TNFRSF9 (F), and TNFRSF14 (G) between loss and no-loss groups. H-J. The correlation analysis between the methylation and mRNA expression levels of HLA-B (H), CXCL10 (I), and CD40 (J).

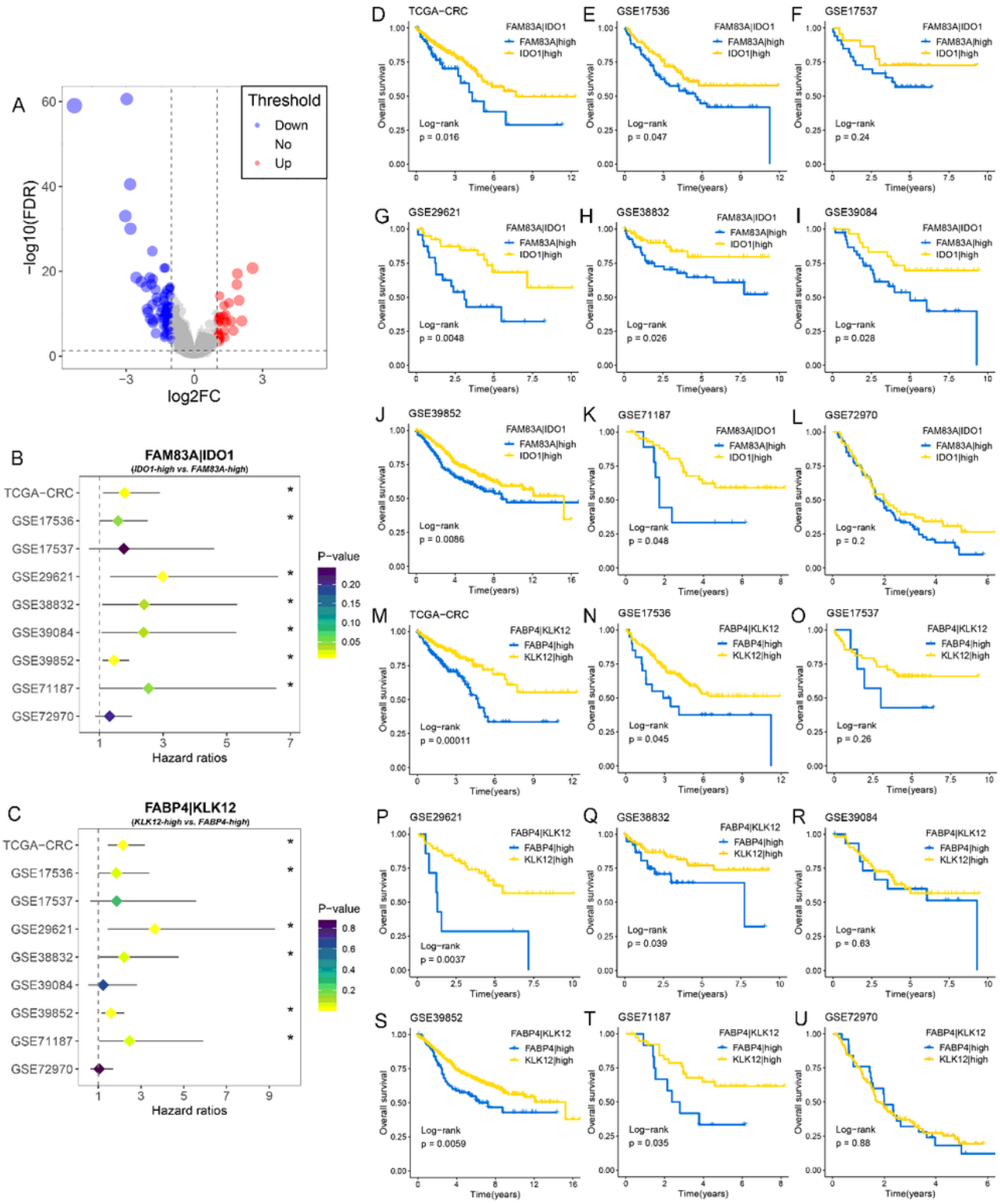

Figure 7

Identification of gene pairs with the ability to predict prognosis of CRC patients. A. Volcano plot of differentially expressed genes (DEGs) between MSC-1 and MSC-2. The abscissa is log2FC, and the ordinate is −log10 (FDR). The red and blue points in the plot represent DEGs with statistical significance (FDR < 0.05 and |log2FC| > 1). B. Forest plot of IDO1-high versus FAM83A-high groups in nine cohorts. C. Forest plot of KLK12-high versus FABP4-high groups in nine cohorts. D-L. Kaplan–Meier survival analysis

for FAM83A|IDO1 in the TCGA-CRC (D), GSE17536 (E), GSE17537 (F), GSE29621 (G), GSE38832 (H), GSE39084 (I), GSE39852 (J), GSE71187 (K), and GSE72970 (L) cohorts. M-U. Kaplan–Meier survival analysis for FABP4|KLK12 in the TCGA-CRC (M), GSE17536 (N), GSE17537 (O), GSE29621 (P), GSE38832 (Q), GSE39084 (R), GSE39852 (S), GSE71187 (T), and GSE72970 (U) cohorts.

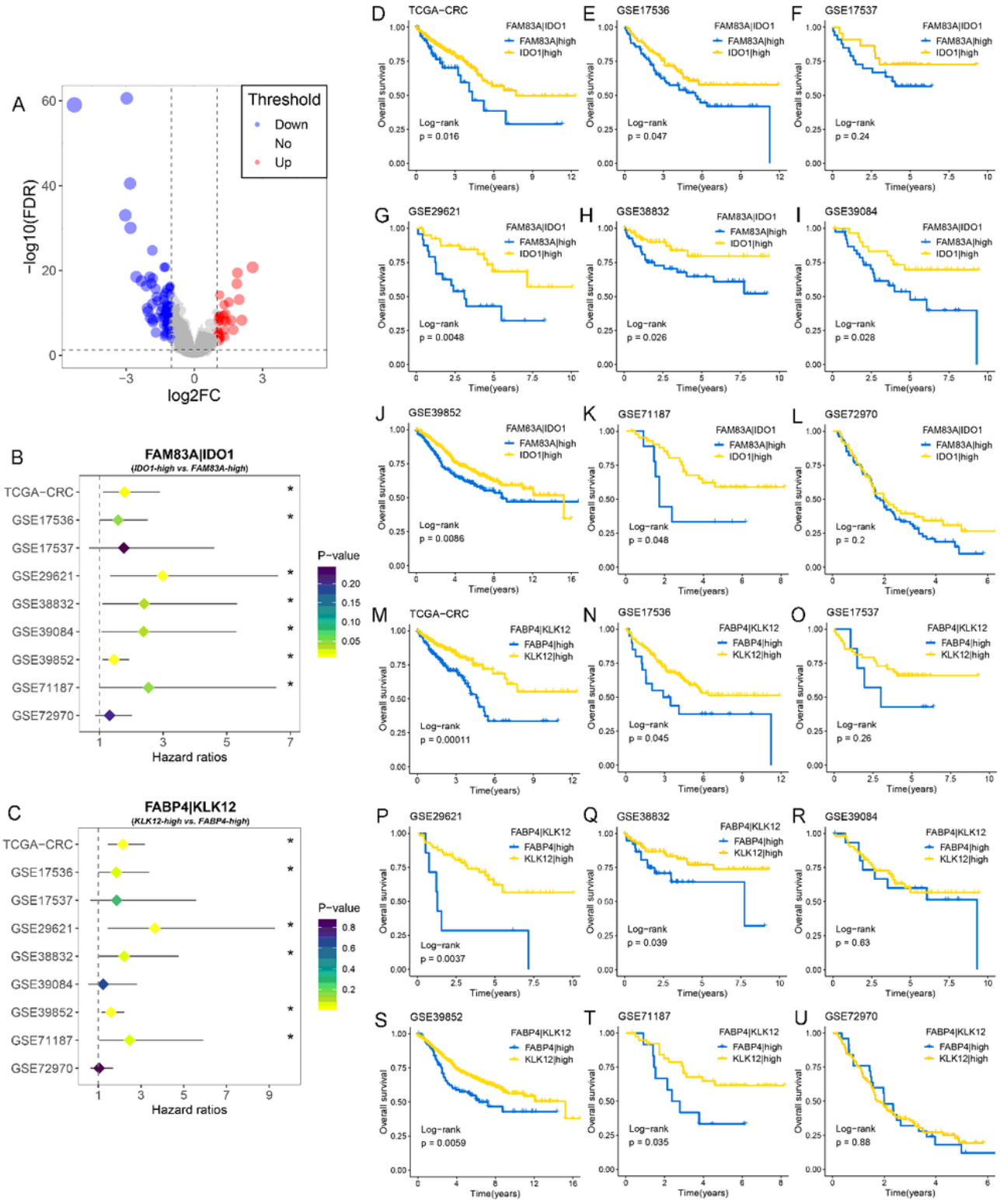

Figure 7

Identification of gene pairs with the ability to predict prognosis of CRC patients. A. Volcano plot of differentially expressed genes (DEGs) between MSC-1 and MSC-2. The abscissa is log2FC, and the ordinate is −log10 (FDR). The red and blue points in the plot represent DEGs with statistical significance (FDR < 0.05 and |log2FC| > 1). B. Forest plot of IDO1-high versus FAM83A-high groups in nine cohorts. C. Forest plot of KLK12-high versus FABP4-high groups in nine cohorts. D-L. Kaplan–Meier survival analysis for FAM83A|IDO1 in the TCGA-CRC (D), GSE17536 (E), GSE17537 (F), GSE29621 (G), GSE38832 (H), GSE39084 (I), GSE39852 (J), GSE71187 (K), and GSE72970 (L) cohorts. M-U. Kaplan–Meier survival analysis for FABP4|KLK12 in the TCGA-CRC (M), GSE17536 (N), GSE17537 (O), GSE29621 (P), GSE38832 (Q), GSE39084 (R), GSE39852 (S), GSE71187 (T), and GSE72970 (U) cohorts.

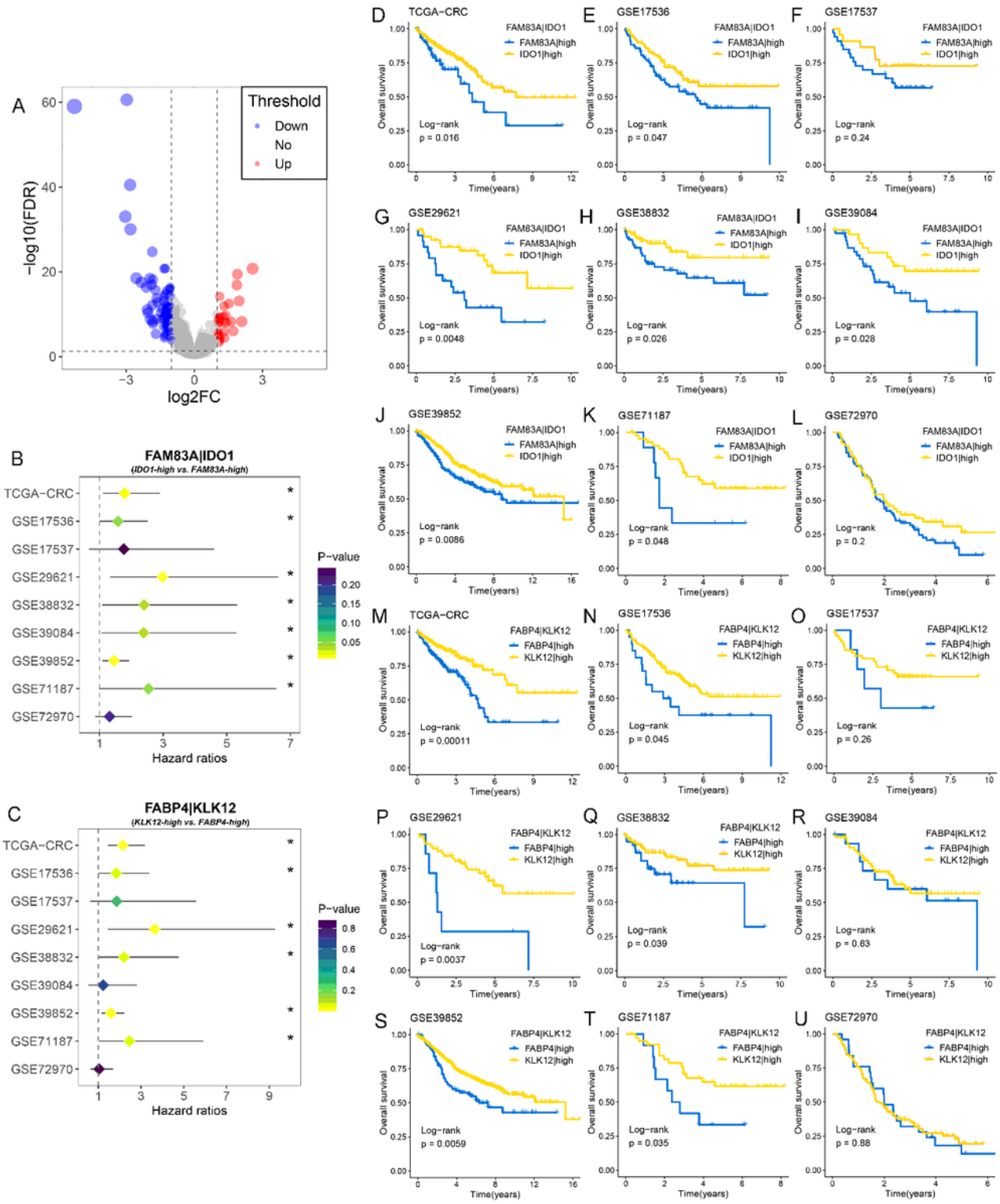

Figure 7

Identification of gene pairs with the ability to predict prognosis of CRC patients. A. Volcano plot of differentially expressed genes (DEGs) between MSC-1 and MSC-2. The abscissa is log2FC, and the ordinate is −log10 (FDR). The red and blue points in the plot represent DEGs with statistical significance (FDR < 0.05 and |log2FC| > 1). B. Forest plot of IDO1-high versus FAM83A-high groups in nine cohorts. C. Forest plot of KLK12-high versus FABP4-high groups in nine cohorts. D-L. Kaplan–Meier survival analysis

for FAM83A|IDO1 in the TCGA-CRC (D), GSE17536 (E), GSE17537 (F), GSE29621 (G), GSE38832 (H), GSE39084 (I), GSE39852 (J), GSE71187 (K), and GSE72970 (L) cohorts. M-U. Kaplan–Meier survival analysis for FABP4|KLK12 in the TCGA-CRC (M), GSE17536 (N), GSE17537 (O), GSE29621 (P), GSE38832 (Q), GSE39084 (R), GSE39852 (S), GSE71187 (T), and GSE72970 (U) cohorts.

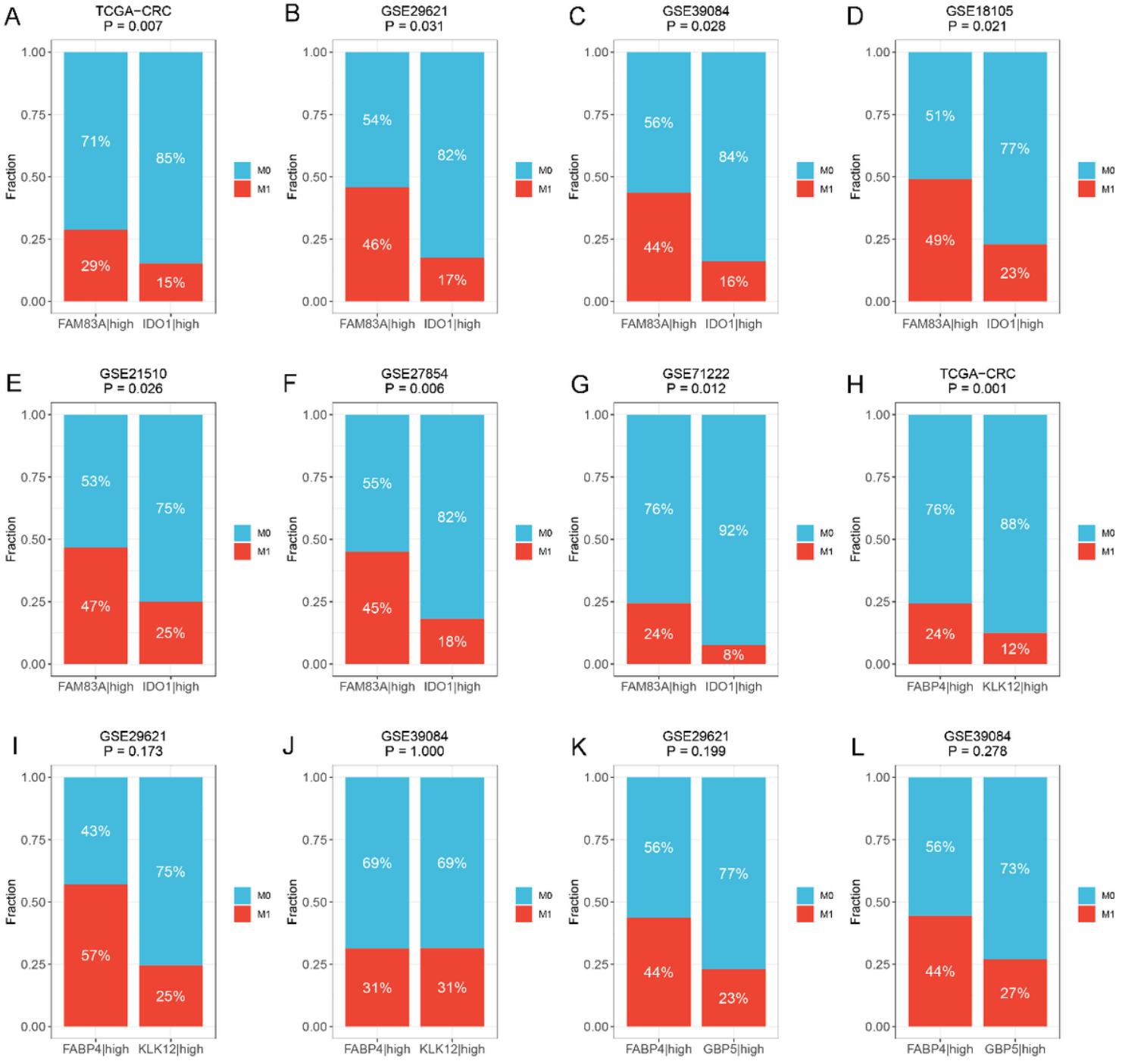

Figure 8

The predictive ability of three prognostic relevant gene pairs for distant metastasis. A-G. The relative proportion of patients with distant metastasis between FAM83A|high and IDO1|high groups in the TCGA-CRC (A), GSE29621 (B), GSE39084 (C), GSE18105 (D), GSE27854 (F), and GSE71222 (G) cohorts. H-J. The relative proportion of patients with distant metastasis between FABP4|high and KLK12|high groups

in the TCGA-CRC (H), GSE29621 (I), and GSE39084 (J) cohorts. K-L. The relative proportion of patients with distant metastasis between FABP4|high and GBP5|high groups in GSE29621 (K) and GSE39084 (L) cohorts. M0, no metastasis; M1, metastasis.

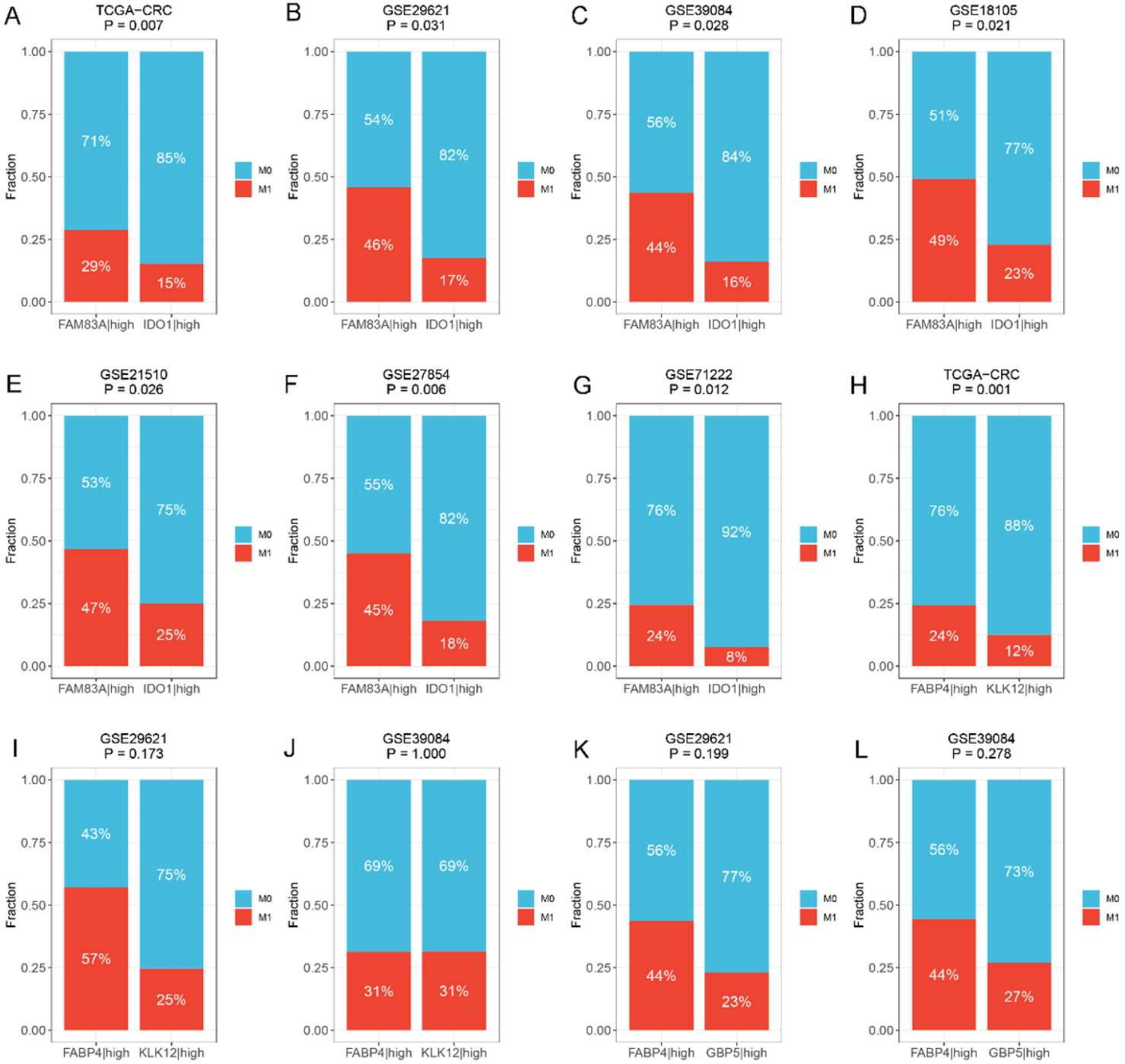

Figure 8

The predictive ability of three prognostic relevant gene pairs for distant metastasis. A-G. The relative proportion of patients with distant metastasis between FAM83A|high and IDO1|high groups in the TCGA-CRC (A), GSE29621 (B), GSE39084 (C), GSE18105 (D), GSE27854 (F), and GSE71222 (G) cohorts. H-J. The relative proportion of patients with distant metastasis between FABP4|high and KLK12|high groups in the TCGA-CRC (H), GSE29621 (I), and GSE39084 (J) cohorts. K-L. The relative proportion of patients

with distant metastasis between FABP4|high and GBP5|high groups in GSE29621 (K) and GSE39084 (L) cohorts. M0, no metastasis; M1, metastasis.

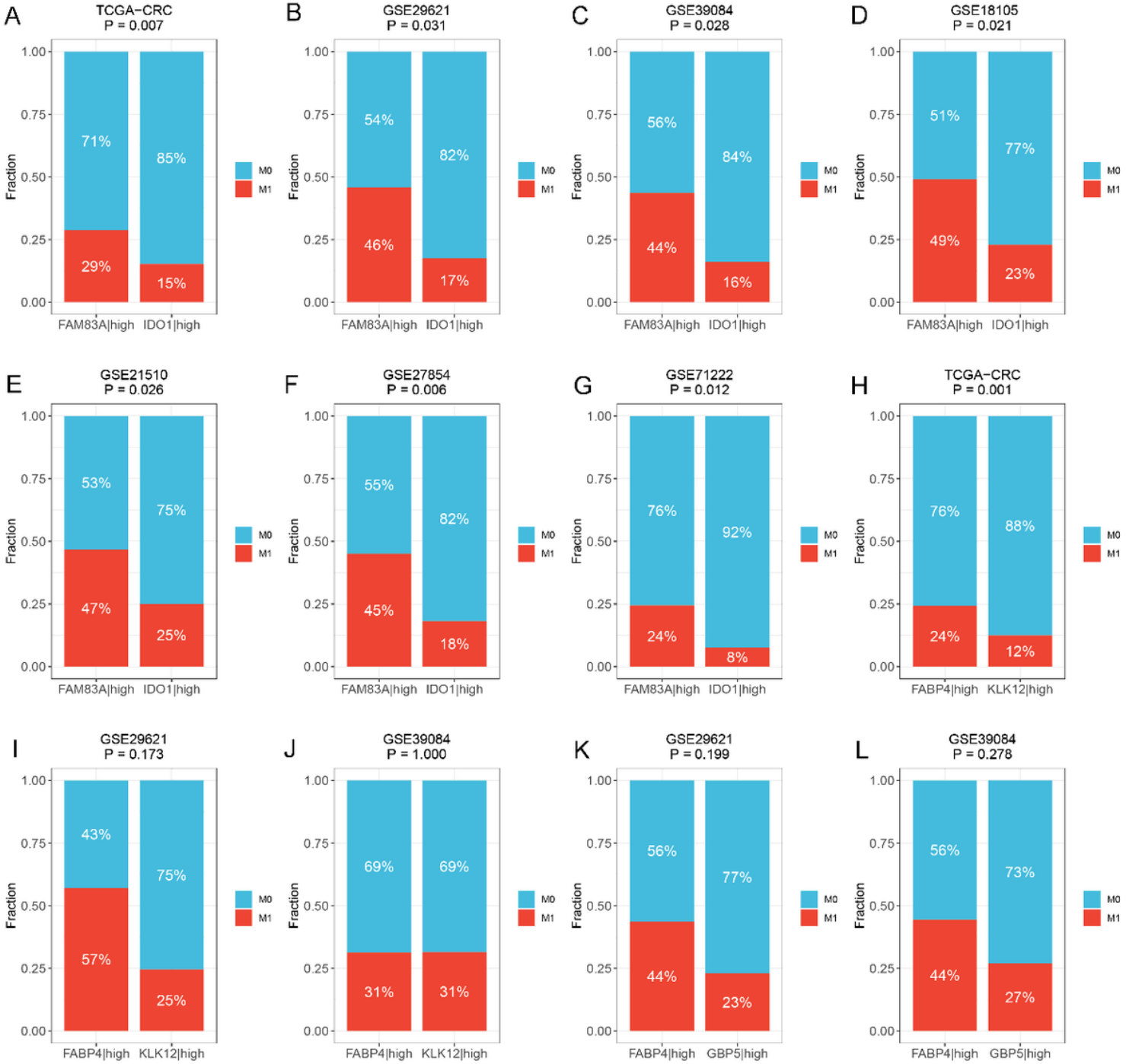

Figure 8

The predictive ability of three prognostic relevant gene pairs for distant metastasis. A-G. The relative proportion of patients with distant metastasis between FAM83A|high and IDO1|high groups in the TCGA-CRC (A), GSE29621 (B), GSE39084 (C), GSE18105 (D), GSE27854 (F), and GSE71222 (G) cohorts. H-J. The relative proportion of patients with distant metastasis between FABP4|high and KLK12|high groups in the TCGA-CRC (H), GSE29621 (I), and GSE39084 (J) cohorts. K-L. The relative proportion of patients

with distant metastasis between FABP4|high and GBP5|high groups in GSE29621 (K) and GSE39084 (L) cohorts. M0, no metastasis; M1, metastasis.

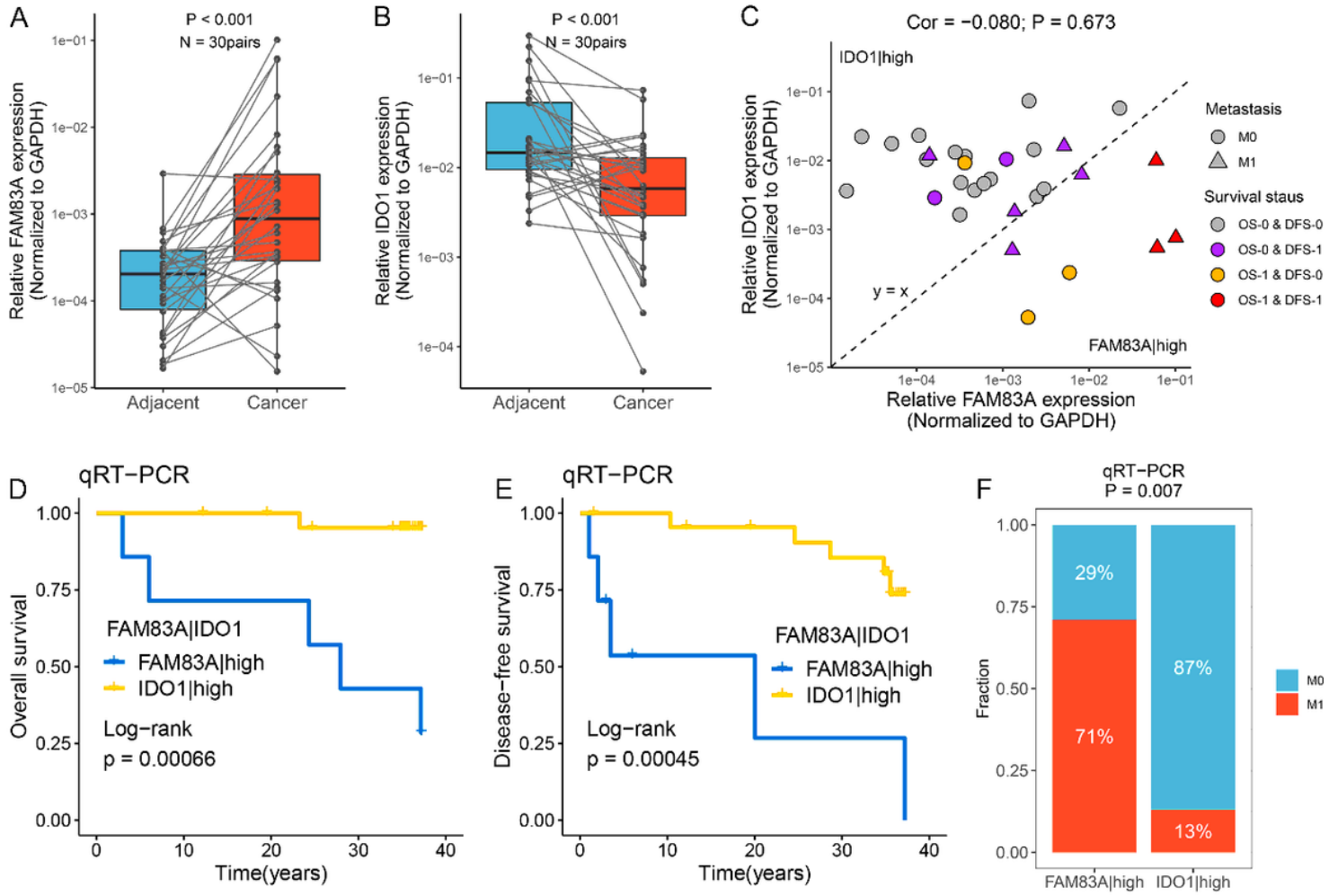

Figure 9

Verified the role of FAM83A|IDO1 in prognosis and metastasis using qRT-PCR. A-B. The expression difference of FAM83A (A) and IDO1 (B) between two subtypes. C. The mRNA expression of FAM83A and IDO1 as well as the clinical outcomes in our cohort. The abscissa is the expression of FAM83A, and the ordinate is the expression of IDO1. Under the line y = x, FAM83A > IDO1, while above it, FAM83A < IDO1. M0, no metastasis; M1, metastasis. OS-0, alive; OS-1, death or censoring; DFS-0, disease free; DFS-1, disease or censoring. D-E. Kaplan–Meier analysis of OS (D) and DFS (E) for FAM83A|IDO1 in our cohort. F. The relative proportion of patients with distant metastasis between FAM83A|high and IDO1|high groups in our cohort.

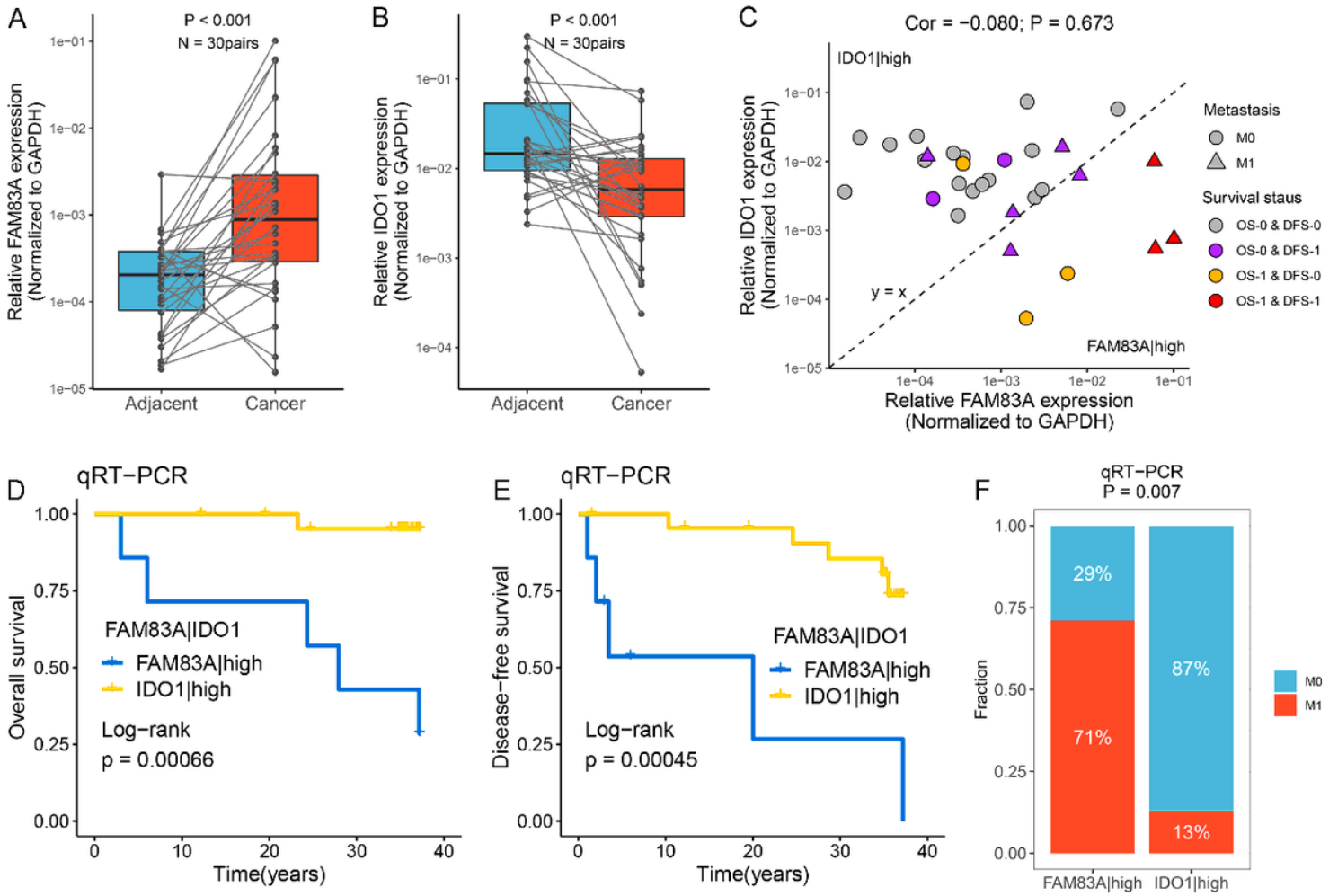

Figure 9

Verified the role of FAM83A|IDO1 in prognosis and metastasis using qRT-PCR. A-B. The expression difference of FAM83A (A) and IDO1 (B) between two subtypes. C. The mRNA expression of FAM83A and IDO1 as well as the clinical outcomes in our cohort. The abscissa is the expression of FAM83A, and the ordinate is the expression of IDO1. Under the line y = x, FAM83A > IDO1, while above it, FAM83A < IDO1. M0, no metastasis; M1, metastasis. OS-0, alive; OS-1, death or censoring; DFS-0, disease free; DFS-1, disease or censoring. D-E. Kaplan–Meier analysis of OS (D) and DFS (E) for FAM83A|IDO1 in our cohort. F. The relative proportion of patients with distant metastasis between FAM83A|high and IDO1|high groups in our cohort.

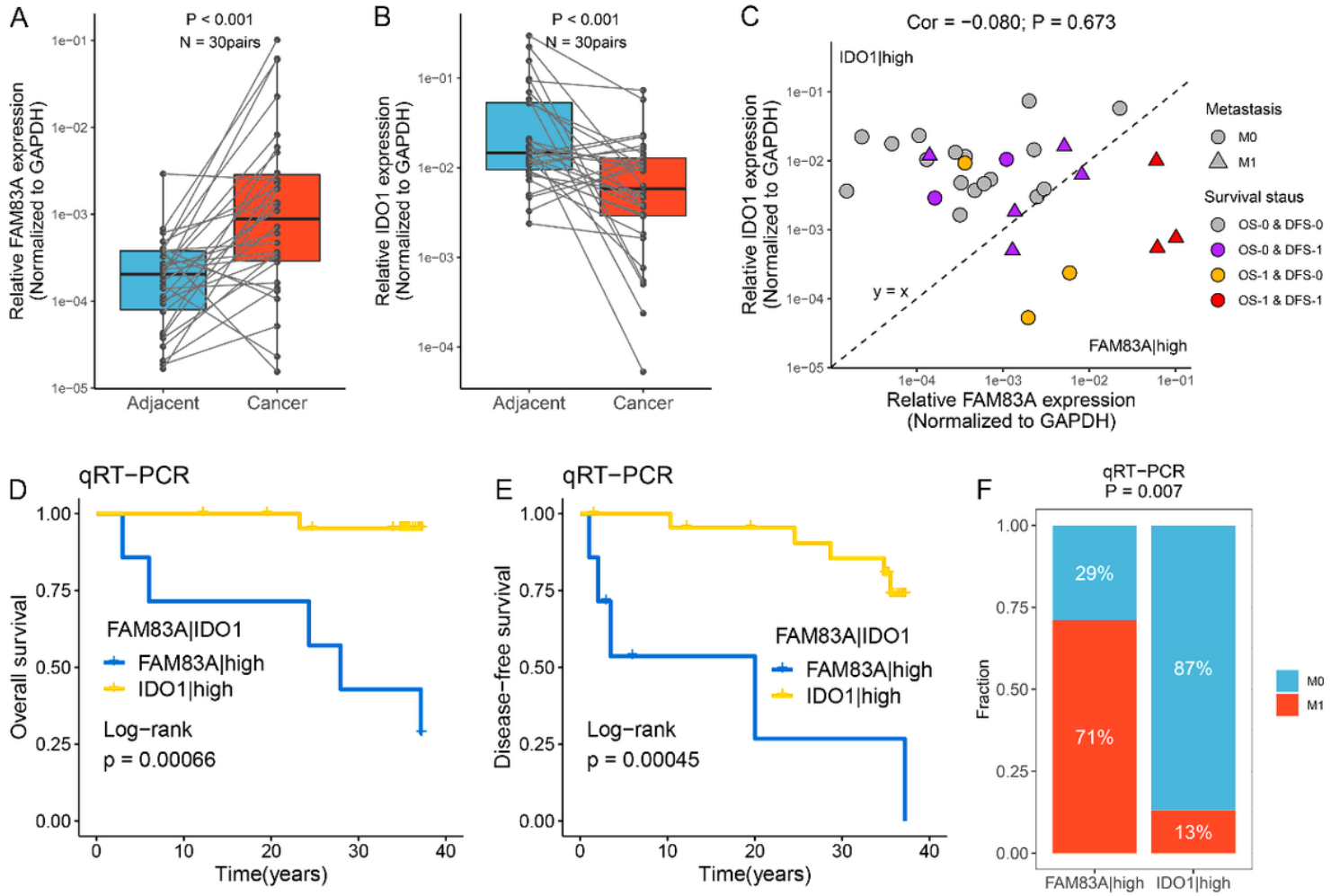

Figure 9

Verified the role of FAM83A|IDO1 in prognosis and metastasis using qRT-PCR. A-B. The expression difference of FAM83A (A) and IDO1 (B) between two subtypes. C. The mRNA expression of FAM83A and IDO1 as well as the clinical outcomes in our cohort. The abscissa is the expression of FAM83A, and the ordinate is the expression of IDO1. Under the line y = x, FAM83A > IDO1, while above it, FAM83A < IDO1. M0, no metastasis; M1, metastasis. OS-0, alive; OS-1, death or censoring; DFS-0, disease free; DFS-1, disease or censoring. D-E. Kaplan–Meier analysis of OS (D) and DFS (E) for FAM83A|IDO1 in our cohort. F. The relative proportion of patients with distant metastasis between FAM83A|high and IDO1|high groups in our cohort.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigureS1.pdf
- FigureS1.pdf
- FigureS1.pdf

- FigureS2.pdf
- FigureS2.pdf
- FigureS2.pdf
- FigureS3.pdf
- FigureS3.pdf
- FigureS3.pdf
- FigureS4.pdf
- FigureS4.pdf
- FigureS4.pdf
- FigureS5.pdf
- FigureS5.pdf
- FigureS5.pdf
- FigureS6.pdf
- FigureS6.pdf
- FigureS6.pdf
- FigureS7.pdf
- FigureS7.pdf
- FigureS7.pdf
- FigureS8.pdf
- FigureS8.pdf
- FigureS8.pdf
- SupplementaryTable.xlsx
- SupplementaryTable.xlsx
- SupplementaryTable.xlsx