A Mutational Signature that Predicts Prognosis and Benefit of Immune Checkpoint Blockade in Colorectal Cancer

Liang Xu (xuliang26@mail2.sysu.edu.cn)
Affiliated Gastrointestinal Hospital of Sun Yat-sen University: Sun Yat-sen University Sixth Affiliated Hospital

Yanyun Lin
Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Department of Colorectal Surgery, Guangdong Institute of Gastroenterology, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

Xijie Chen
Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Department of Colorectal Surgery, Guangdong Institute of Gastroenterology, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

Lisheng Zheng
State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Department of Clinical Laboratory, Sun Yat-sen University Cancer Center, Guangzhou, China

Yufeng Cheng
Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Department of Colorectal Surgery, Guangdong Institute of Gastroenterology, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

Jiancong Hu
Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Department of Colorectal Surgery, Guangdong Institute of Gastroenterology, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

Bin Zheng
Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Department of Colorectal Surgery, Guangdong Institute of Gastroenterology, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

Bin Zhang
Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Department of Colorectal Surgery, Guangdong Institute of Gastroenterology, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

Guanman Li

Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Department of Colorectal Surgery, Guangdong Institute of Gastroenterology, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

Zengjie Chi

Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Department of Colorectal Surgery, Guangdong Institute of Gastroenterology, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

Shuang Guo

Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Department of Colorectal Surgery, Guangdong Institute of Gastroenterology, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

Danling Liu

Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Department of Colorectal Surgery, Guangdong Institute of Gastroenterology, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

Xiaosheng He

Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Department of Colorectal Surgery, Guangdong Institute of Gastroenterology, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

Ping Lan

Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Department of Colorectal Surgery, Guangdong Institute of Gastroenterology, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

---

Research

Keywords: Colorectal cancer, Mutational signature, Immunity, Prognosis

Posted Date: December 11th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-122186/v1

Abstract

Background: Colorectal cancer (CRC) is characterized by broad genomic and transcriptional heterogeneity. However, the genomic basis of this variability remains poorly understood. Our pilot study identified mutated genes were associated with immune infiltration. This study aims to explore a novel mutational signature (MS) in tumor microenvironment (TME) of CRC.

Methods: We integrated single nucleotide variation and transcriptome data and collected corresponding clinicopathologic information from 1,133 and 588 CRC patients of Memorial Sloan Kettering Cancer Center and The Cancer Genome Atlas databases, respectively. Single sample gene set enrichment analysis (ssGSEA) was used to identify the subtypes of CRC based on the immune genomes of 29 immune signatures. CIBERSORT was used to analyze the infiltration of 22 immune cell types in the TME and immune-related gene expression CRC tissues.

Results: In the training cohort, we identified a novel MS consisting of 27 genes and generated a prognostic model that classifies patients into high- and low-risk groups. The low-risk group was associated with better survival and more tumor mutational burden, microsatellite instability, and mismatch repair deficiency. The data were all verified in the validation set. Further analysis revealed that the MS was associated with tumor immunogenicity and immunocyte infiltration, and the determined risk score (RS) could be an index for the immunity level.

Conclusion: We identified a MS that could assist clinicians to select immunotherapy responsive patients and the combination of RS and TNM stage could provide comprehensive prognostic information for CRC.

Background

Colorectal cancer (CRC) is considered as a genetic disease, which arises from the stepwise accumulation of genetic and epigenetic alterations [1, 2]. It is known that these alterations promote the dysplasia and tumorigenesis of CRC [3, 4], but the genomic basis of this variability remains poorly understood [5]. TNM staging system is currently regarded as the standard for the staging of patients with CRC [6, 7], but it is limited by variations among patients with the same tumor stage. Previous studies have shown that treatment response and survival rate of CRC patients depend not only on tumor staging but also on heterogeneous and epigenetic molecular features [8–10]. Thus, it is significant to identify an effective system that better predicts overall survival (OS) and treatment selection of patients with CRC.

Although profiling studies have been carried out to identify patterns of gene and which might predict CRC survival and recurrence [11–14], expression profile is greatly influenced by physiological and pathological conditions that lead to poor reproducibility in the process of library construction. Moreover, exaction to RNA samples tends to degrade, leading to deviations and inaccuracies in the results of subsequent analyses. Consequently, RNA expression profiles are rarely used to predict patient survival and devise medication plans. In contrast, DNA is stable during extraction and detection. Furthermore, structural changes of amino acids and proteins caused by mutations play a significant role in tumor progression

[15]. Thus, further analysis and validation in larger, independent cohorts in combination with mutated genes to predict prognosis are essential prior to application in a clinical setting.

The somatic mutations in a tumor are caused by multiple mutational processes [16, 17]. Different mutational processes often generate distinct combinations of mutation types, termed a ‘signature’ [18]. The effect of such mutation processes can be modeled by a mutational signature, of which two different conceptualizations exist, as in the models introduced by Alexandrov et al. [19] and Shiraishi et al.[20]. Glowing evidence have revealed that tumor mutational burden (TMB) is correlated with response to programmed cell death 1 (PD-1) inhibitior and programmed cell death ligand 1 (PD-L1) in tumor microenvironment (TME), which is so called immune checkpoint blockade [21]. Therefore, an understanding of cancer characteristics based on TMB and mutational signature could provide new insights into mutation-driven tumorigenesis and progression (22).

Reliable prognostic biomarkers are needed to select patients at high-risk for a poor prognosis. It is particularly important to identify mutated genes that are associated with the immune microenvironment to achieve more precise application of immunotherapy. In our study, we constructed a predictive model characterized by mutated genes alone to estimate patient survival outcomes and immunity levels. Furthermore, we characterized 27 mutated genes at the molecular level and identified the mutational signature associated with the TMB, microsatellite instability (MSI), and mismatch repair deficiency (dMMR) in CRC. Gene set enrichment analysis (GSEA) revealed how these genes are involved in the immune response, and further analysis indicated that the risk scores represents an index of the immunity level. Thus, this model could offer opportunities to stratify CRC patients for optimal treatment plans based on genomic subtyping.

Methods

Patients and datasets

This study used data from two separate cohorts: the MSKCC cohort [23] as training cohort and the Cancer Genome Atlas (TCGA) cohort as validation cohort. The training cohort, MSKCC, contains data of 1133 colorectal adenocarcinomas samples. The simple nucleotide variation (SNV) and corresponding clinical information were collected from the cBioPortal (https://www.cbioportal.org/study?id=crc_msk_2017) [24]. The validation cohort, TCGA, contains 596 CRC tissues. The transcriptome expression profiles, simple nucleotide variation and clinical data were downloaded from the Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov/). The expression data was HTSeq-FPKM and count type. The detailed clinical information and molecular features are listed in supplementary material, Table S1.

Construction and validation of the prognostic mutational signature

The SNV data of the MSKCC cohort was analyzed using the MSK-IMPACT, a capture-based next-generation sequencing platform that can detect somatic mutations. As result, about 521 mutated genes were selected as potential candidate genes to construct a prognostic signature. To minimize the over-fitting risk, a penalized regression analysis was applied to construct a prognostic model. The LASSO algorithm test within the R package "glmnet" was used for variable selection and shrinkage, while the Cox model was used for the coefficient determination of the prognostic model. The generated prognostic model was applied to calculate the CRC risk-score of patients in both training and validation set. In order to find out the optimal cutoff value between low or high-risk subgroups, an optimal ROC cutoff was used in the training dataset. The point representing optimal ROC cutoff was chosen. The Kaplan-Meier survival curve with the log-rank test and ROC curve were applied to evaluate the predicting power of the model in training and validation cohorts using R packages “survival” and “survivalROC”. To examine the prognostic value of our risk-score model as an independent factor, the predicted risk-score and clinical characteristics selected with univariate model were analyzed using multivariate Cox model. The significant variables remained in the final Cox model were used to establish a prognostic nomogram using the “rms” package in R. Calibration curves for three-year and five-year were plotted to conform the predicted accuracy of the nomogram model.

Functional annotation and enrichment analyses

GO and KEGG pathway analyses were conducted for selected genes within the mutational signature using the R package "clusterProfiler" [25], and results were visualized by “GOplot” R package. To compare the potential immune mechanisms between the subgroups, single-sample Gene Set Enrichment Analysis (ssGSEA) was performed for both low and high-risk groups using the “gsva” R package [26].

Identification of mutational landscape

To determine whether the mutational landscape is different between the low and high-risk subgroups. We downloaded the SNV data of two cohorts from datasets. The r package "maftools" was used to visually delineate the mutational landscape for low and high-risk groups [27].

Evaluation of immune cell infiltration

The relationship between signature and immune microenvironment was examined as follow: Based on the ssGSEA result, the dataset of MSKCC cohort and TCGA were clustered into three robust clusters representing low, medium, and high immunity using R package "sparcl". The cluster distributions of low and high-risk were compared with each other. Next, the immune scores of different immune microenvironment states were calculated using the ESTIMATE algorithm test within the R package "estimate" [28]. The scores of each subgroups were compared with each other. Finally, the relative proportions of 22 different infiltrating immune cell types were estimated using normalized transcriptome data and CIBERSORT algorithm as described before [29]. CIBERSORT is a biology tool that uses the deconvolution method to analyze bulky gene expression data of 22 immune cell types. Different infiltrating immune cells between low and high-risk groups were identified and considered further research.

Statistical analysis

Statistical analysis was conducted using the statistical packages for R software (version 3.4.0). Means with standard deviations (SD) was calculated for continuous values, while frequencies were determined for categorical values. The significance between two different groups composed of continuous values was examined using Student-t tests or Wilcoxon rank-sum test, while the significance between the rates of different groups was determined using Chi-squared test. Kaplan-Meier analyses with the log-rank test were performed to access the difference between the survival rates of each group. Multivariate analysis was conducted for the variables, which were significantly associated with disease free survival in the univariate Cox regression model univariate analyses, using the Cox proportional hazards regression model. Hazard ratios and 95% confidence intervals were provided for the multivariate analysis. For functional annotation and enrichment analysis, FDR (false discovery rate) < 0.05 was selected as the threshold to identify significant terms or pathways. For the CIBERSORT algorithm test, only cases with CIBERSORT p-value < 0.05 were included in the corresponding analysis. Unless specially mentioned, all statistical tests were two sides and p-value < 0.05 were considered as statistically significant.

Results

Flowchart of mutational signatures stabilization and clinical characteristics.

A flowchart demonstrated the procedure how to analyze the mutational signatures and their correlation with clinical characteristics and immunity, as well as prognostic models based on the mutated genes is shown in Fig. 1. This study obtained data from two cohort studies, Memorial Sloan Kettering Cancer Center (MSKCC) database (n = 1,133) and The Cancer Genome Atlas (TCGA) database (n = 588), respectively. The training cohort, MSKCC cohorts, comprised 729 male patients (64.34%) and 404 female patients (35.66%), while the validation cohort, TCGA cohort, comprised 309 male patients (52.55%) and 279 female patients (47.45%). The clinical characteristics of all CRC patients were listed in Supplementary Table S1.

Construction and verification of the mutational signature prognostic classifier.

A mutational signature including 521 genes was identified to be related to survival in CRC by MSKCC cohort, and 27 mutated genes which were included in the classifier were identified by least absolute shrinkage and selection operator (LASSO) analysis (Fig. 2A and B). The coefficients of the 27 mutated genes were shown in Table S2. Using Kaplan-Meier analysis, the mutational signature effectively stratified CRCs into high- and low-risk groups, and the classified high-risk group showed a poorer OS compared with the low-risk group, which verified in both MSKCC and TCGA cohorts (Fig. 2E and F). The

receiver operating characteristic (ROC) curves indicated that the classifier had a strong predictive ability, as the area under the curve (AUC) values for 1-, 3-, and 5-year OS were 0.712, 0.670, and 0.682, respectively (Fig. 2G). To assess the predictive power of the classifier, we compared the area under ROC curves between the classifier and risk score, tumor location, M stage, and TNM stage. The result showed that the classifier had a better predictive power and accuracy than other clinical features (Fig. 2H). In multivariate Cox analysis of both the MSKCC and TCGA cohorts, the classifier was identified as an unfavorable prognostic factor (Fig. 2I and J). The results in the univariate and multivariate analyses of prognostic factors shown in Table 1 revealed mutational signature is an independent, unfavorable prognostic indicator for CRC in both the MSKCC and TCGA cohorts.

Construction of a predictive nomogram in CRC.

Using the data of the training cohort, a nomogram was generated to predict the OS (Fig. 3A). The predictors included tumor location, M stage, TNM stage, and risk score, among which the risk score had the highest C-index. The calibration plots for the 3- and 5-year OS were well predicted in the training cohort (C-index = 0.666) and the validation cohort (C-index = 0.689) (Fig. 3B and C). The predictive power of the nomogram comprising the mutational signature was compared to clinicopathological risk factors using ROC analysis. The result indicated that the OS was more accurately predicted by the nomogram than by the risk factors in both cohorts (Fig. 3D).

Decision curve analysis was used to quantify clinical application by net benefits at different threshold probability in our nomogram model. Here we found that threshold probabilities of 0 ~ 0.43 and 0 ~ 0.65 were the most beneficial for predicting 3- and 5-year OS, respectively (Supplementary Fig. S1A). Gene ontology (GO) analysis based on the 27 mutated genes demonstrated that mutated genes were mainly enriched in protein binding, beta-catenin binding, nucleus, nucleoplasm, and beta-catenin destruction complex assembly (Supplementary Fig. S1b). Kyoto Encyclopedia of Genes and Genomes (KEGG) functional enrichment analyses revealed that these 27 mutated genes were associated with endometrial cancer, acute myeloid leukemia, pathways in cancer, signaling pathways regulating pluripotency of stem cells, CRC, ErbB signaling pathway, prostate cancer, and thyroid cancer (Supplementary Fig. S1C).

Mutational landscape of significantly mutated genes in defined high- and low-risk subgroups.

To explore the differences of genomic alterations between the defined high- and low-risk groups, we analyzed the data containing somatic mutations from TCGA (https://portal.gdc.cancer.gov/) database. First, comparison according to mutation frequency revealed a significant enrichment of different mutations between high- and low-risk groups (Fig. 4A and E). The most frequently found mutation types were missense mutations, nonsense mutations and frameshift deletions. Analyzing the mutation frequency of both subgroups, a larger number of mutations were found in the low-risk group as compared

with the high-risk group (Fig. 4A). More than 95% genes had a higher mutation rate in the low-risk group as compared with the high-risk group (Supplementary Fig. S2A and B). In addition, the high- and low-risk groups had a significant different distribution of the top 10 mutated genes (Fig. 4C and G). These results suggested that there were significant differences in the mutated genes between the high- and low-risk groups.

A significant enrichment of oncogenic alterations in such genes as BRAF, ZFHX3, SOX9 and MTOR were found in right-sided primary tumors, while oncogenic alteration of APC was primary found in the left-sided primary tumors. Analyzing mutated genes in MSI and MSS CRC patients, it revealed a higher altered frequency of genes including BRAF, TCF7L2, ZFHX3, MTOR and DNMT1 in MSI patient group as compared with MSS patient group, though a significant enrichment of oncogenic alteration in APC gene was found in MSS patients (Fig. 4D and H).

We observed significantly higher TMB in the low-risk group as compared with the high-risk group. Since mutational signatures are significantly correlated with TMB, which is positively correlated with tumor immune signatures and immunotherapy response, it can be speculated that mutational signature may be related to tumor immune activity and further affect immunotherapy response.

The mutational signature was associated with the genomic features of MSI and dMMR in CRC.

The MSI status is critical when considering immunotherapy and chemotherapeutic drugs as options for CRC patients [30]. MMR is the process by which potentially mutagenic misincorporation errors that occur during normal DNA replication are corrected and the absence of MMR results in increased accumulation of mutations [31, 32]. To further characterize the classified risk groups, we examined the association between the defined risk groups and other clinical characteristics using data of patients from both training and validation cohorts. The result showed that the outcome of the risk score was highly correlated with tumor location, hyper-mutation, MSI status, and TNM stage (Table 2). The risk score was observed to be significantly associated with the status of MSI/dMMR in both training and validation cohorts (Fig. 5A and B). In line with previous observation, the status of MSI was more common in low-risk group as compared with high-risk group (Fig. 5C). In addition, left-side tumors and TNM stage III-IV were more common in the high-risk group compared with the low-risk group (Supplementary Fig. 1A and B).

In TNM stage III-IV patients, the risk score was much higher than in that of patients with TNM stage I-II (Supplementary Fig. S1C). In addition, the defined low-risk group was significantly associated with hyper-mutation (Fig. 5D). Furthermore, we observed that high-risk group exhibited significantly higher rate of low mutation than low-risk group (Fig. 5E), and the number of mutations was higher in the low-risk group than that in the high-risk group (Fig. 5F). These results showed the mutational signature was associated with TMB, MSI status, and dMMR in CRC. It demonstrated a potential value of this mutation signature model for characterization of immune environment and prediction of immunotherapy outcome.

Mutation signature was associated with immune activity by immunogenic profiling identification.

To further clarify the relationship between mutational signature and immune-phenotyping, we analyzed single nucleotide variation (SNV) and transcriptome data in TCGA database. Immune activity differences between the high- and low-risk groups were determined by analyzing 29 immune-associated gene sets, which represented diverse immune cell types, functions, and pathways [33]. These gene sets were analyzed using the single sample gene set enrichment analysis (ssGSEA), an extension of GSEA, which could calculate separate enrichment scores for each pairing of a sample and gene set to quantify the activity or enrichment levels of immune cells, functions, or pathways in cancer samples [34].

On the basis of ssGSEA scores, we hierarchically clustered all CRC samples in TCGA dataset, and defined the three clusters as Immunity-High, Immunity-Medium, and Immunity-Low. Tumor purity, stromal score, and immune score were analyzed for each CRC sample based on the Estimation of Stromal and Immune cells in Malignant Tumor tissues using Expression data (ESTIMATE) algorithm. A heatmap of the infiltration levels and scores of each sample of immune cells in the three subtypes was shown in Fig. 6A. The results showed that the stromal score was significantly higher in the Immunity-High cluster and significantly lower in the Immunity-Low cluster. Immunity scores and ESTIMATE scores were gradually reduced from Immunity-H cluster to Immunity-L cluster. However, the opposite trend was observed for tumor purity in comparisons of the three CRC subtypes (Fig. 6A). Principal component analysis (PCA) revealed marked differences between the three clusters (Fig. 6B), indicating that Immunity-H cluster contained the largest number of immune cells and stromal cells, while Immunity-L cluster contained the largest number of tumor cells (Fig. 6D). Furthermore, we found significant higher expression of most human leukocyte antigen (HLA) genes in Immunity-H cluster as compared with Immunity-L cluster (Fig. 6C). Moreover, a significantly higher expression level of PD-L1 gene was found in Immunity-H cluster while Immunity-H cluster was correlated with better survival outcome as compared with Immunity-L cluster (Fig. 6E and 6F). According to distribution of these three clusters, we found that Immunity-H cluster was significantly enriched in the low-risk group (Fig. 6G), indicating that patients in the low-risk group might benefit more from PD-L1 inhibitor treatment.

To assess whether risk score was highly correlated with the immunity, we performed consensus molecular subgroups (CMS) classification [35], which give a more profound biological insight into metastatic CRC carcinogenesis, immunity typing, and has a strong prognostic effect. CMS1 is defined by an upregulation of immune genes and is highly associated with microsatellite instability (MSI-h) [36], while CMS4 is defined by an activated tissue growth factor (TGF)-β pathway and by epithelial-mesenchymal transition (EMT) making it in general more chemo-resistant. As expected, patients were divided into four clusters (Fig. 6H), and the distribution analysis revealed that the CMS4 was significantly more representable for the high-risk group as compared to CMS1, while the low-risk group was more represented by CMS1 subgroup (Fig. 6J). In addition, a significantly higher expression level of PD-L1 gene was found in the CMS1 subgroup as compared with other CMS subgroups (Fig. 6I). Taken together, the

data suggested that the low-risk group was mainly represented by CMS1 and had a higher PD-L1 expression, which might benefit more from PD-L1 inhibitor treatment.

Composition of immune cell profiles in the high- and low-risk groups.

To further investigate the potential predictive value of the mutational signature for the immune status, we examined possible associations between the risk score and immune status. The risk score was negatively correlated with the TMB score (Fig. 7A) while TMB was positively correlated with immune score (Fig. 7B). Therefore, we postulated that the risk score was negatively associated with immune score. Since a high immune score was related with a better survival outcome (Fig. 7C), the low-risk group may also be associated with a better survival.

To Fig. out the infiltrated immune cell composition in the defined risk groups, we analyzed the expression signature matrix of 22 infiltrated immune cell types in tumor samples from the TCGA cohort using the CIBERSORT test. Among the total samples, 63 low-risk and 63 high-risk samples were found to be eligible for further analysis. The different immune cell fractions were weakly correlated with each other in tumor tissues in the TCGA cohort (Fig. 7D and 7E). Regarding to tumor-infiltrating immune cells in CRC microenvironment, reduced number of activated CD4 + memory T cells, but increased number of macrophages M0 were found in the high-risk group (Fig. 7F) as compared with the low-risk group. Finally, we analyzed the pathways that were significantly enriched in the high- and low-risk groups and found an enriched immunologic pathway in the high-risk samples (Fig. 7G). Taken together, these data suggested that mutational signature consisted of genes that are important regulatory components of the immune cell activation mechanisms. The predictive power of the mutational signature for OS might be dependent on the immune status of TME.

Discussion

CRC is considered as a genetic disease, which arises from the stepwise accumulation of genetic and epigenetic alterations that might promote the dysplasia and tumorigenesis of CRC. Advances in molecular biology stimulate the generation of large amounts of data that were used to construct multigene profiles, which can be used for risk stratification and guidance for chemotherapy treatment in various types of cancers [37–40]. Therefore, exploring the dysregulated genes involved in carcinogenesis and disease development might help to improve prognostic and therapeutic strategies for CRC patients.

In this study, we generated a novel prognostic model based on a mutational signature classifier to predict the CRC overall survival and the efficacy of immunotherapy. The mutational signature classifier consisted of 27 mutated genes including APC and TCF7L2 that are relevant to the WNT signaling pathway and influence the cancer cell metastatic ability [41–43], and BRAF and NRAS that are involved in the EGFR signaling pathway and associated with drug-resistance [44–47]. Using the generated prognostic model,

the CRC patients from the cohorts were categorized into high- and low-risk groups. Comparing the global heterogeneity between high- and low-risk groups, the risk score was shown to be correlated with known predictive factor for the carcinogenesis and the therapy outcome such as the TNM stage, MSI status, hyper-mutation, and TMB. Furthermore, we demonstrated that the nomogram comprising the identified mutational signature classifier could better predict the OS as compared to clinic-pathological risk factors. This may due to the property of the mutational signature, which reflects the biological heterogeneity of these tumors. This new nomogram including the mutational signature might provide a simple and accurate method for predicting prognoses in CRC.

The immune status within TME plays a pivotal role during the tumorigenesis, and the immune response is a complex process in which various immune cells interact and play different roles. Studies have showed that the immune status could be a better prognostic predictor than the TNM stage, since the tumor progression was significantly dependent on the density of host cytotoxic- and memory T cell, higher density of these T cells was correlated with better survival outcome [44–47]. Tumor infiltrating lymphocytes (TILs) are immune cells leave the blood stream and migrate towards tumor; this population of immune cells contains T cells, B cells, and NK cells, which could exhibit anti-tumor functions. Some studies revealed that TILs are highly heterogeneous in intra-tumor and para-tumor areas [51, 52], and could be associated with prolonged survival [53]. Therefore, better understanding of the immune status of the TME and exploring the distribution and function of immune cells are critical to improve the efficacy of immunotherapy in cancer.

In our study, significant higher amount of activated CD4 + memory T cells among TILs was found in the low-risk group. Further classification, previous data revealed that activated CD4 + memory T cells were mainly detected in early stage of tumor progression [54]. Since the low-risk group was correlated with a better survival outcome, it is reasonable to suggest this activated CD4 + memory T cell population may exhibit an anti-tumor effect in early stage of CRC.

Immunotherapy has raised as a novel effective treatment against CRC, however, the current standard therapeutic guidelines based on the TNM stage cannot reflect the information of host immune system response. In clinic, tests are required before taking immunotherapy, and only when certain conditions and levels are met might immunotherapy drugs be effective in patients. At present, the most commonly used clinical detection method is MSI status; the higher the degree of microsatellite instability, the more genetic errors the patient shows, and the greater the mutation load, which in turn triggers attack of the immune system on tumor cells. However, only 40% of CRC patients with MSI-H can benefit from immunotherapy. Our prediction model can not only predict the prognosis of CRC, but also further evaluate immune infiltration, accurately screen the population of immunotherapy subjects, and improve the efficacy of immunotherapy.

Conclusions

This study presents a novel valuable mutational signature that could be used to predict OS and support immunotherapy selection, but several limitations still need to be noted. First, our data were obtained from the MSKCC and TCGA datasets, which was not multicenter cohorts, our mutational signature and nomogram require further validation in prospective studies and multicenter clinical trials. Second, the TCGA database mainly contains data from people of European descent, which means that the result from these data cannot be directly extrapolated to other racial groups. Third, the biological contribution of candidate genes of the mutational signature, such as ZFHX3 and DNTM1, to OS remain unclear. Further investigations to elucidate the biological mechanisms of these genes might provide novel targets and treatment strategies.

Despite these limitations, we have identified a novel mutational signature, which can generate a prognostic tool to effectively classify CRC patients into groups with different OS risks. Moreover, the mutational signature classifier can be used to predict the effective patients to immunotherapy and nomogram comprising the mutational signature could help clinicians in directing personalized therapeutic regimen selection for patients with advanced CRC.

Declarations

Authors' contributions

Study concepts: P.L., XS.H., L.X.

Study design: P.L., XS.H., L.X.

Literature research: YY.L., XJ.C.

Data acquisition: L.X., YY.L., XJ.C.

Data analysis/interpretation: YY.L.,GM.L.

Statistical analysis: XJ.C., ZJ.C.

Manuscript preparation: L.X., B.Z.

Manuscript definition of intellectual content: LS.Z., B.Z.

Manuscript editing: YF.C., JC.C., S.G., DL.L.

Author details

[1]Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Department of Colorectal Surgery, Guangdong Institute of Gastroenterology, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China. [2]State Key Laboratory of Oncology in South China, Collaborative Innovation Center

for Cancer Medicine, Department of Clinical Laboratory, Sun Yat-sen University Cancer Center, Guangzhou, China.

Acknowledgments

Not applicable

Conflicts of interest

The authors declare that they have no competing interests.

Availability of data and materials

Due to ethical restrictions, the raw data underlying this paper are available upon request to the corresponding author.

Consent for publication

Not applicable.

Funding

This work was supported by grants from the National Key R&D Program of China (No. 2017YFC1308800), the Sun Yat-sen University 5010 Project (N0.2010012), the National Natural Science Foundation of China (32000555), and the Natural Science Foundation of Guangdong Province (Z20190107202209743, Z20190115202112705).

Ethics approval and consent to participate

Not applicable.

References

1. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LJ, Kinzler KW: **Cancer genome landscapes.** *Science* 2013, **339:**1546-1558.
2. Bolli N, Avet-Loiseau H, Wedge DC, Van Loo P, Alexandrov LB, Martincorena I, Dawson KJ, Iorio F, Nik-Zainal S, Bignell GR, et al: **Heterogeneity of genomic evolution and mutational profiles in multiple myeloma.** *Nat Commun* 2014, **5:**2997.

3. Okugawa Y, Grady WM, Goel A: **Epigenetic Alterations in Colorectal Cancer: Emerging Biomarkers.** *Gastroenterology* 2015, **149:**1204-1225.
4. Lao VV, Grady WM: **Epigenetics and colorectal cancer.** *Nat Rev Gastroenterol Hepatol* 2011, **8:**686-700.
5. Grady WM, Carethers JM: **Genomic and epigenetic instability in colorectal cancer pathogenesis.** *Gastroenterology* 2008, **135:**1079-1099.
6. Schneider NI, Langner C: **Prognostic stratification of colorectal cancer patients: current perspectives.** *Cancer Manag Res* 2014, **6:**291-300.
7. Puppa G, Sonzogni A, Colombari R, Pelosi G: **TNM staging system of colorectal carcinoma: a critical appraisal of challenging issues.** *Arch Pathol Lab Med* 2010, **134:**837-852.
8. Sagaert X, Vanstapel A, Verbeek S: **Tumor Heterogeneity in Colorectal Cancer: What Do We Know So Far?** *Pathobiology* 2018, **85:**72-84.
9. Joung JG, Oh BY, Hong HK, Al-Khalidi H, Al-Alem F, Lee HO, Bae JS, Kim J, Cha HU, Alotaibi M, et al: **Tumor Heterogeneity Predicts Metastatic Potential in Colorectal Cancer.** *Clin Cancer Res* 2017, **23:**7209-7216.
10. Ma J, Setton J, Lee NY, Riaz N, Powell SN: **The therapeutic significance of mutational signatures from DNA repair deficiency in cancer.** *Nat Commun* 2018, **9:**3292.
11. Matsuoka T, Yashiro M: **Biomarkers of gastric cancer: Current topics and future perspective.** *World J Gastroenterol* 2018, **24:**2818-2832.
12. Xu L, Li X, Chu ES, Zhao G, Go MY, Tao Q, Jin H, Zeng Z, Sung JJ, Yu J: **Epigenetic inactivation of BCL6B, a novel functional tumour suppressor for gastric cancer, is associated with poor survival.** *Gut* 2012, **61:**977-985.
13. Bertoli G, Cava C, Castiglioni I: **MicroRNAs as Biomarkers for Diagnosis, Prognosis and Theranostics in Prostate Cancer.** *Int J Mol Sci* 2016, **17:**421.
14. Lin L, Liu Y, Pan C, Zhang J, Zhao Y, Shao R, Huang Z, Su Y, Shi M, Bin J, et al: **Gastric cancer cells escape metabolic stress via the DLC3/MACC1 axis.** *Theranostics* 2019, **9:**2100-2114.
15. Chen H, Chong W, Yang X, Zhang Y, Sang S, Li X, Lu M: **Age-related mutational signature negatively associated with immune activity and survival outcome in triple-negative breast cancer.** *Oncoimmunology* 2020, **9:**1788252.
16. Kruger S, Piro RM: **decompTumor2Sig: identification of mutational signatures active in individual tumors.** *Bmc Bioinformatics* 2019, **20:**152.
17. Alexandrov LB, Nik-Zainal S, Siu HC, Leung SY, Stratton MR: **A mutational signature in gastric cancer suggests therapeutic strategies.** *Nat Commun* 2015, **6:**8683.
18. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR: **Deciphering signatures of mutational processes operative in human cancer.** *Cell Rep* 2013, **3:**246-259.
19. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, et al: **Signatures of mutational processes in human cancer.** *Nature* 2013, **500:**415-

421.

20. Shiraishi Y, Tremmel G, Miyano S, Stephens M: **A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures.** *Plos Genet* 2015, **11:**e1005657.
21. Havel JJ, Chowell D, Chan TA: **The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy.** *Nat Rev Cancer* 2019, **19:**133-150.
22. Roberts SA, Gordenin DA: **Hypermutation in human cancer genomes: footprints and mechanisms.** *Nat Rev Cancer* 2014, **14:**786-800.
23. Yaeger R, Chatila WK, Lipsyc MD, Hechtman JF, Cercek A, Sanchez-Vega F, Jayakumaran G, Middha S, Zehir A, Donoghue M, et al: **Clinical Sequencing Defines the Genomic Landscape of Metastatic Colorectal Cancer.** *Cancer Cell* 2018, **33:**125-136.
24. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, et al: **The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.** *Cancer Discov* 2012, **2:**401-404.
25. Yu G, Wang LG, Han Y, He QY: **clusterProfiler: an R package for comparing biological themes among gene clusters.** *Omics* 2012, **16:**284-287.
26. Hanzelmann S, Castelo R, Guinney J: **GSVA: gene set variation analysis for microarray and RNA-seq data.** *Bmc Bioinformatics* 2013, **14:**7.
27. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP: **Maftools: efficient and comprehensive analysis of somatic variants in cancer.** *Genome Res* 2018, **28:**1747-1756.
28. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, Trevino V, Shen H, Laird PW, Levine DA, et al: **Inferring tumour purity and stromal and immune cell admixture from expression data.** *Nat Commun* 2013, **4:**2612.
29. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA: **Robust enumeration of cell subsets from tissue expression profiles.** *Nat Methods* 2015, **12:**453-457.
30. Wang Z, Tang X, Wu X, Yang M, Wang D: **Mismatch repair status between primary colorectal tumor and metastatic tumor, a retrospective consistent study.** *Biosci Rep* 2019, **39**.
31. Smith CE, Mendillo ML, Bowen N, Hombauer H, Campbell CS, Desai A, Putnam CD, Kolodner RD: **Dominant mutations in S. cerevisiae PMS1 identify the Mlh1-Pms1 endonuclease active site and an exonuclease 1-independent mismatch repair pathway.** *Plos Genet* 2013, **9:**e1003869.
32. Sinicrope FA, Mahoney MR, Smyrk TC, Thibodeau SN, Warren RS, Bertagnolli MM, Nelson GD, Goldberg RM, Sargent DJ, Alberts SR: **Prognostic impact of deficient DNA mismatch repair in patients with stage III colon cancer from a randomized trial of FOLFOX-based adjuvant chemotherapy.** *J Clin Oncol* 2013, **31:**3664-3672.
33. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N: **Molecular and genetic properties of tumors associated with local immune cytolytic activity.** *Cell* 2015, **160:**48-61.
34. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou YT, Porta-Pardo E, Gao GF, Plaisier CL, Eddy JA, et al: **The Immune Landscape of Cancer.** *Immunity* 2019, **51:**411-412.

35. Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, et al: **The consensus molecular subtypes of colorectal cancer.** *Nat Med* 2015, **21:**1350-1356.

36. Dienstmann R, Vermeulen L, Guinney J, Kopetz S, Tejpar S, Tabernero J: **Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer.** *Nat Rev Cancer* 2017, **17:**79-92.

37. van der Hoeven JJ: **[70-Gene signature as an aid to treatment decisions in early-stage breast cancer].** *Ned Tijdschr Geneeskd* 2017, **161:**D1369.

38. Taneja SS: **Re: Development and Validation of a 24-Gene Predictor of Response to Postoperative Radiotherapy in Prostate Cancer: A Matched, Retrospective Analysis.** *J Urol* 2017, **197:**1264-1266.

39. Rini B, Goddard A, Knezevic D, Maddala T, Zhou M, Aydin H, Campbell S, Elson P, Koscielny S, Lopatin M, et al: **A 16-gene assay to predict recurrence after surgery in localised renal cell carcinoma: development and validation studies.** *Lancet Oncol* 2015, **16:**676-685.

40. Scott DW, Chan FC, Hong F, Rogic S, Tan KL, Meissner B, Ben-Neriah S, Boyle M, Kridel R, Telenius A, et al: **Gene expression-based model using formalin-fixed paraffin-embedded biopsies predicts overall survival in advanced-stage classical Hodgkin lymphoma.** *J Clin Oncol* 2013, **31:**692-700.

41. Ji L, Lu B, Wang Z, Yang Z, Reece-Hoyes J, Russ C, Xu W, Cong F: **Identification of ICAT as an APC Inhibitor, Revealing Wnt-Dependent Inhibition of APC-Axin Interaction.** *Mol Cell* 2018, **72:**37-47.

42. Wenzel J, Rose K, Haghighi EB, Lamprecht C, Rauen G, Freihen V, Kesselring R, Boerries M, Hecht A: **Loss of the nuclear Wnt pathway effector TCF7L2 promotes migration and invasion of human colorectal cancer cells.** *Oncogene* 2020, **39:**3893-3909.

43. Shulewitz M, Soloviev I, Wu T, Koeppen H, Polakis P, Sakanaka C: **Repressor roles for TCF-4 and Sfrp1 in Wnt signaling in breast cancer.** *Oncogene* 2006, **25:**4361-4369.

44. Corcoran RB, Andre T, Atreya CE, Schellens J, Yoshino T, Bendell JC, Hollebecque A, McRee AJ, Siena S, Middleton G, et al: **Combined BRAF, EGFR, and MEK Inhibition in Patients with BRAF(V600E)-Mutant Colorectal Cancer.** *Cancer Discov* 2018, **8:**428-443.

45. Kim AS, Bartley AN, Bridge JA, Kamel-Reid S, Lazar AJ, Lindeman NI, Long TA, Merker JD, Rai AJ, Rimm DL, et al: **Comparison of Laboratory-Developed Tests and FDA-Approved Assays for BRAF, EGFR, and KRAS Testing.** *Jama Oncol* 2018, **4:**838-841.

46. Adams R, Brown E, Brown L, Butler R, Falk S, Fisher D, Kaplan R, Quirke P, Richman S, Samuel L, et al: **Inhibition of EGFR, HER2, and HER3 signalling in patients with colorectal cancer wild-type for BRAF, PIK3CA, KRAS, and NRAS (FOCUS4-D): a phase 2-3 randomised trial.** *Lancet Gastroenterol Hepatol* 2018, **3:**162-171.

47. Van Cutsem E, Lenz HJ, Kohne CH, Heinemann V, Tejpar S, Melezinek I, Beier F, Stroh C, Rougier P, van Krieken JH, Ciardiello F: **Fluorouracil, leucovorin, and irinotecan plus cetuximab treatment and RAS mutations in colorectal cancer.** *J Clin Oncol* 2015, **33:**692-700.

48. Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pages C, Tosolini M, Camus M, Berger A, Wind P, et al: **Type, density, and location of immune cells within human colorectal tumors**

predict clinical outcome. *Science* 2006, **313:**1960-1964.

49. Mlecnik B, Tosolini M, Kirilovsky A, Berger A, Bindea G, Meatchi T, Bruneval P, Trajanoski Z, Fridman WH, Pages F, Galon J: **Histopathologic-based prognostic factors of colorectal cancers are associated with the state of the local immune reaction.** *J Clin Oncol* 2011, **29:**610-618.
50. Pages F, Kirilovsky A, Mlecnik B, Asslaber M, Tosolini M, Bindea G, Lagorce C, Wind P, Marliot F, Bruneval P, et al: **In situ cytotoxic and memory T cells predict outcome in patients with early-stage colorectal cancer.** *J Clin Oncol* 2009, **27:**5944-5951.
51. Ascierto ML, De Giorgi V, Liu Q, Bedognetti D, Spivey TL, Murtas D, Uccellini L, Ayotte BD, Stroncek DF, Chouchane L, et al: **An immunologic portrait of cancer.** *J Transl Med* 2011, **9:**146.
52. Pages F, Galon J, Dieu-Nosjean MC, Tartour E, Sautes-Fridman C, Fridman WH: **Immune infiltration in human tumors: a prognostic factor that should not be ignored.** *Oncogene* 2010, **29:**1093-1102.
53. Fridman WH, Pages F, Sautes-Fridman C, Galon J: **The immune contexture in human tumours: impact on clinical outcome.** *Nat Rev Cancer* 2012, **12:**298-306.
54. Ge P, Wang W, Li L, Zhang G, Gao Z, Tang Z, Dang X, Wu Y: **Profiles of immune cell infiltration and immune-related genes in the tumor microenvironment of colorectal cancer.** *Biomed Pharmacother* 2019, **118:**109228.

Tables

Due to technical limitations, table 1 and table 2 are only available as a download in the Supplemental Files section.

Figures

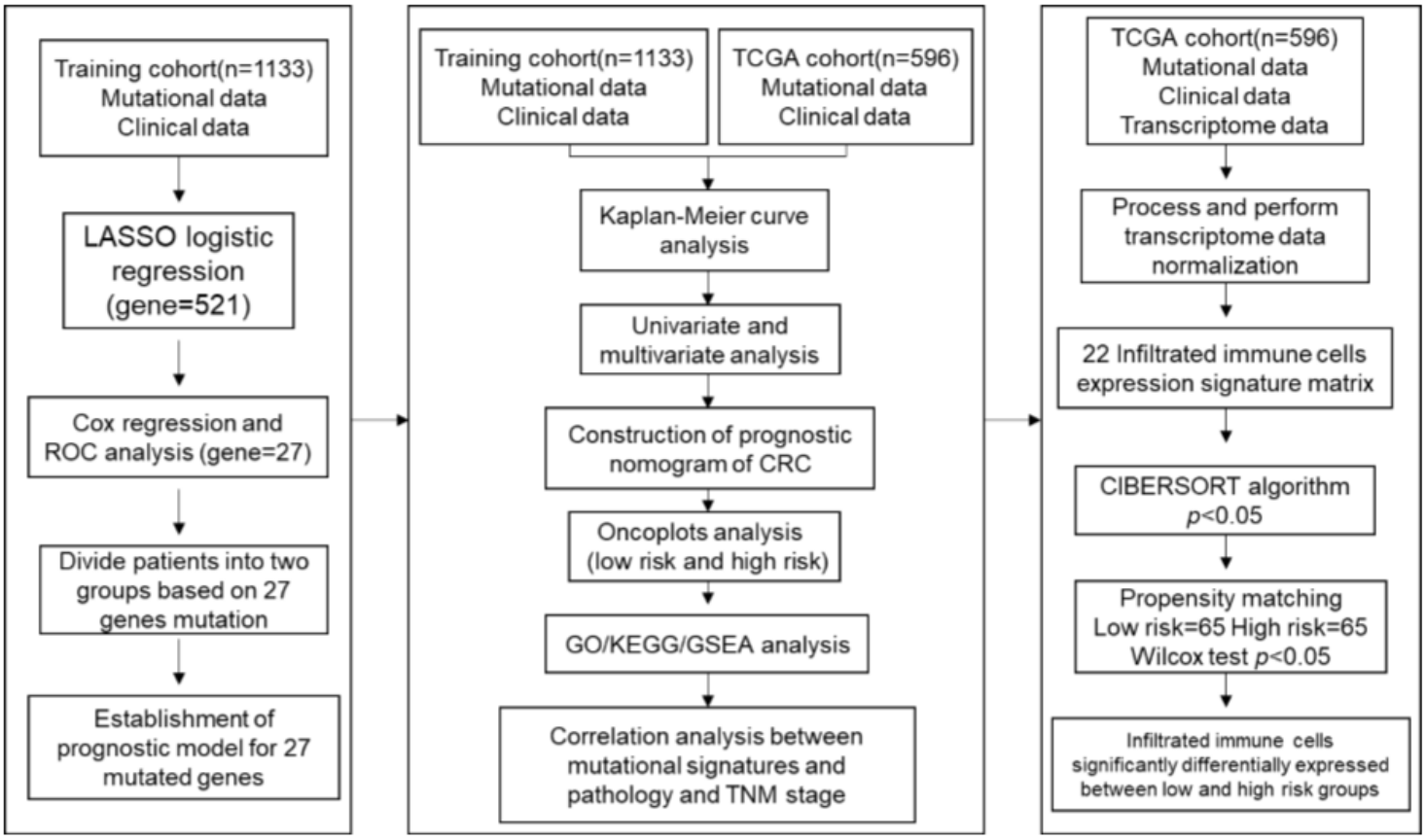

Figure 1

Flowchart detailing the procedure of analyzing mutational signature and their correlation with clinical characteristics and immunity, as well as prognostic models of mutated genes.

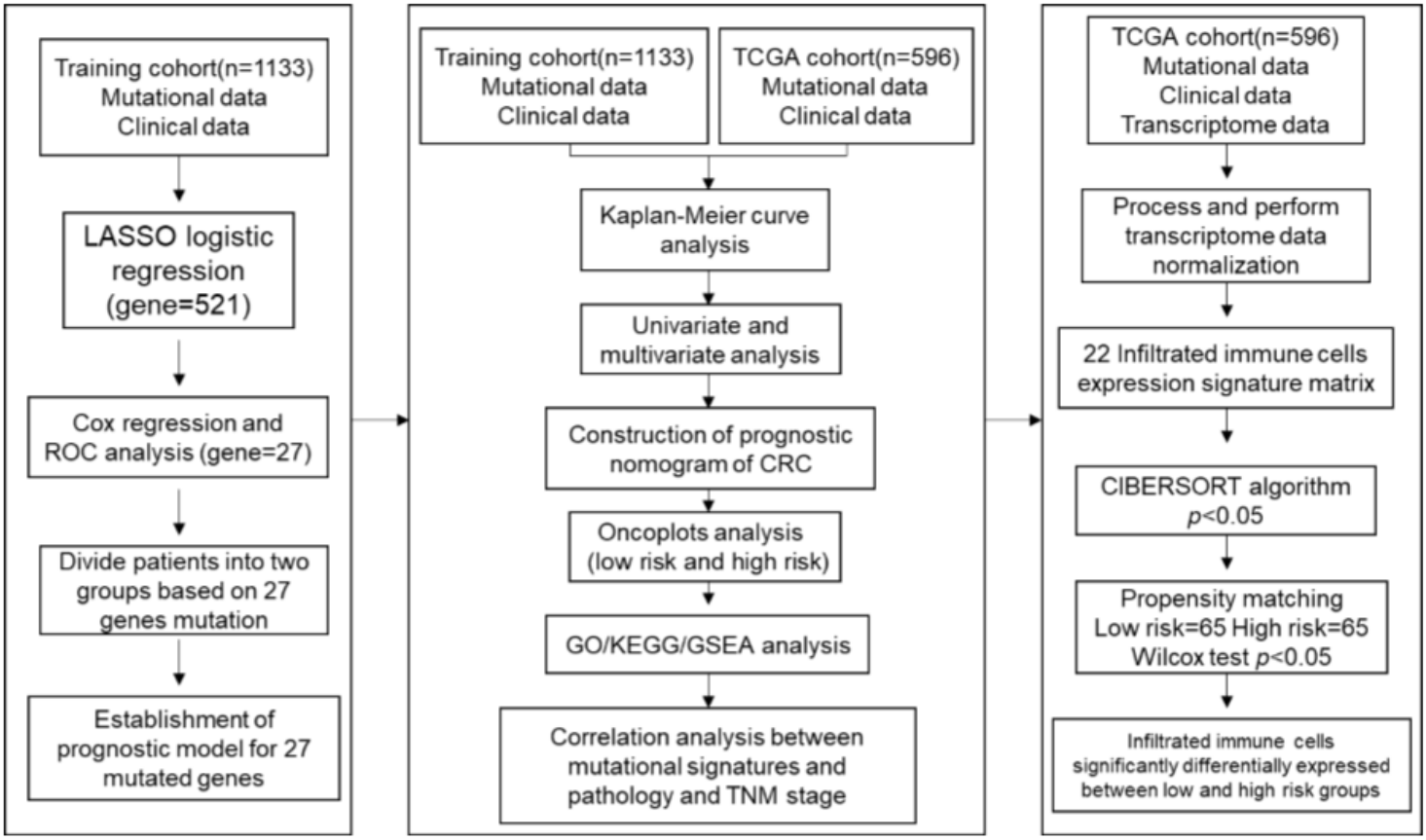

Figure 1

Flowchart detailing the procedure of analyzing mutational signature and their correlation with clinical characteristics and immunity, as well as prognostic models of mutated genes.

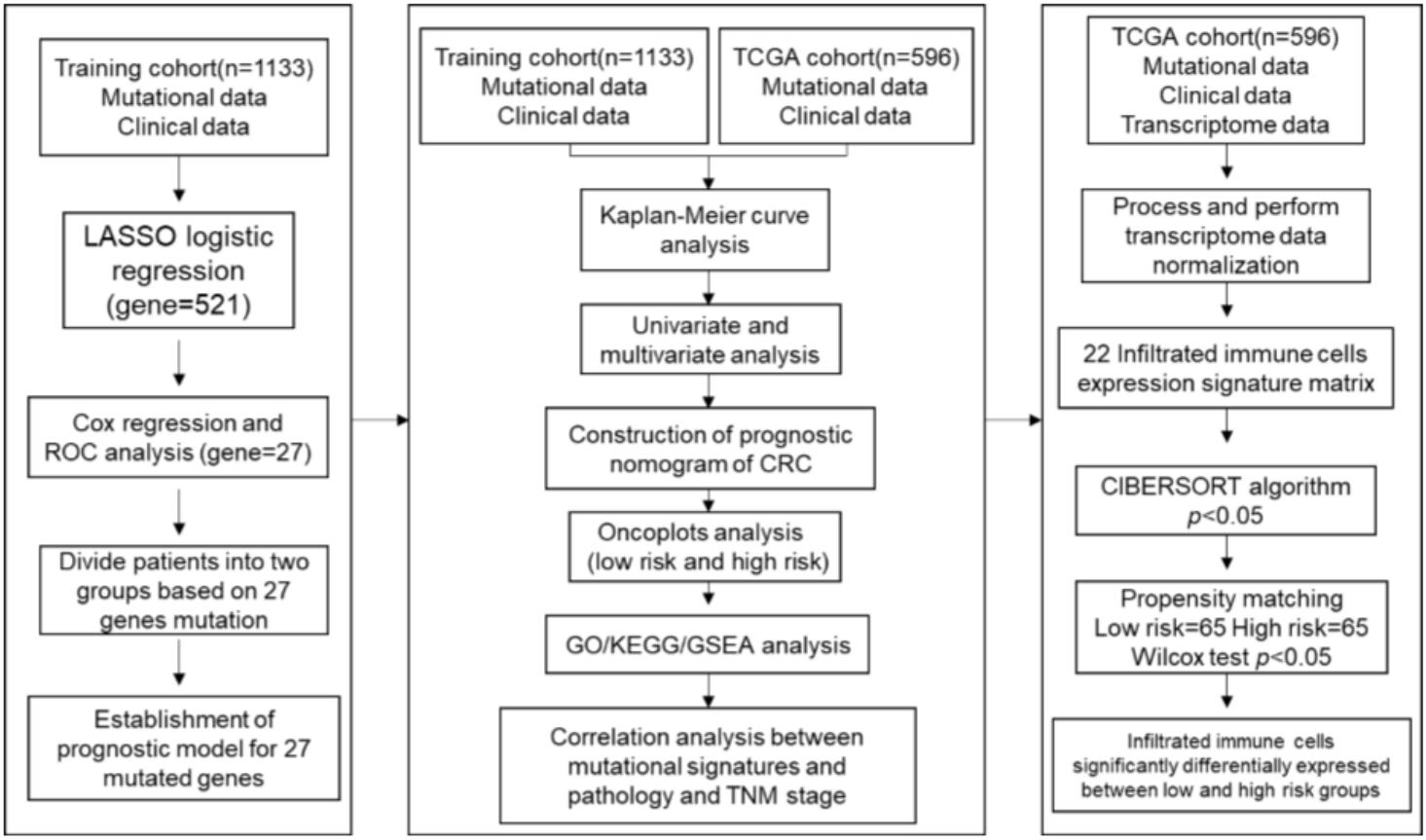

Figure 1

Flowchart detailing the procedure of analyzing mutational signature and their correlation with clinical characteristics and immunity, as well as prognostic models of mutated genes.

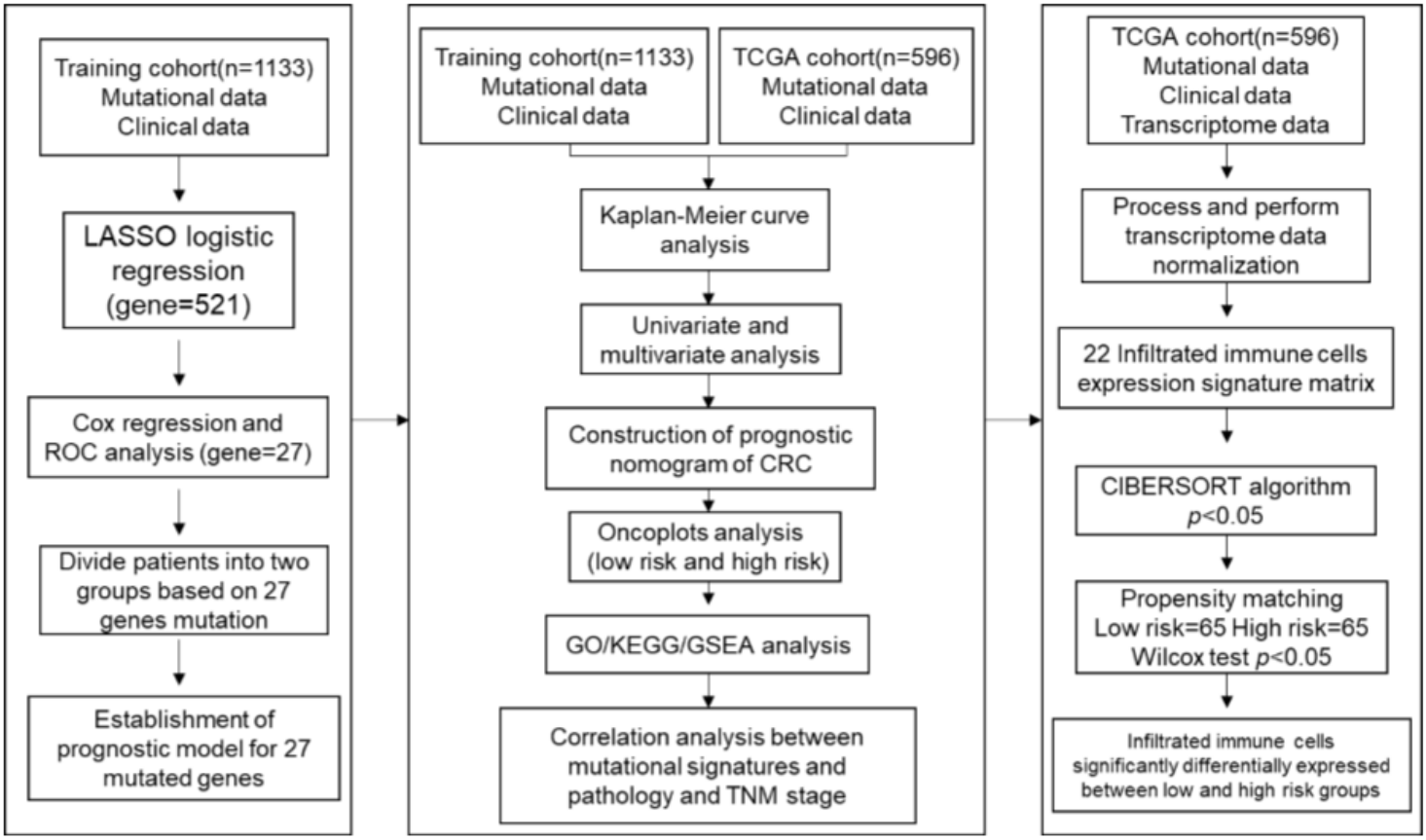

Figure 1

Flowchart detailing the procedure of analyzing mutational signature and their correlation with clinical characteristics and immunity, as well as prognostic models of mutated genes.

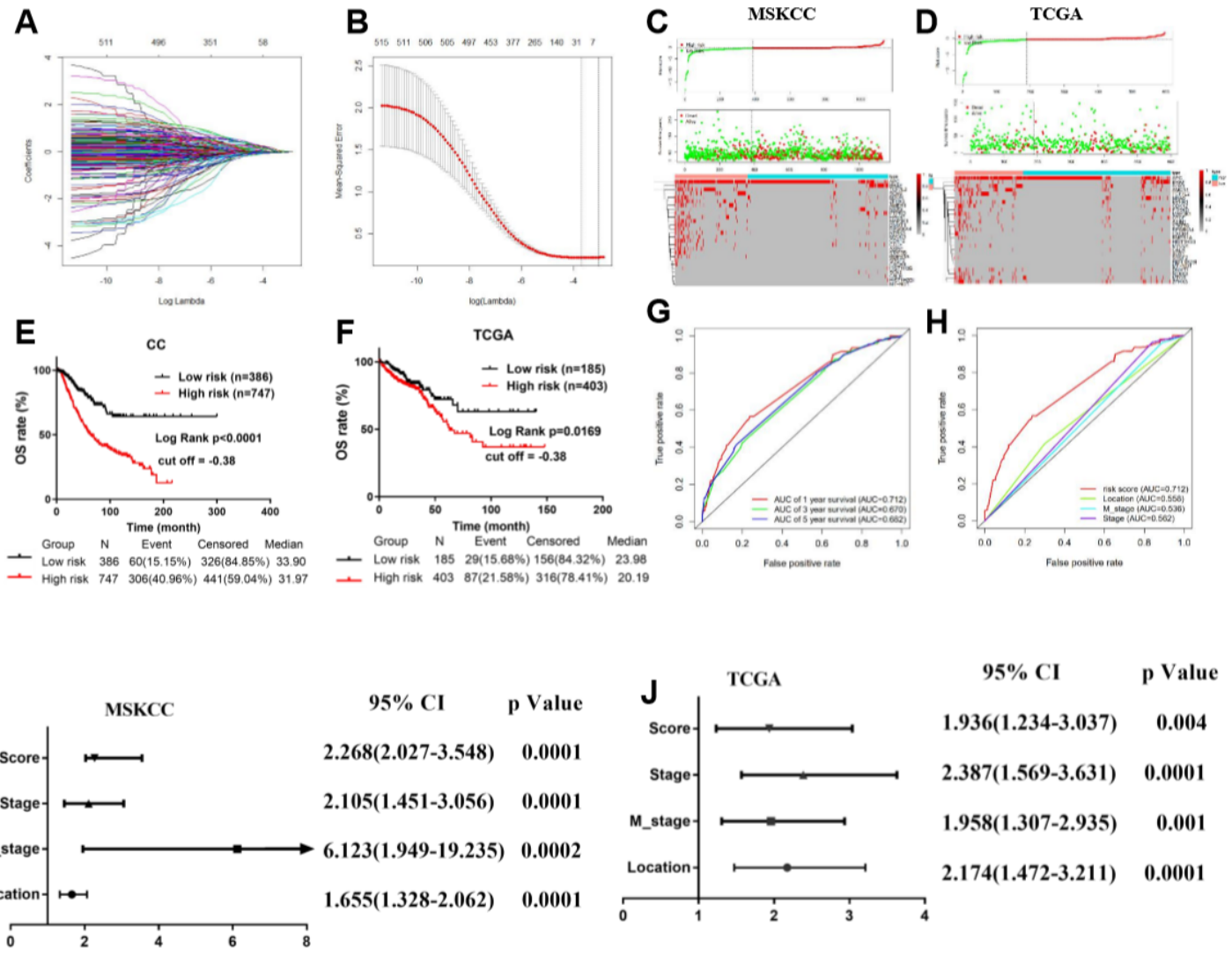

Figure 2

Identification of mutational signature and its prognostic value in CRC. A. and B. Determination of the number of factors by the LASSO analysis. C. and D. The distribution of risk score, survival duration and status of patients, and a heatmap of mutated genes in the classifier. E. Kaplan-Meier curve for prognostic model showing the overall survival based on relative high- and low-risk patients for OS in the training cohort. F. Kaplan-Meier curve for prognostic model showing the overall survival based on relative high- and low-risk patients for OS in the validation cohort. G. ROC curve analysis of the signature in 1-year, 3-year, and 5-year in the MSKCC cohort, AUC, area under the curve. H. ROC curve analysis of the risk score, tumor location, M stage, and TNM stage in the MSKCC cohort, AUC, area under the curve. I. and J. Clinical pathologic features and mutational signature were selected for multivariate Cox regression analysis to build a predictive model for OS in MSKCC and TCGA.

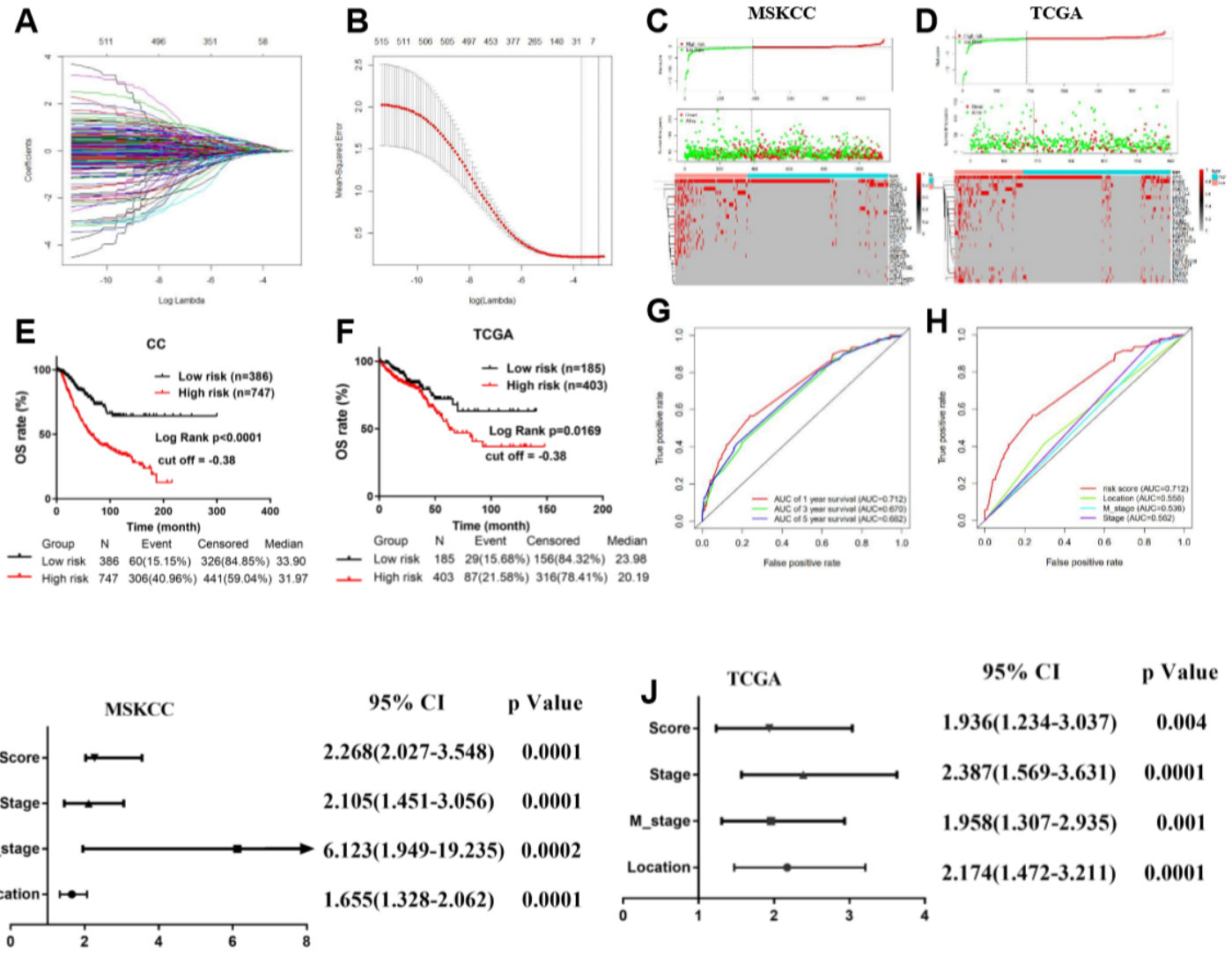

Figure 2

Identification of mutational signature and its prognostic value in CRC. A. and B. Determination of the number of factors by the LASSO analysis. C. and D. The distribution of risk score, survival duration and status of patients, and a heatmap of mutated genes in the classifier. E. Kaplan-Meier curve for prognostic model showing the overall survival based on relative high- and low-risk patients for OS in the training cohort. F. Kaplan-Meier curve for prognostic model showing the overall survival based on relative high- and low-risk patients for OS in the validation cohort. G. ROC curve analysis of the signature in 1-year, 3-year, and 5-year in the MSKCC cohort, AUC, area under the curve. H. ROC curve analysis of the risk score, tumor location, M stage, and TNM stage in the MSKCC cohort, AUC, area under the curve. I. and J. Clinical pathologic features and mutational signature were selected for multivariate Cox regression analysis to build a predictive model for OS in MSKCC and TCGA.

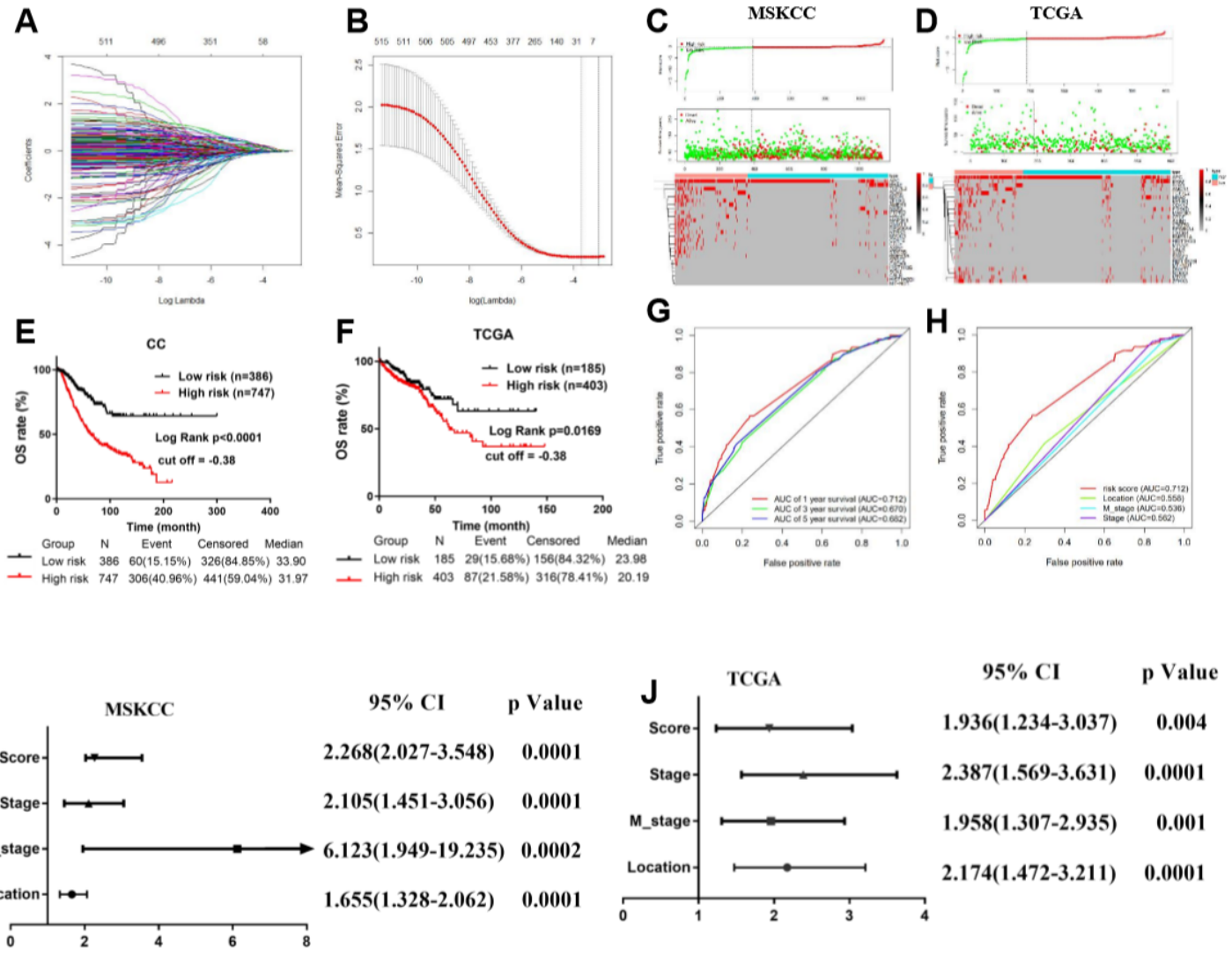

Figure 2

Identification of mutational signature and its prognostic value in CRC. A. and B. Determination of the number of factors by the LASSO analysis. C. and D. The distribution of risk score, survival duration and status of patients, and a heatmap of mutated genes in the classifier. E. Kaplan-Meier curve for prognostic model showing the overall survival based on relative high- and low-risk patients for OS in the training cohort. F. Kaplan-Meier curve for prognostic model showing the overall survival based on relative high- and low-risk patients for OS in the validation cohort. G. ROC curve analysis of the signature in 1-year, 3-year, and 5-year in the MSKCC cohort, AUC, area under the curve. H. ROC curve analysis of the risk score, tumor location, M stage, and TNM stage in the MSKCC cohort, AUC, area under the curve. I. and J. Clinical pathologic features and mutational signature were selected for multivariate Cox regression analysis to build a predictive model for OS in MSKCC and TCGA.

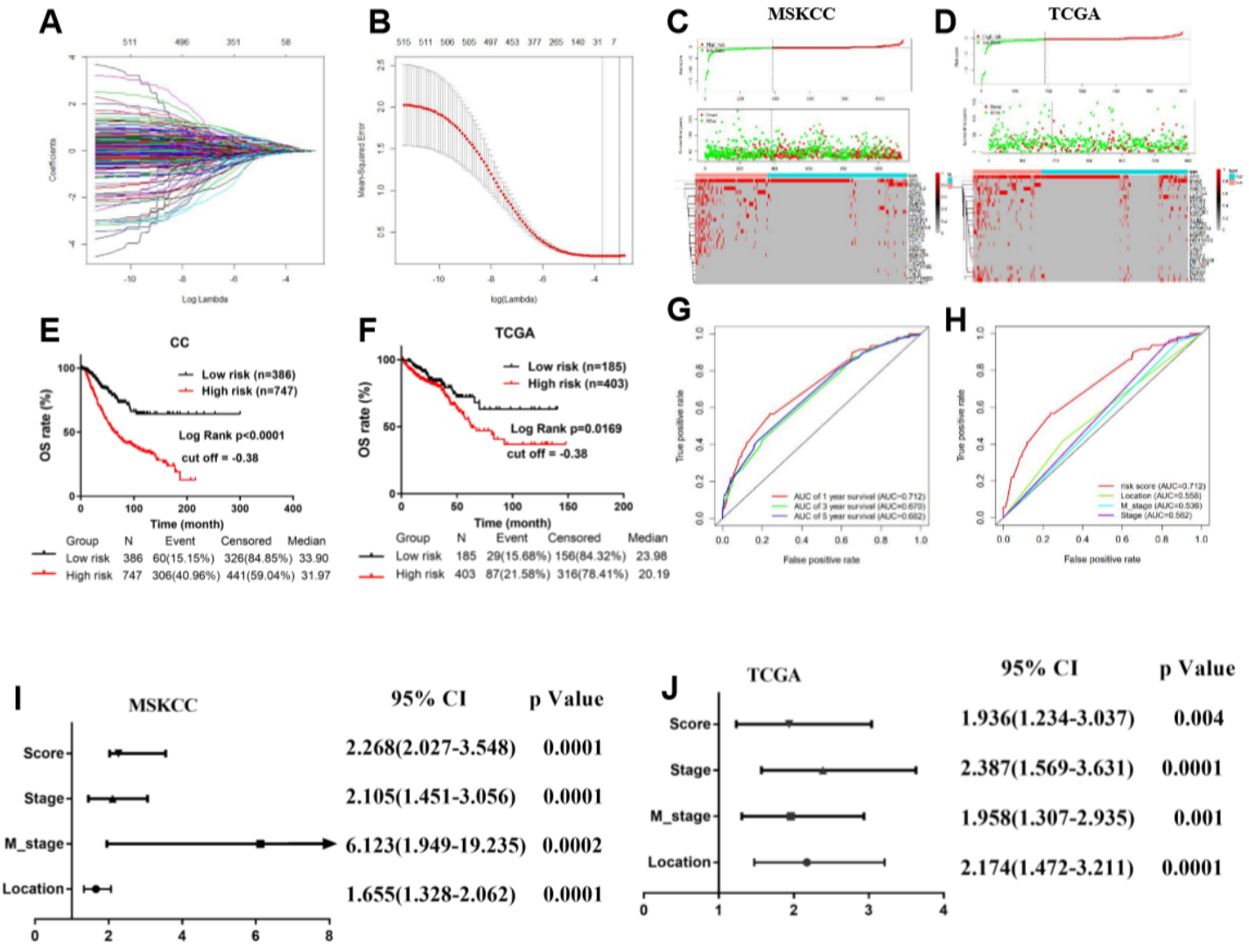

Figure 2

Identification of mutational signature and its prognostic value in CRC. A. and B. Determination of the number of factors by the LASSO analysis. C. and D. The distribution of risk score, survival duration and status of patients, and a heatmap of mutated genes in the classifier. E. Kaplan-Meier curve for prognostic model showing the overall survival based on relative high- and low-risk patients for OS in the training cohort. F. Kaplan-Meier curve for prognostic model showing the overall survival based on relative high- and low-risk patients for OS in the validation cohort. G. ROC curve analysis of the signature in 1-year, 3-year, and 5-year in the MSKCC cohort, AUC, area under the curve. H. ROC curve analysis of the risk score, tumor location, M stage, and TNM stage in the MSKCC cohort, AUC, area under the curve. I. and J. Clinical pathologic features and mutational signature were selected for multivariate Cox regression analysis to build a predictive model for OS in MSKCC and TCGA.

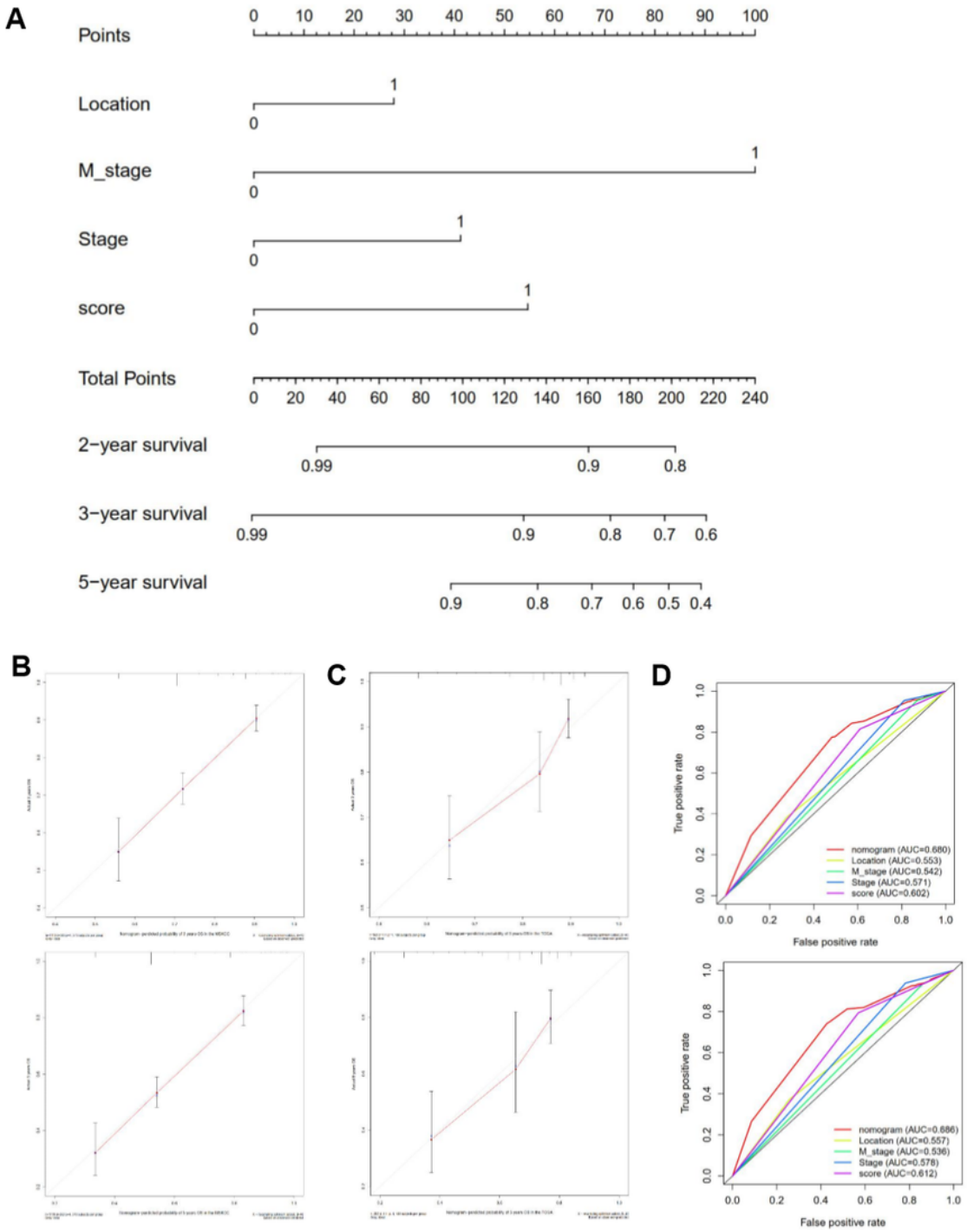

Figure 3

Univariate and multivariate COX regression analyses of clinical factors and independence associated with prognosis/Nomogram A to predict the risk of overall survival in CRC. A. Nomogram to predict distant metastasis-free survival. B. Calibration curves of the nomogram to predict overall survival at 3, and 5 years in the MSKCC cohort. C. Calibration curves of the nomogram to predict overall survival at 3, and 5

years in the TCGA cohort. D. ROC curve analysis of the nomogram, risk score, tumor location, M stage, and TNM stage at 3, and 5 years in the MSKCC cohort, AUC, area under the curve.

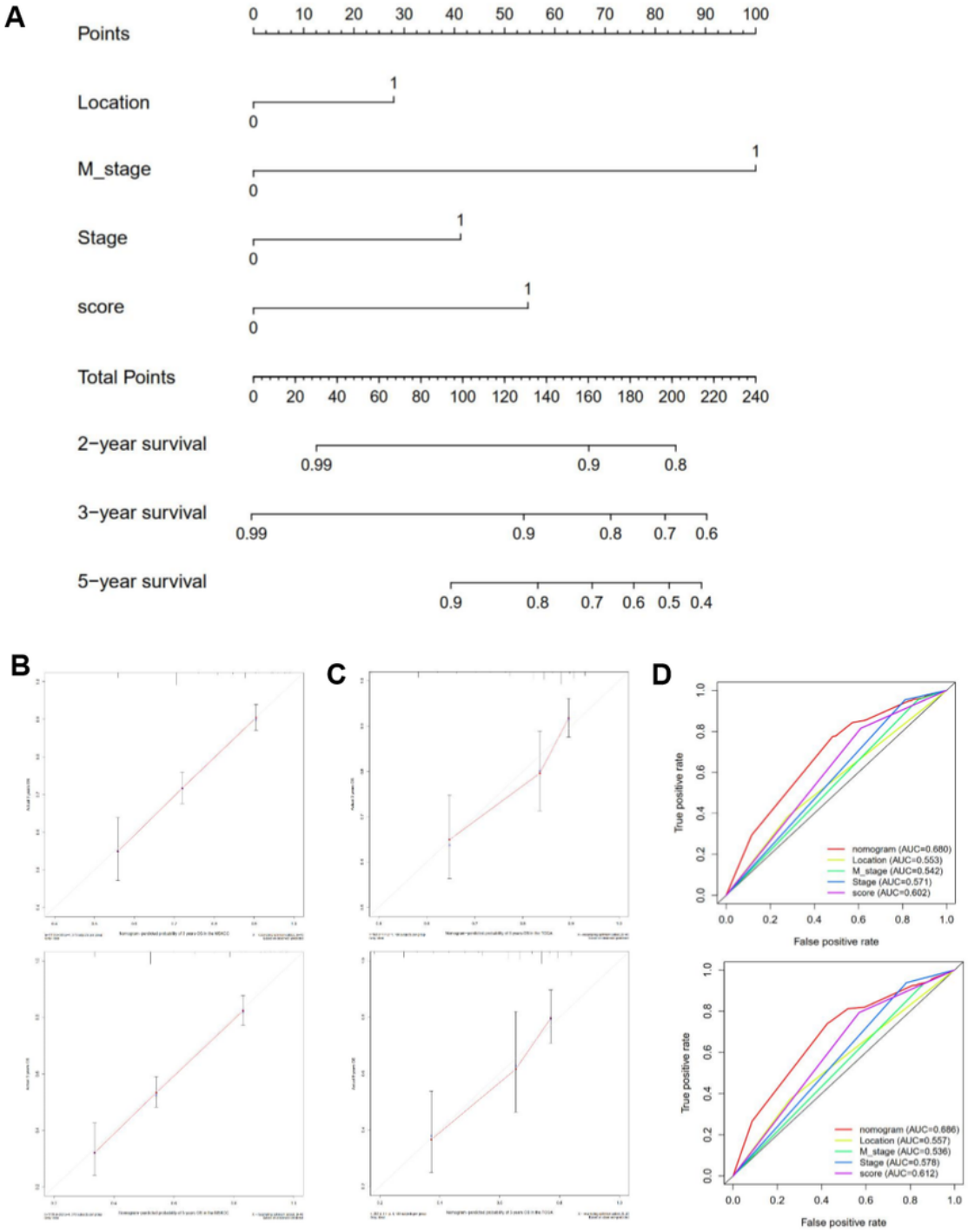

Figure 3

Univariate and multivariate COX regression analyses of clinical factors and independence associated with prognosis/Nomogram A to predict the risk of overall survival in CRC. A. Nomogram to predict distant metastasis-free survival. B. Calibration curves of the nomogram to predict overall survival at 3, and 5

years in the MSKCC cohort. C. Calibration curves of the nomogram to predict overall survival at 3, and 5 years in the TCGA cohort. D. ROC curve analysis of the nomogram, risk score, tumor location, M stage, and TNM stage at 3, and 5 years in the MSKCC cohort, AUC, area under the curve.

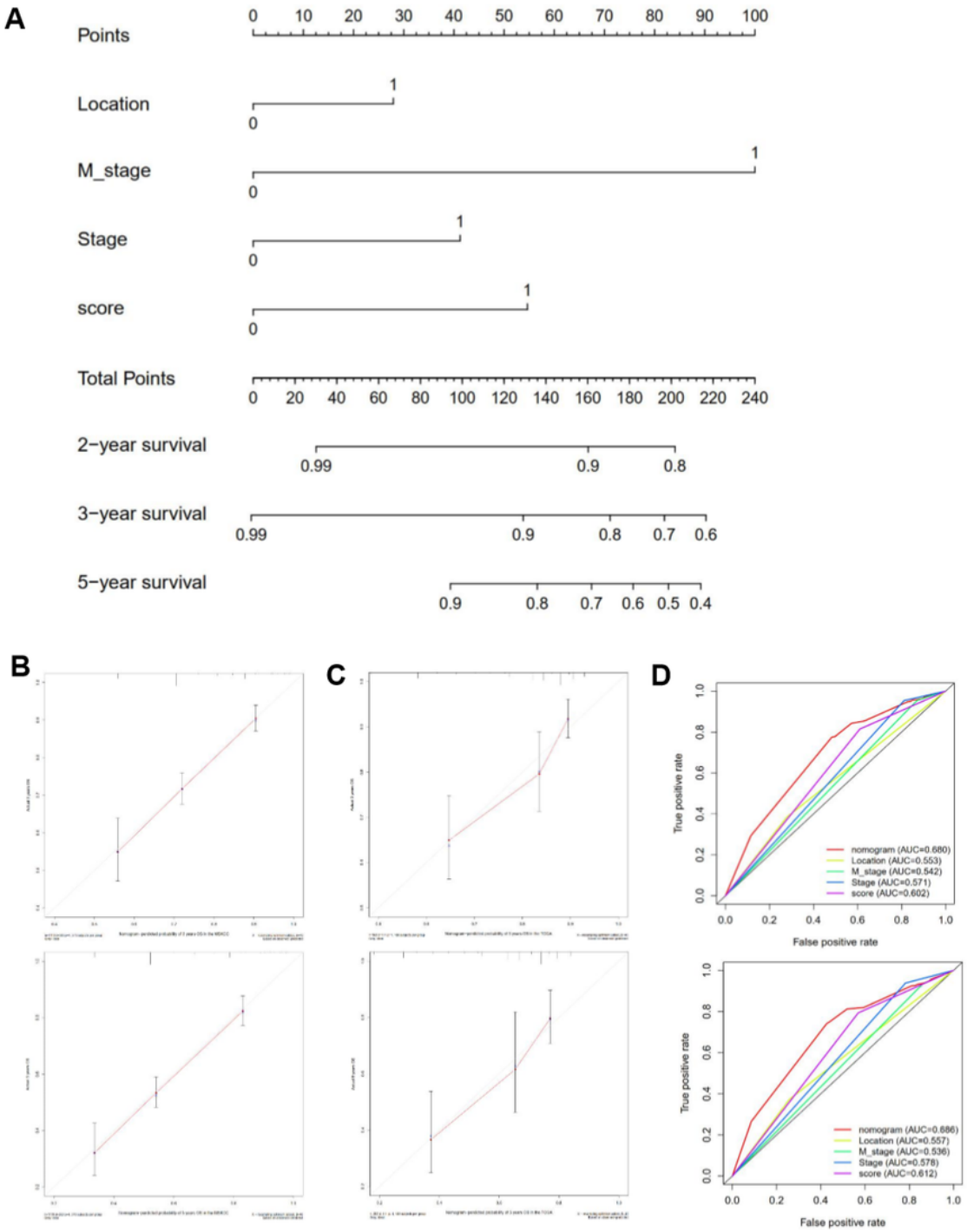

Figure 3

Univariate and multivariate COX regression analyses of clinical factors and independence associated with prognosis/Nomogram A to predict the risk of overall survival in CRC. A. Nomogram to predict distant

metastasis-free survival. B. Calibration curves of the nomogram to predict overall survival at 3, and 5 years in the MSKCC cohort. C. Calibration curves of the nomogram to predict overall survival at 3, and 5 years in the TCGA cohort. D. ROC curve analysis of the nomogram, risk score, tumor location, M stage, and TNM stage at 3, and 5 years in the MSKCC cohort, AUC, area under the curve.

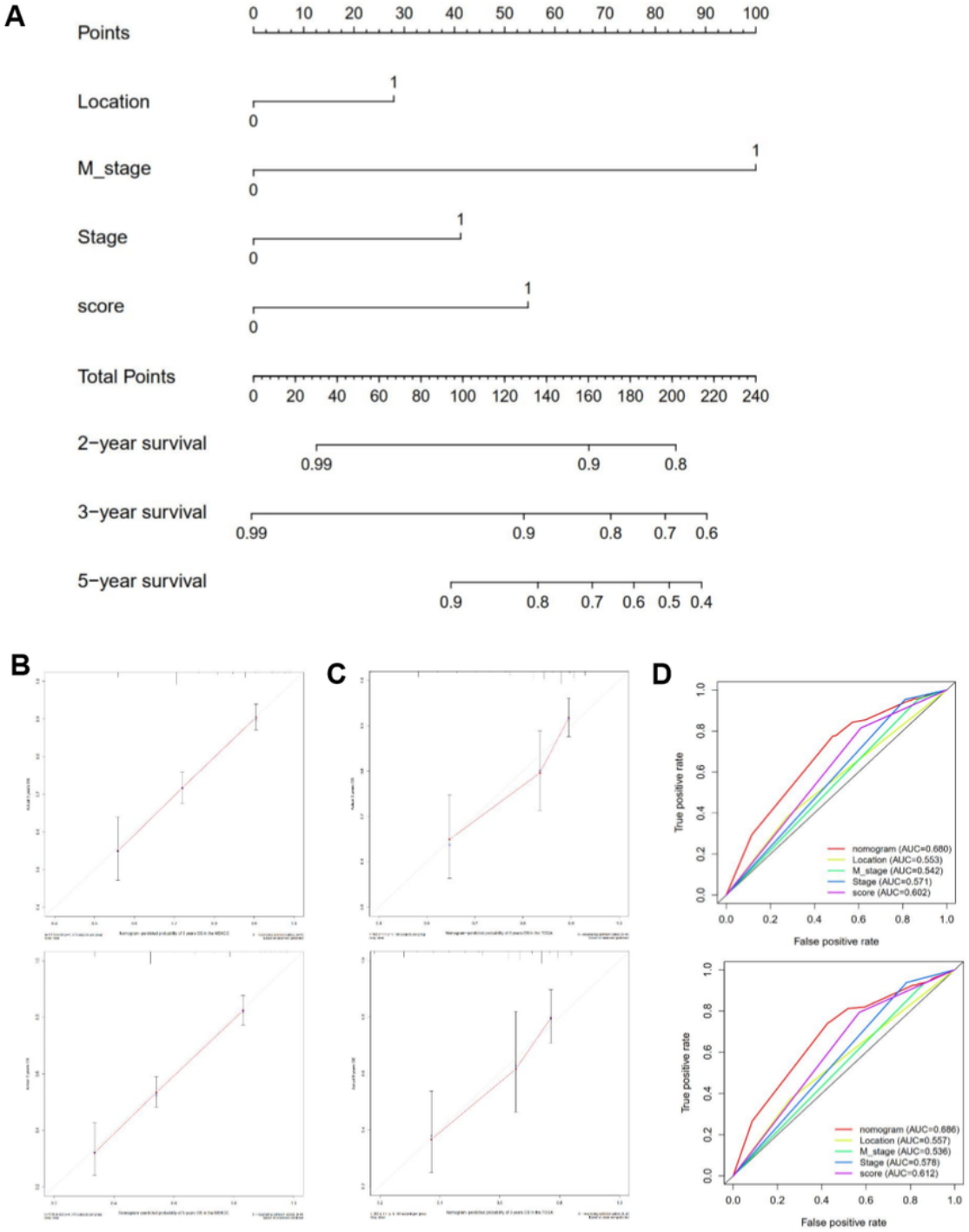

Figure 3

Univariate and multivariate COX regression analyses of clinical factors and independence associated with prognosis/Nomogram A to predict the risk of overall survival in CRC. A. Nomogram to predict distant metastasis-free survival. B. Calibration curves of the nomogram to predict overall survival at 3, and 5 years in the MSKCC cohort. C. Calibration curves of the nomogram to predict overall survival at 3, and 5 years in the TCGA cohort. D. ROC curve analysis of the nomogram, risk score, tumor location, M stage, and TNM stage at 3, and 5 years in the MSKCC cohort, AUC, area under the curve.

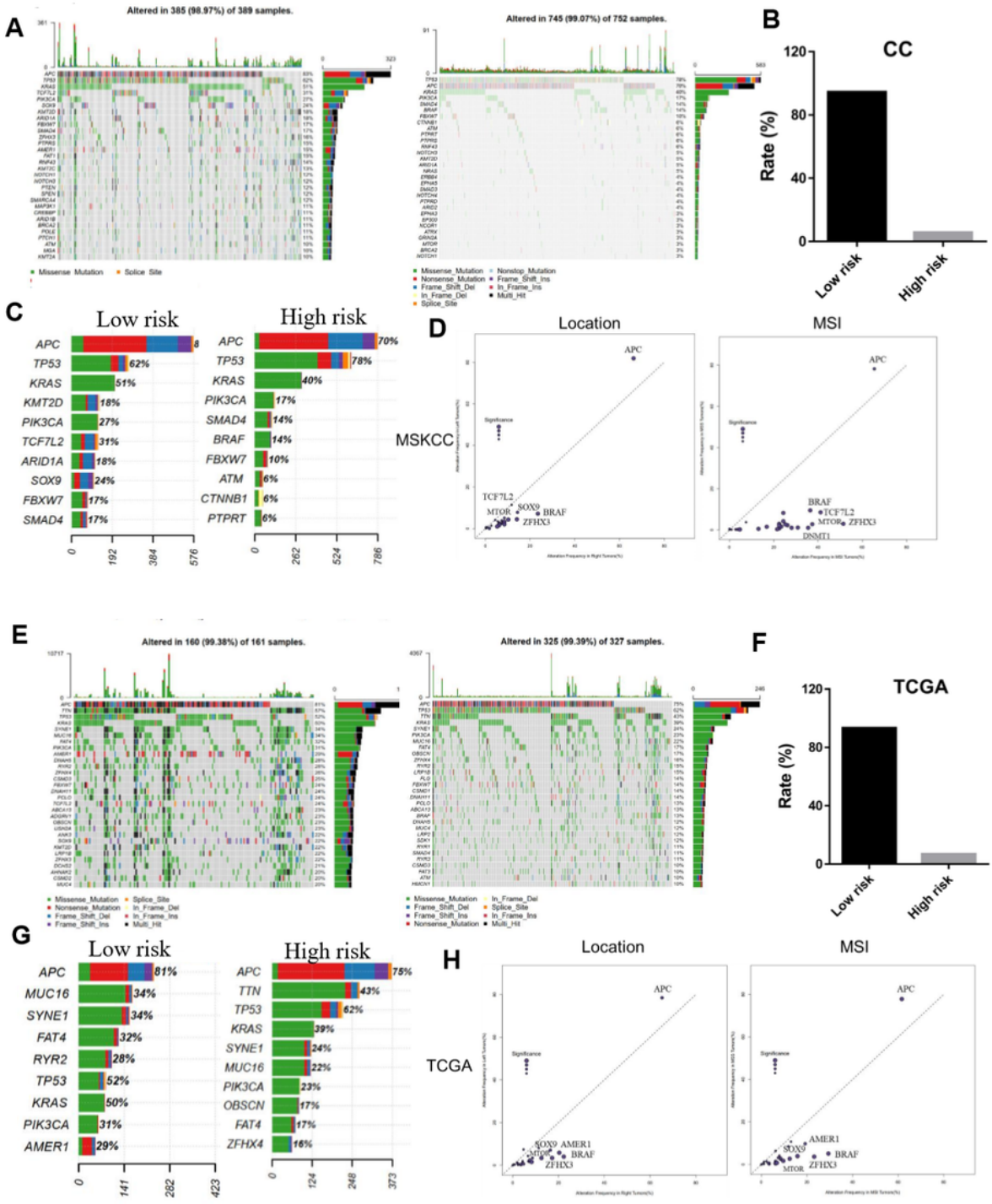

Figure 4

Mutational landscape of significantly mutated genes in the training cohort and verification cohort. A. Top 30 genes with the most significant mutations in MSKCC cohort. The bar chart above shows the total number of synonymous and non-synonymous mutations in each patient's top 30 genes. The bar chart on the right shows the number of samples in which the 30 genes were mutated at low risk and high-risk groups. The different colors in the thermogram indicate the type of mutation; gray indicates no mutation. B. The low-risk group with more mutations, while the high-risk group with fewer mutations in MSKCC cohort. C. The top 10 most high mutation genes in low risk and high-risk group in MSKCC cohort. D. Genomic alteration enrichment analysis by primary tumor site in location and molecular subtype in MSKCC cohort. E. Top 30 genes with the most significant mutations in TCGA cohort. The bar chart above shows the total number of synonymous and non-synonymous mutations in each patient's top 30 genes. The bar chart on the right shows the number of samples in which the 30 genes were mutated at low risk and high-risk groups. The different colors in the thermogram indicate the type of mutation; gray indicates no mutation. F. The low-risk group with more mutations, while the high-risk group with fewer mutations in TCGA cohort. G. The top 10 most high mutation genes in low risk and high-risk group in TCGA cohort. H. Genomic alteration enrichment analysis by primary tumor site in location and molecular subtype in TCGA cohort.

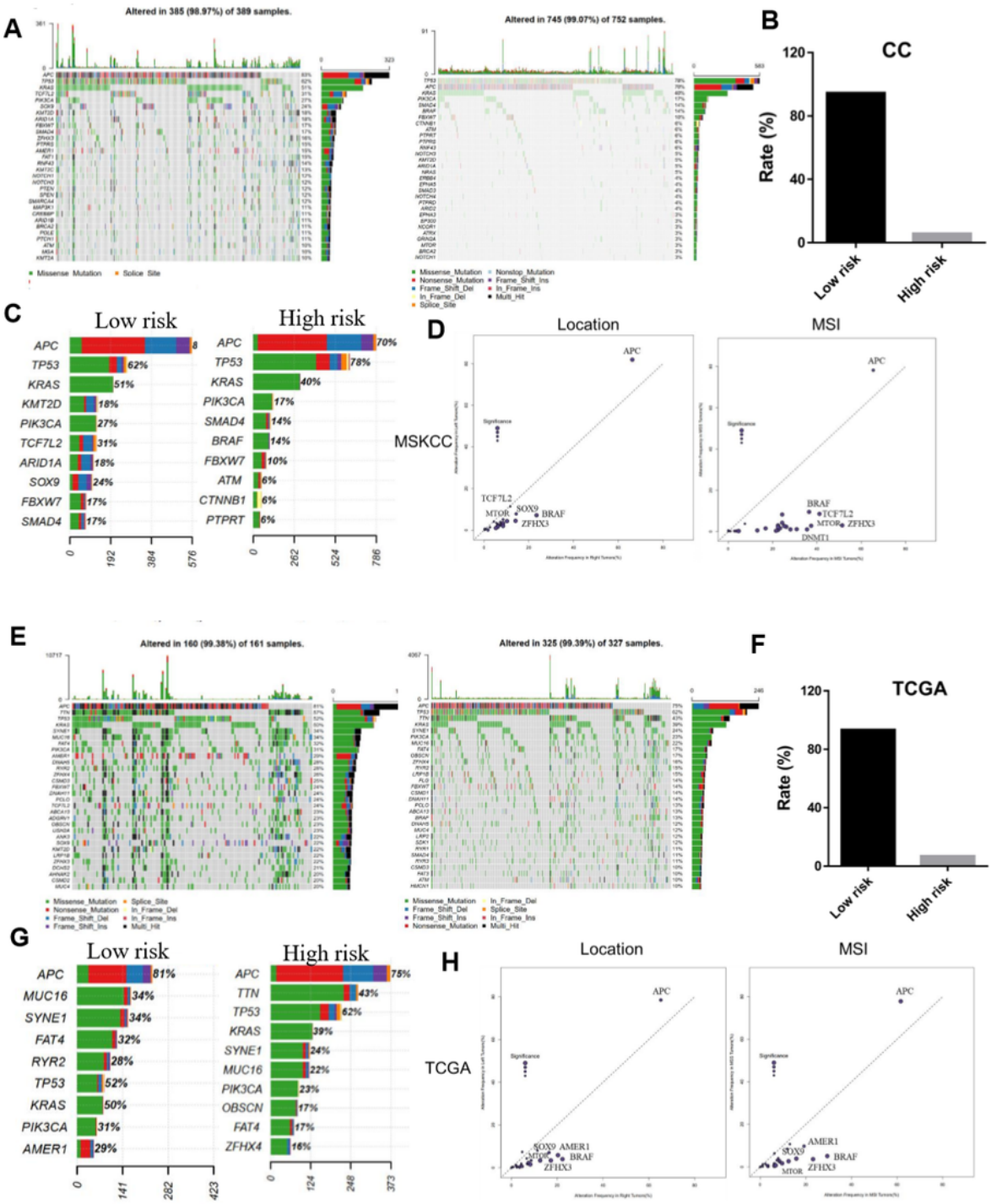

Figure 4

Mutational landscape of significantly mutated genes in the training cohort and verification cohort. A. Top 30 genes with the most significant mutations in MSKCC cohort. The bar chart above shows the total number of synonymous and non-synonymous mutations in each patient's top 30 genes. The bar chart on the right shows the number of samples in which the 30 genes were mutated at low risk and high-risk groups. The different colors in the thermogram indicate the type of mutation; gray indicates no mutation.

B. The low-risk group with more mutations, while the high-risk group with fewer mutations in MSKCC cohort. C. The top 10 most high mutation genes in low risk and high-risk group in MSKCC cohort. D. Genomic alteration enrichment analysis by primary tumor site in location and molecular subtype in MSKCC cohort. E. Top 30 genes with the most significant mutations in TCGA cohort. The bar chart above shows the total number of synonymous and non-synonymous mutations in each patient's top 30 genes. The bar chart on the right shows the number of samples in which the 30 genes were mutated at low risk and high-risk groups. The different colors in the thermogram indicate the type of mutation; gray indicates no mutation. F. The low-risk group with more mutations, while the high-risk group with fewer mutations in TCGA cohort. G. The top 10 most high mutation genes in low risk and high-risk group in TCGA cohort. H. Genomic alteration enrichment analysis by primary tumor site in location and molecular subtype in TCGA cohort.

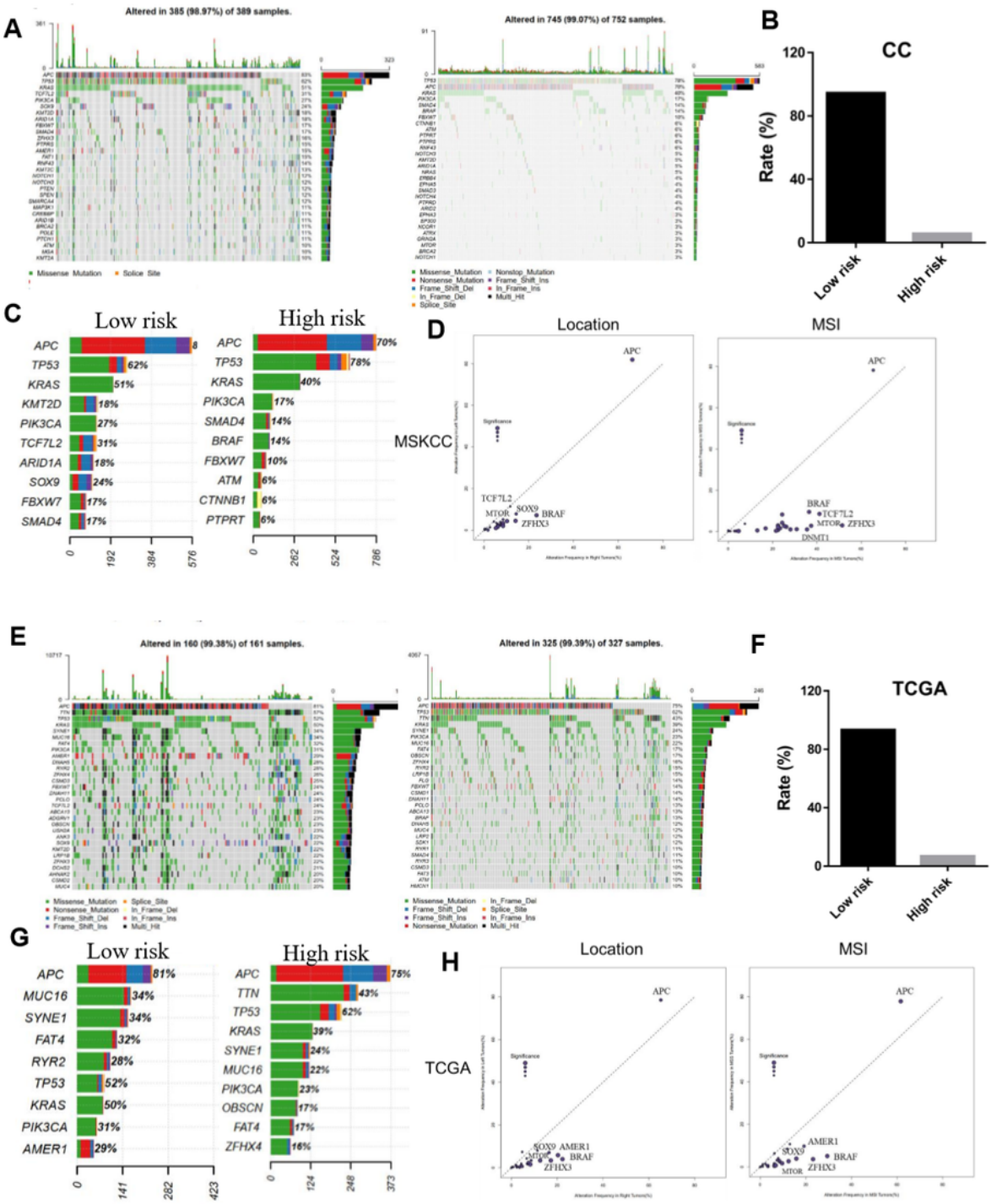

Figure 4

Mutational landscape of significantly mutated genes in the training cohort and verification cohort. A. Top 30 genes with the most significant mutations in MSKCC cohort. The bar chart above shows the total number of synonymous and non-synonymous mutations in each patient's top 30 genes. The bar chart on the right shows the number of samples in which the 30 genes were mutated at low risk and high-risk groups. The different colors in the thermogram indicate the type of mutation; gray indicates no mutation.

B. The low-risk group with more mutations, while the high-risk group with fewer mutations in MSKCC cohort. C. The top 10 most high mutation genes in low risk and high-risk group in MSKCC cohort. D. Genomic alteration enrichment analysis by primary tumor site in location and molecular subtype in MSKCC cohort. E. Top 30 genes with the most significant mutations in TCGA cohort. The bar chart above shows the total number of synonymous and non-synonymous mutations in each patient's top 30 genes. The bar chart on the right shows the number of samples in which the 30 genes were mutated at low risk and high-risk groups. The different colors in the thermogram indicate the type of mutation; gray indicates no mutation. F. The low-risk group with more mutations, while the high-risk group with fewer mutations in TCGA cohort. G. The top 10 most high mutation genes in low risk and high-risk group in TCGA cohort. H. Genomic alteration enrichment analysis by primary tumor site in location and molecular subtype in TCGA cohort.

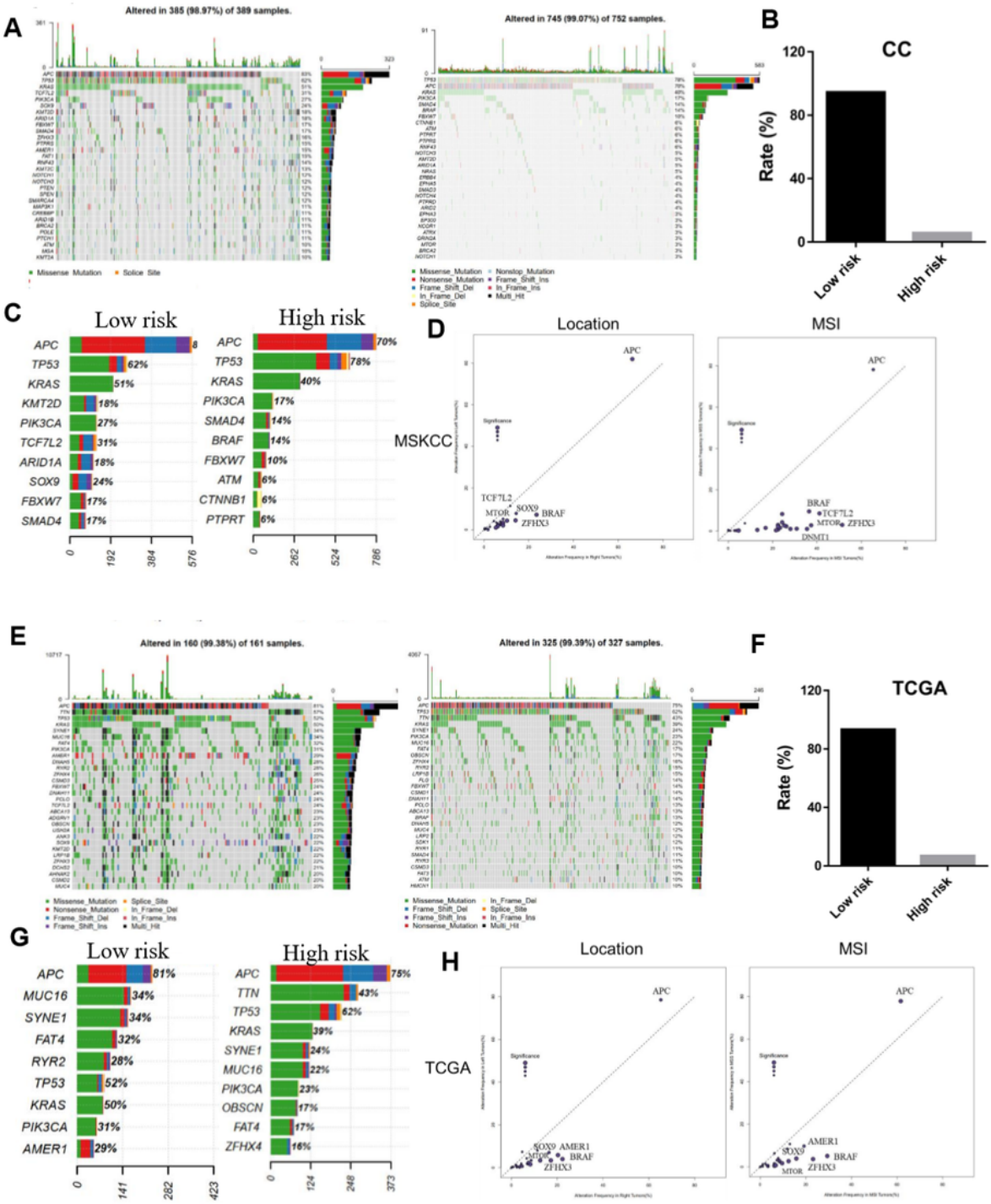

Figure 4

Mutational landscape of significantly mutated genes in the training cohort and verification cohort. A. Top 30 genes with the most significant mutations in MSKCC cohort. The bar chart above shows the total number of synonymous and non-synonymous mutations in each patient's top 30 genes. The bar chart on the right shows the number of samples in which the 30 genes were mutated at low risk and high-risk groups. The different colors in the thermogram indicate the type of mutation; gray indicates no mutation.

B. The low-risk group with more mutations, while the high-risk group with fewer mutations in MSKCC cohort. C. The top 10 most high mutation genes in low risk and high-risk group in MSKCC cohort. D. Genomic alteration enrichment analysis by primary tumor site in location and molecular subtype in MSKCC cohort. E. Top 30 genes with the most significant mutations in TCGA cohort. The bar chart above shows the total number of synonymous and non-synonymous mutations in each patient's top 30 genes. The bar chart on the right shows the number of samples in which the 30 genes were mutated at low risk and high-risk groups. The different colors in the thermogram indicate the type of mutation; gray indicates no mutation. F. The low-risk group with more mutations, while the high-risk group with fewer mutations in TCGA cohort. G. The top 10 most high mutation genes in low risk and high-risk group in TCGA cohort. H. Genomic alteration enrichment analysis by primary tumor site in location and molecular subtype in TCGA cohort.

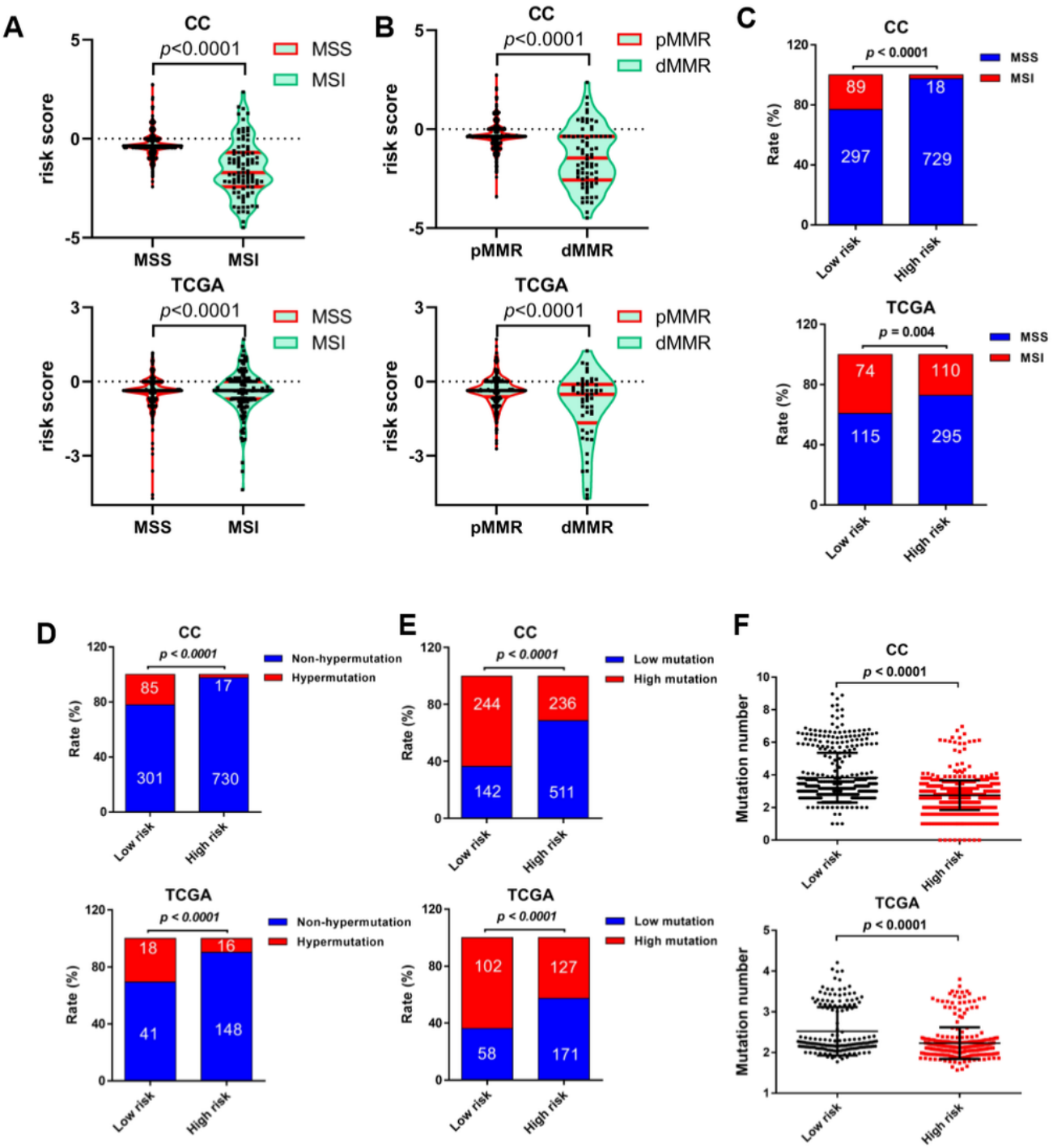

Figure 5

The mutational signature are associated with the microsatellite instability (MSI. and of mismatch repair (MMR. genomic feature of CRC. A. MSI cancers have significantly lower risk score than the MSS cancers both in the training and validation cohorts. B. The dMMR cancers have significantly lower risk score than the pMMR cancers both in the training and validation cohorts. C. The proportion of MSI was significantly increased in the low-risk group both in the training and validation cohorts. D. The proportion of

hypermutation was significantly increased in the low-risk group both in the training and validation cohorts. E. The proportion of high mutation was significantly increased in the low-risk group both in the training and validation cohorts. F. The low-risk cancers have significantly higher mutation number than the high-risk cancers both in the training and validation cohorts.

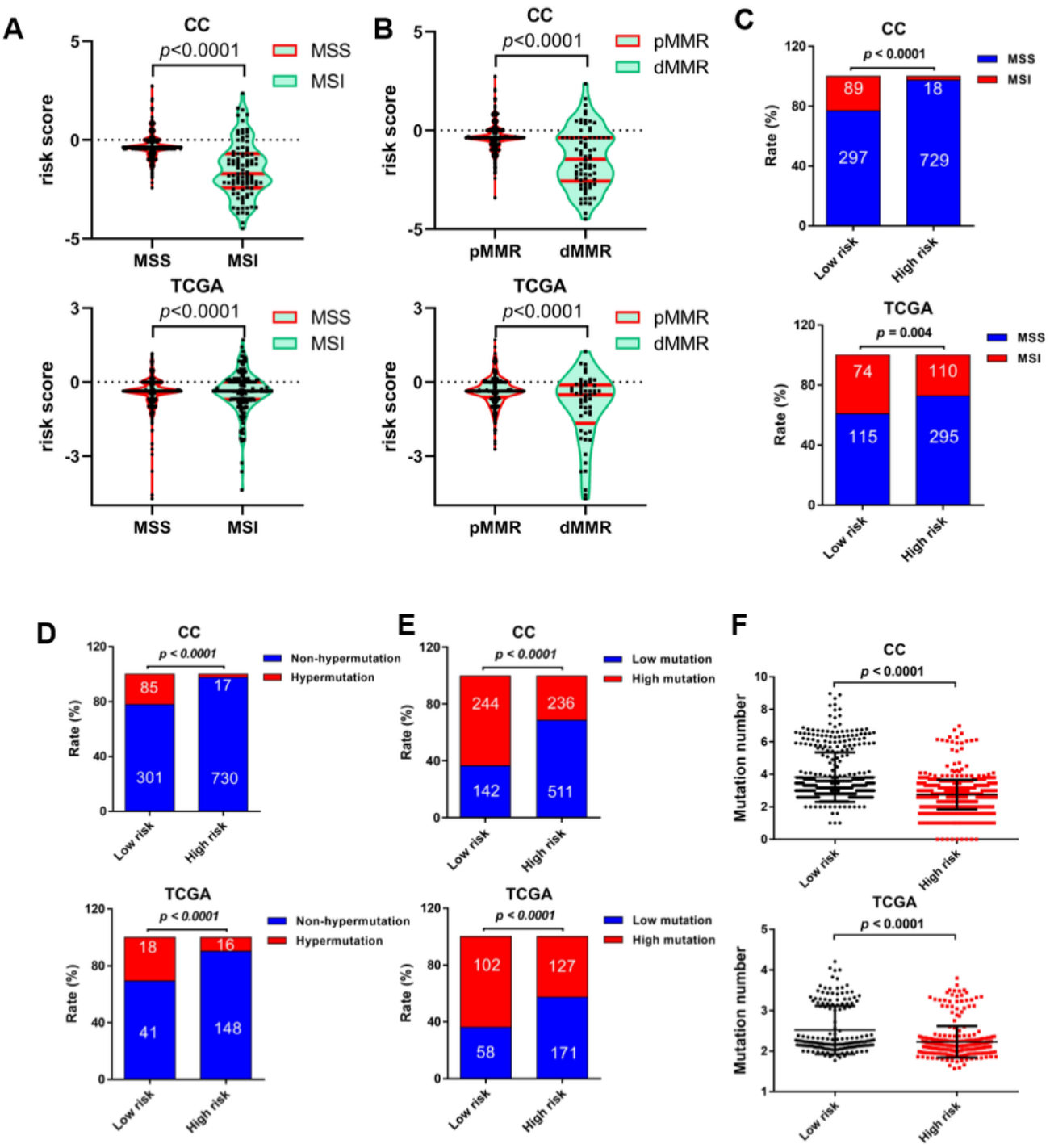

Figure 5

The mutational signature are associated with the microsatellite instability (MSI. and of mismatch repair (MMR. genomic feature of CRC. A. MSI cancers have significantly lower risk score than the MSS cancers both in the training and validation cohorts. B. The dMMR cancers have significantly lower risk score than the pMMR cancers both in the training and validation cohorts. C. The proportion of MSI was significantly increased in the low-risk group both in the training and validation cohorts. D. The proportion of hypermutation was significantly increased in the low-risk group both in the training and validation cohorts. E. The proportion of high mutation was significantly increased in the low-risk group both in the training and validation cohorts. F. The low-risk cancers have significantly higher mutation number than the high-risk cancers both in the training and validation cohorts.

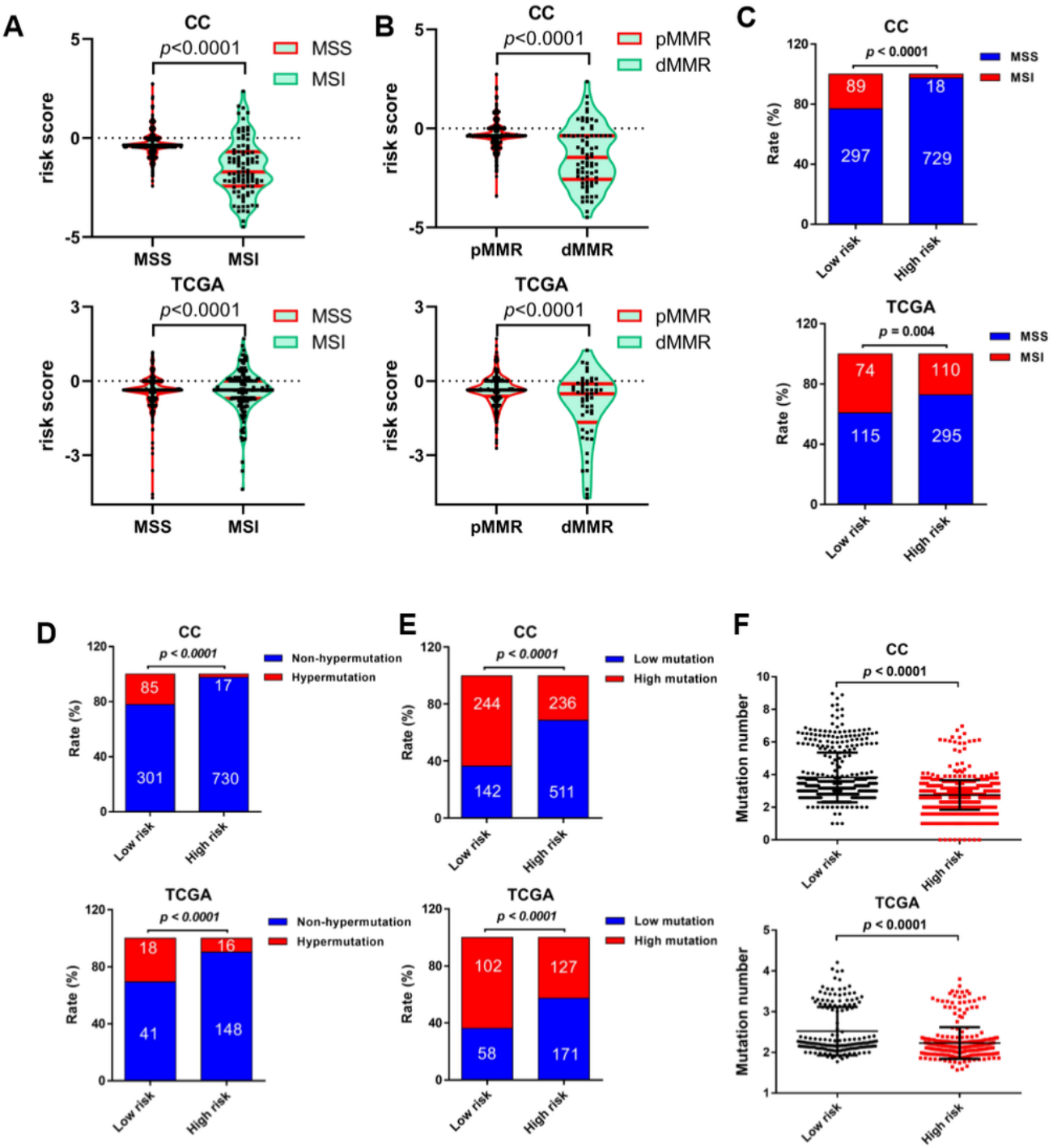

Figure 5

The mutational signature are associated with the microsatellite instability (MSI. and of mismatch repair (MMR. genomic feature of CRC. A. MSI cancers have significantly lower risk score than the MSS cancers both in the training and validation cohorts. B. The dMMR cancers have significantly lower risk score than the pMMR cancers both in the training and validation cohorts. C. The proportion of MSI was significantly increased in the low-risk group both in the training and validation cohorts. D. The proportion of

hypermutation was significantly increased in the low-risk group both in the training and validation cohorts. E. The proportion of high mutation was significantly increased in the low-risk group both in the training and validation cohorts. F. The low-risk cancers have significantly higher mutation number than the high-risk cancers both in the training and validation cohorts.

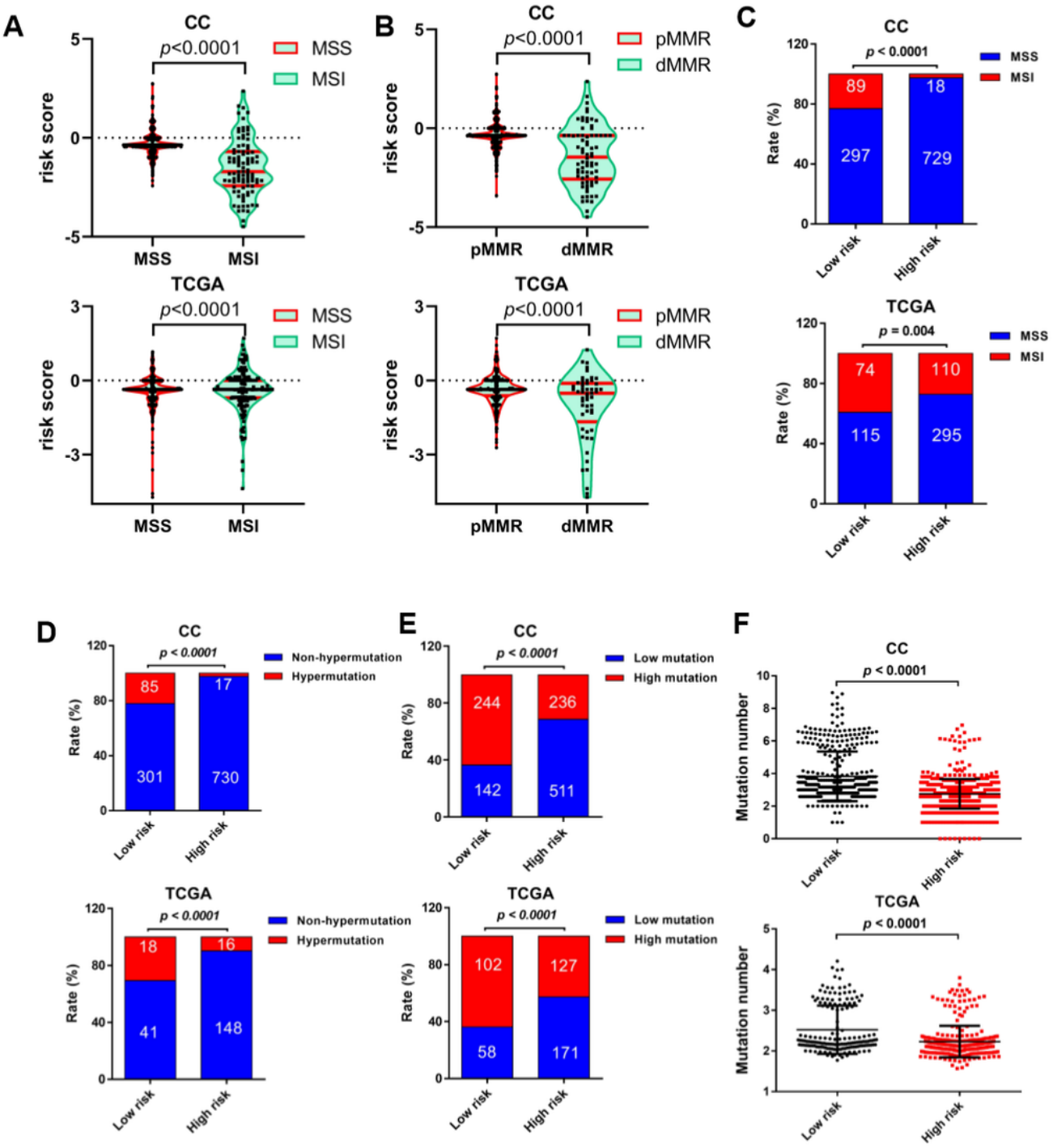

Figure 5

The mutational signature are associated with the microsatellite instability (MSI. and of mismatch repair (MMR. genomic feature of CRC. A. MSI cancers have significantly lower risk score than the MSS cancers both in the training and validation cohorts. B. The dMMR cancers have significantly lower risk score than the pMMR cancers both in the training and validation cohorts. C. The proportion of MSI was significantly increased in the low-risk group both in the training and validation cohorts. D. The proportion of hypermutation was significantly increased in the low-risk group both in the training and validation cohorts. E. The proportion of high mutation was significantly increased in the low-risk group both in the training and validation cohorts. F. The low-risk cancers have significantly higher mutation number than the high-risk cancers both in the training and validation cohorts.

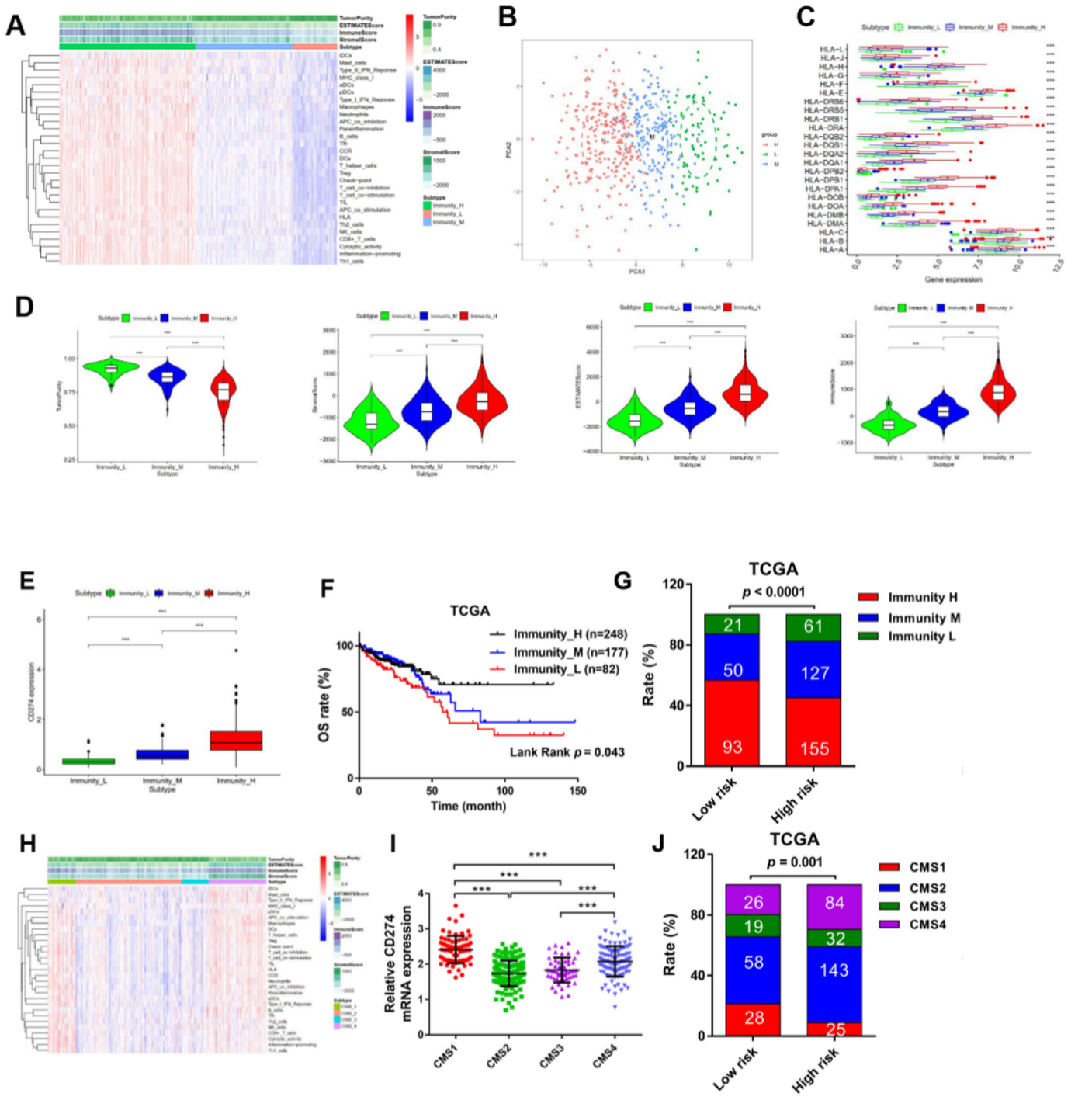

Figure 6

Mutational signature is associated with immune activity in CRC. A. The immune cell infiltration level in each subtype, tumor purity, ESTIMATE score, stromal score, and immune score were evaluated by ESTIMATE. B. PCA analysis of the three clusters. C. Comparison of the expression levels of HLA genes between CRC subtypes (ANOVA test.. D. Comparison of the stromal score, immune score, ESTIMATE score, and tumor purity between CRC subtypes (Mann-Whitney U test.. E. Comparison of PD-L1 (CD274.

expression between CRC subtypes. F. Kaplan-Meier analysis of three immunity cluster. G. The distribution of CRC subtypes in high- and low-risk group. H. The immune cell infiltration level in each CMS subtype, tumor purity, ESTIMATE score, stromal score, and the immune score was evaluated by ESTIMATE algorithm. I. CD274 mRNA expression in CMS subtype. J. The distribution of CMS subtypes in high- and low-risk group.

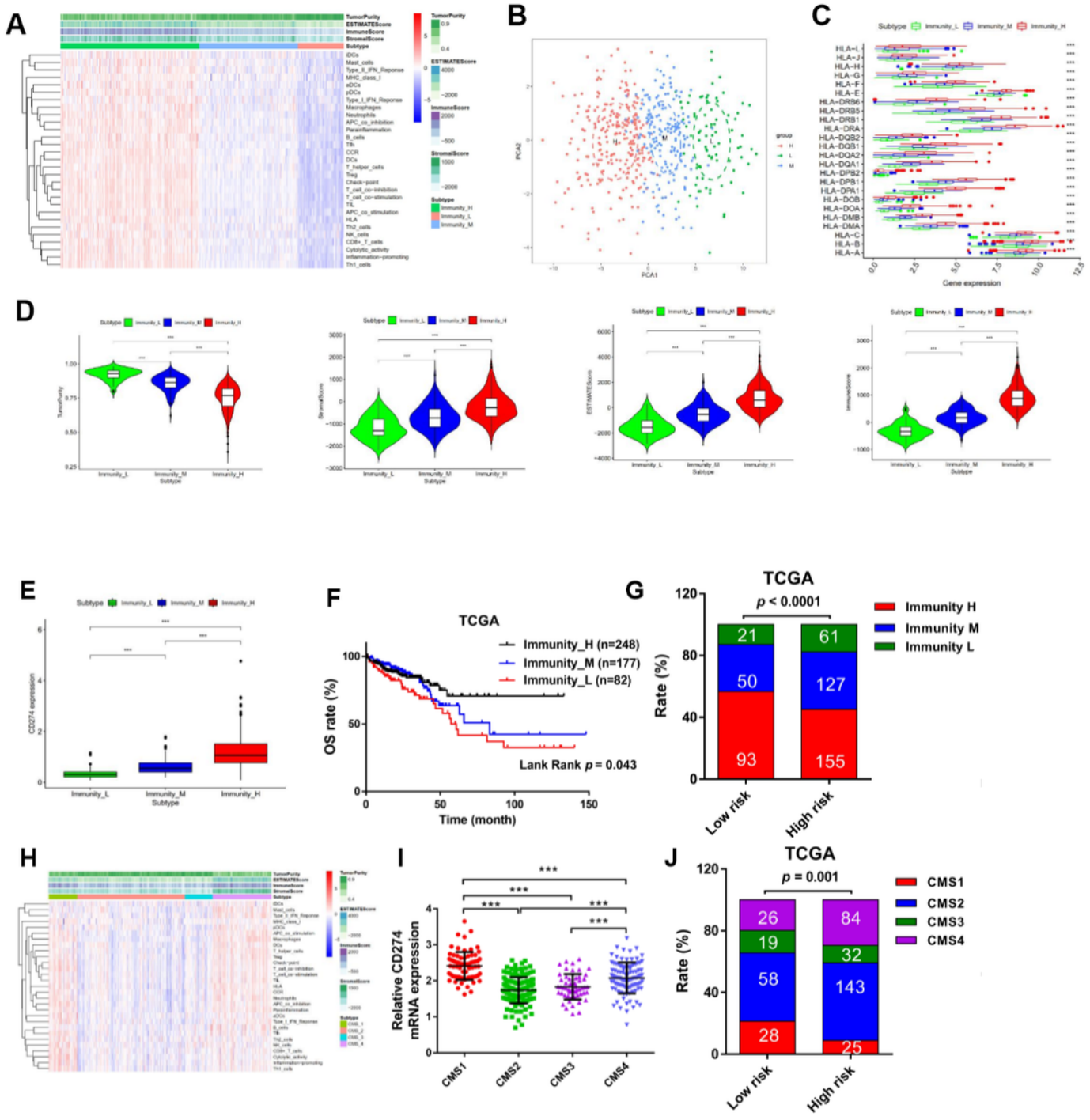

Figure 6

Mutational signature is associated with immune activity in CRC. A. The immune cell infiltration level in each subtype, tumor purity, ESTIMATE score, stromal score, and immune score were evaluated by ESTIMATE. B. PCA analysis of the three clusters. C. Comparison of the expression levels of HLA genes between CRC subtypes (ANOVA test.. D. Comparison of the stromal score, immune score, ESTIMATE score, and tumor purity between CRC subtypes (Mann-Whitney U test.. E. Comparison of PD-L1 (CD274. expression between CRC subtypes. F. Kaplan-Meier analysis of three immunity cluster. G. The distribution of CRC subtypes in high- and low-risk group. H. The immune cell infiltration level in each CMS subtype, tumor purity, ESTIMATE score, stromal score, and the immune score was evaluated by ESTIMATE algorithm. I. CD274 mRNA expression in CMS subtype. J. The distribution of CMS subtypes in high- and low-risk group.

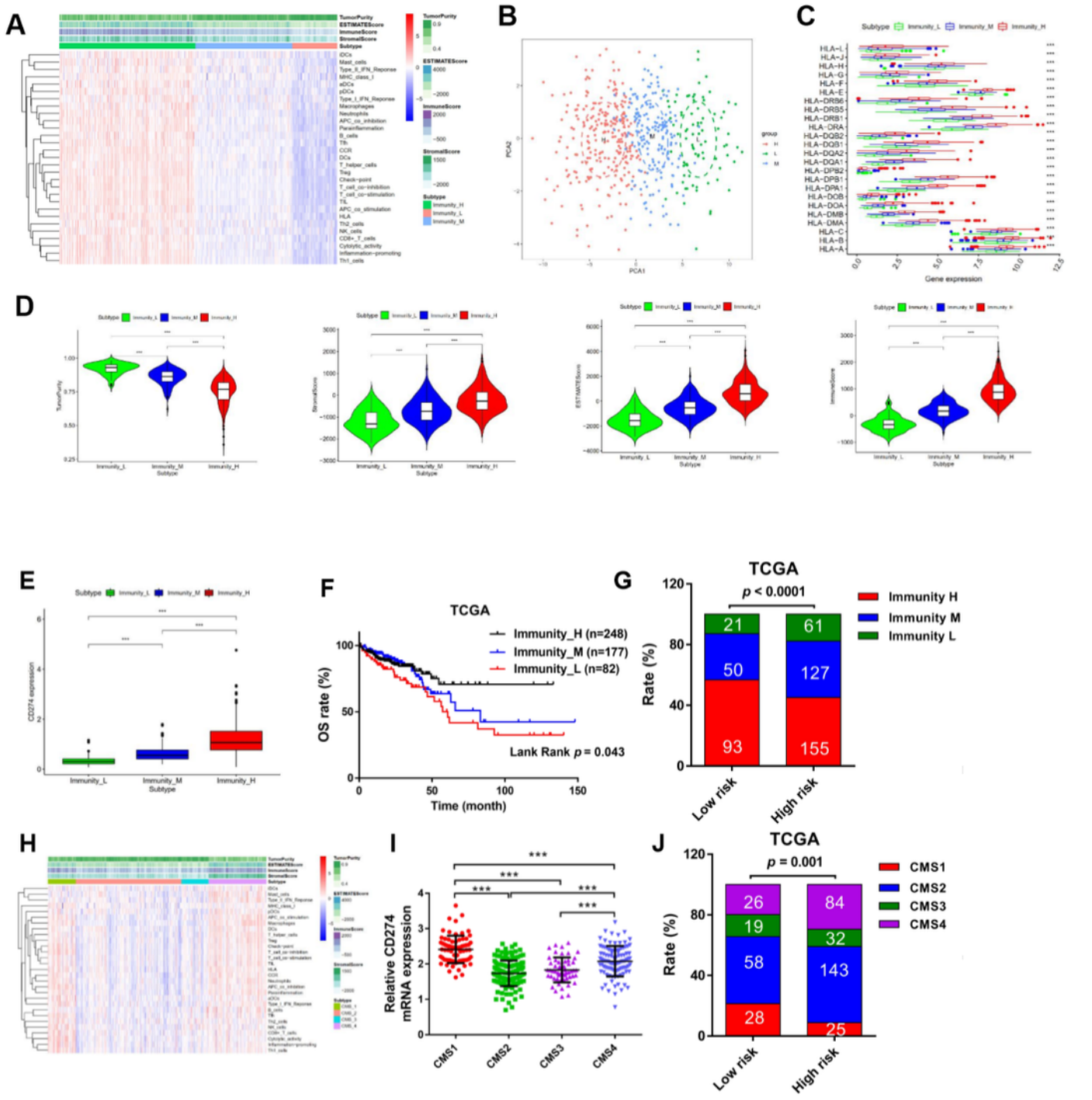

Figure 6

Mutational signature is associated with immune activity in CRC. A. The immune cell infiltration level in each subtype, tumor purity, ESTIMATE score, stromal score, and immune score were evaluated by ESTIMATE. B. PCA analysis of the three clusters. C. Comparison of the expression levels of HLA genes between CRC subtypes (ANOVA test.. D. Comparison of the stromal score, immune score, ESTIMATE score, and tumor purity between CRC subtypes (Mann-Whitney U test.. E. Comparison of PD-L1 (CD274.

expression between CRC subtypes. F. Kaplan-Meier analysis of three immunity cluster. G. The distribution of CRC subtypes in high- and low-risk group. H. The immune cell infiltration level in each CMS subtype, tumor purity, ESTIMATE score, stromal score, and the immune score was evaluated by ESTIMATE algorithm. I. CD274 mRNA expression in CMS subtype. J. The distribution of CMS subtypes in high- and low-risk group.

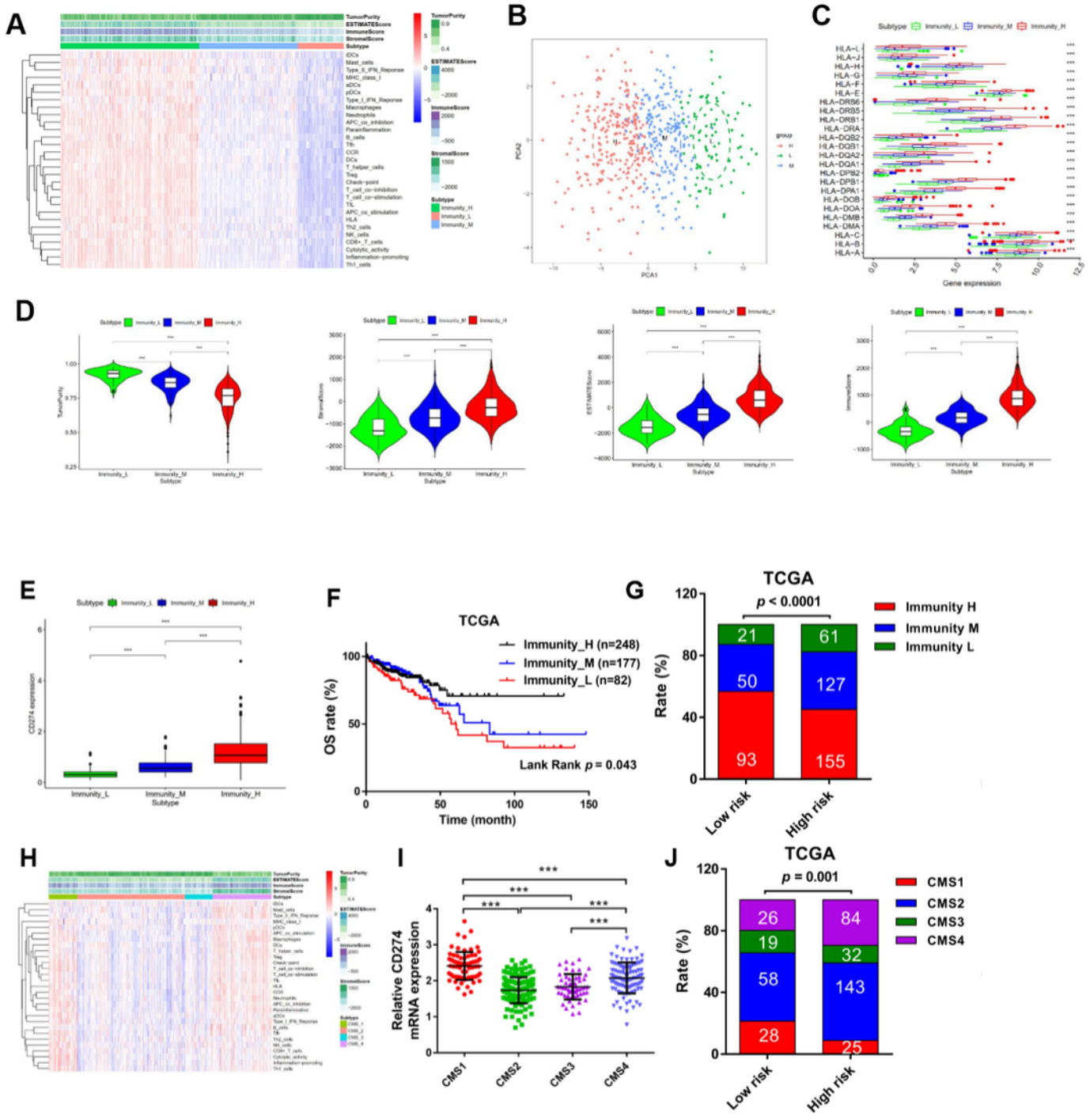

Figure 6

Mutational signature is associated with immune activity in CRC. A. The immune cell infiltration level in each subtype, tumor purity, ESTIMATE score, stromal score, and immune score were evaluated by ESTIMATE. B. PCA analysis of the three clusters. C. Comparison of the expression levels of HLA genes between CRC subtypes (ANOVA test.. D. Comparison of the stromal score, immune score, ESTIMATE score, and tumor purity between CRC subtypes (Mann-Whitney U test.. E. Comparison of PD-L1 (CD274. expression between CRC subtypes. F. Kaplan-Meier analysis of three immunity cluster. G. The distribution of CRC subtypes in high- and low-risk group. H. The immune cell infiltration level in each CMS subtype, tumor purity, ESTIMATE score, stromal score, and the immune score was evaluated by ESTIMATE algorithm. I. CD274 mRNA expression in CMS subtype. J. The distribution of CMS subtypes in high- and low-risk group.

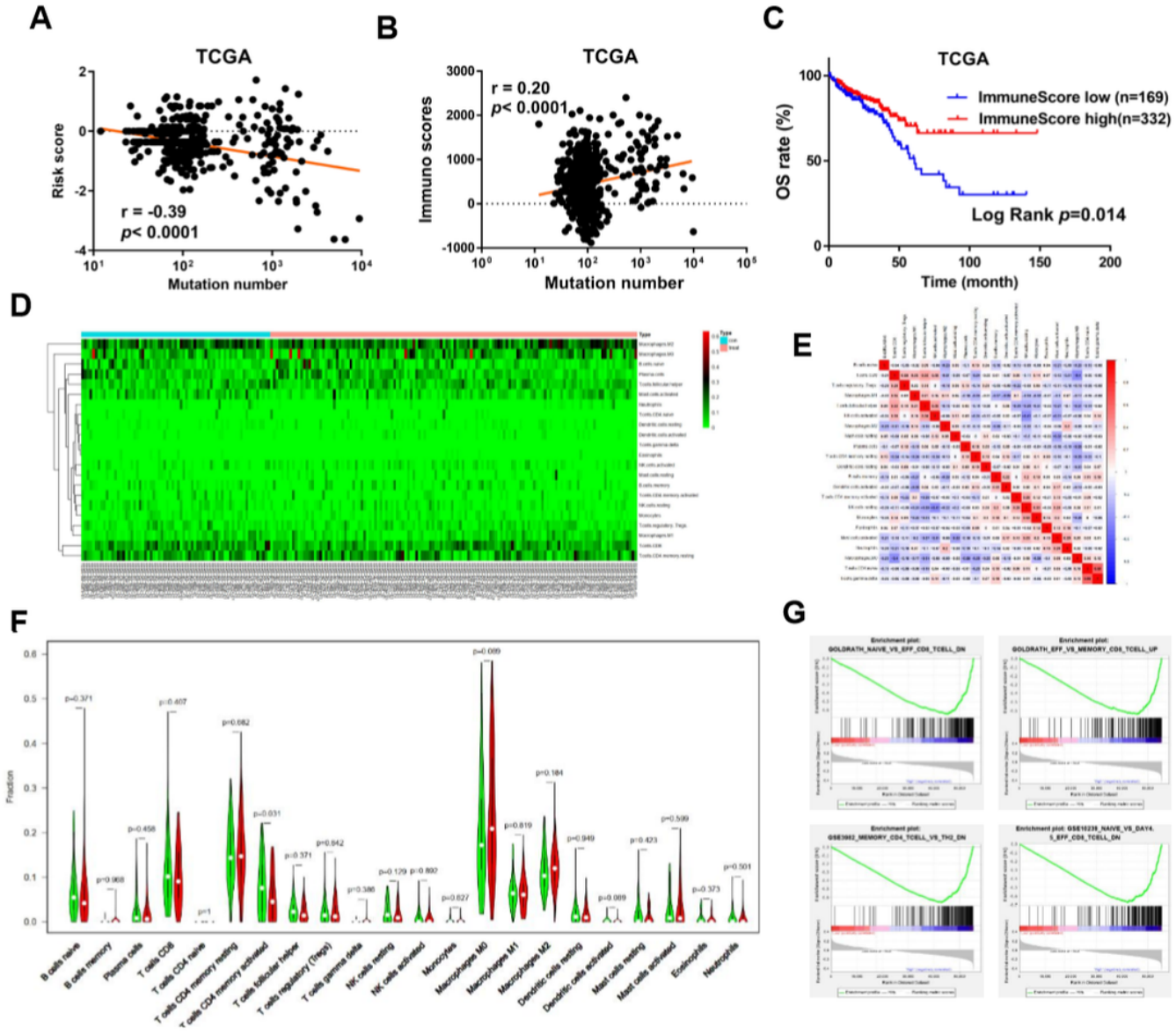

Figure 7

Composition of immune cells at low risk and high-risk tissues in the TCGA cohort. A. Correlation analysis between risk score and TMB in the TCGA dataset. B. Correlation analysis between immune score and

TMB in the TCGA dataset. C. Kaplan-Meier analysis between low and high immune score. D. Fractions of immune cells in 63 high-risk and 63 low-risk groups in the TCGA dataset. E. Correlation of immune cells in the TCGA dataset. F. Comparison of immune cells between high- and low-risk groups in the TCGA dataset. G. GSEA analysis of high- and low-risk group.

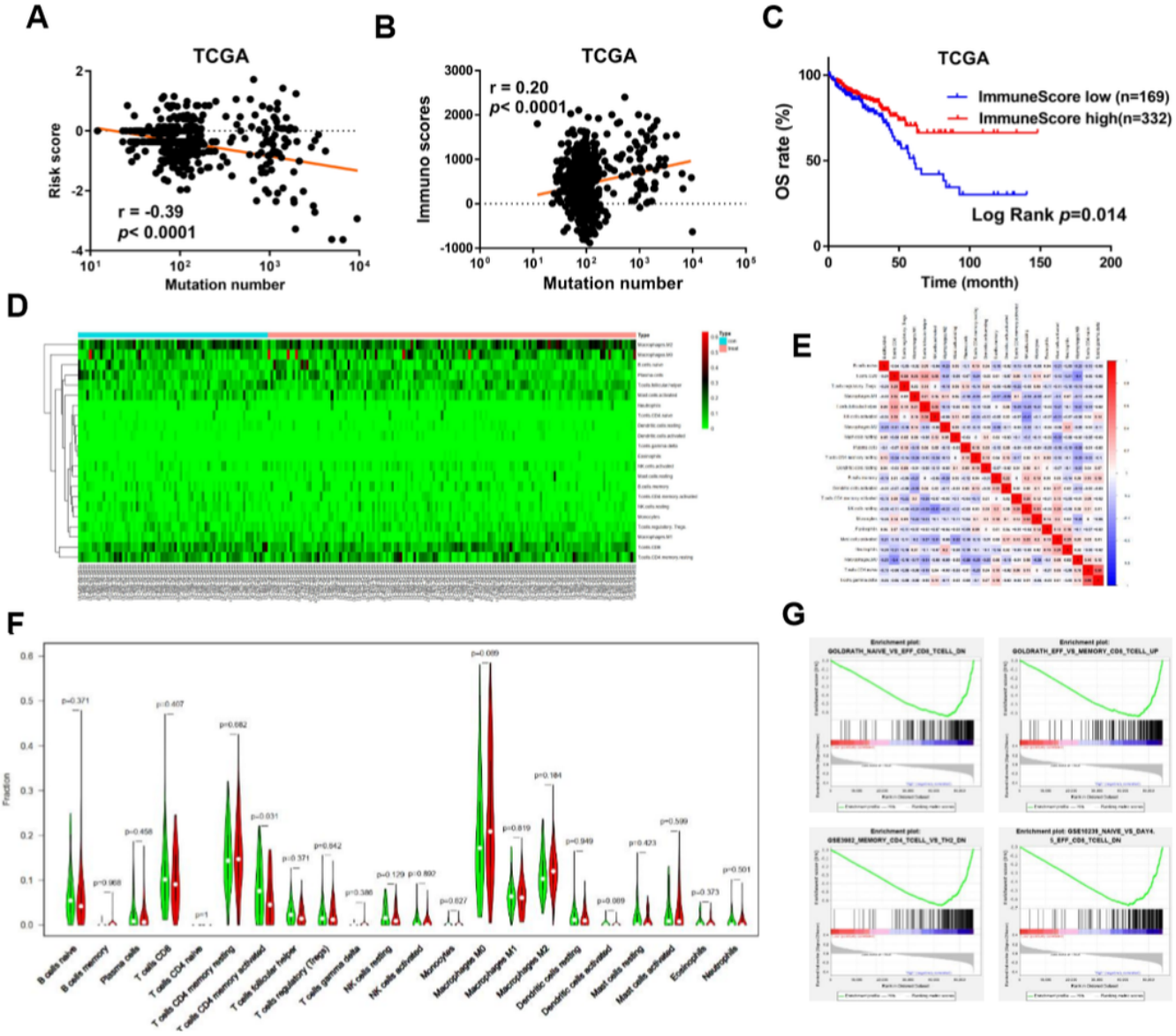

Figure 7

Composition of immune cells at low risk and high-risk tissues in the TCGA cohort. A. Correlation analysis between risk score and TMB in the TCGA dataset. B. Correlation analysis between immune score and TMB in the TCGA dataset. C. Kaplan-Meier analysis between low and high immune score. D. Fractions of immune cells in 63 high-risk and 63 low-risk groups in the TCGA dataset. E. Correlation of immune cells in the TCGA dataset. F. Comparison of immune cells between high- and low-risk groups in the TCGA dataset. G. GSEA analysis of high- and low-risk group.

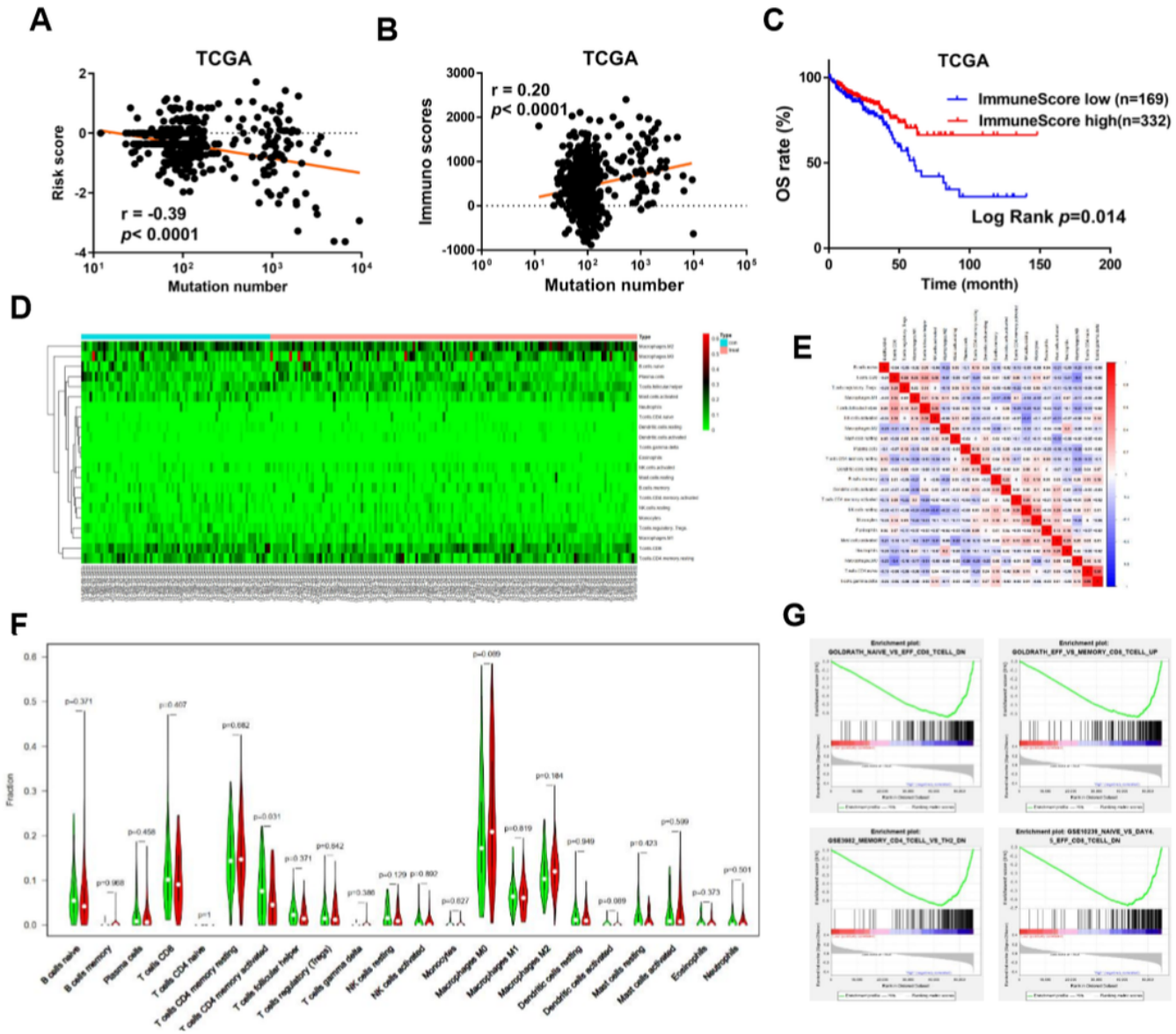

Figure 7

Composition of immune cells at low risk and high-risk tissues in the TCGA cohort. A. Correlation analysis between risk score and TMB in the TCGA dataset. B. Correlation analysis between immune score and TMB in the TCGA dataset. C. Kaplan-Meier analysis between low and high immune score. D. Fractions of immune cells in 63 high-risk and 63 low-risk groups in the TCGA dataset. E. Correlation of immune cells in the TCGA dataset. F. Comparison of immune cells between high- and low-risk groups in the TCGA dataset. G. GSEA analysis of high- and low-risk group.

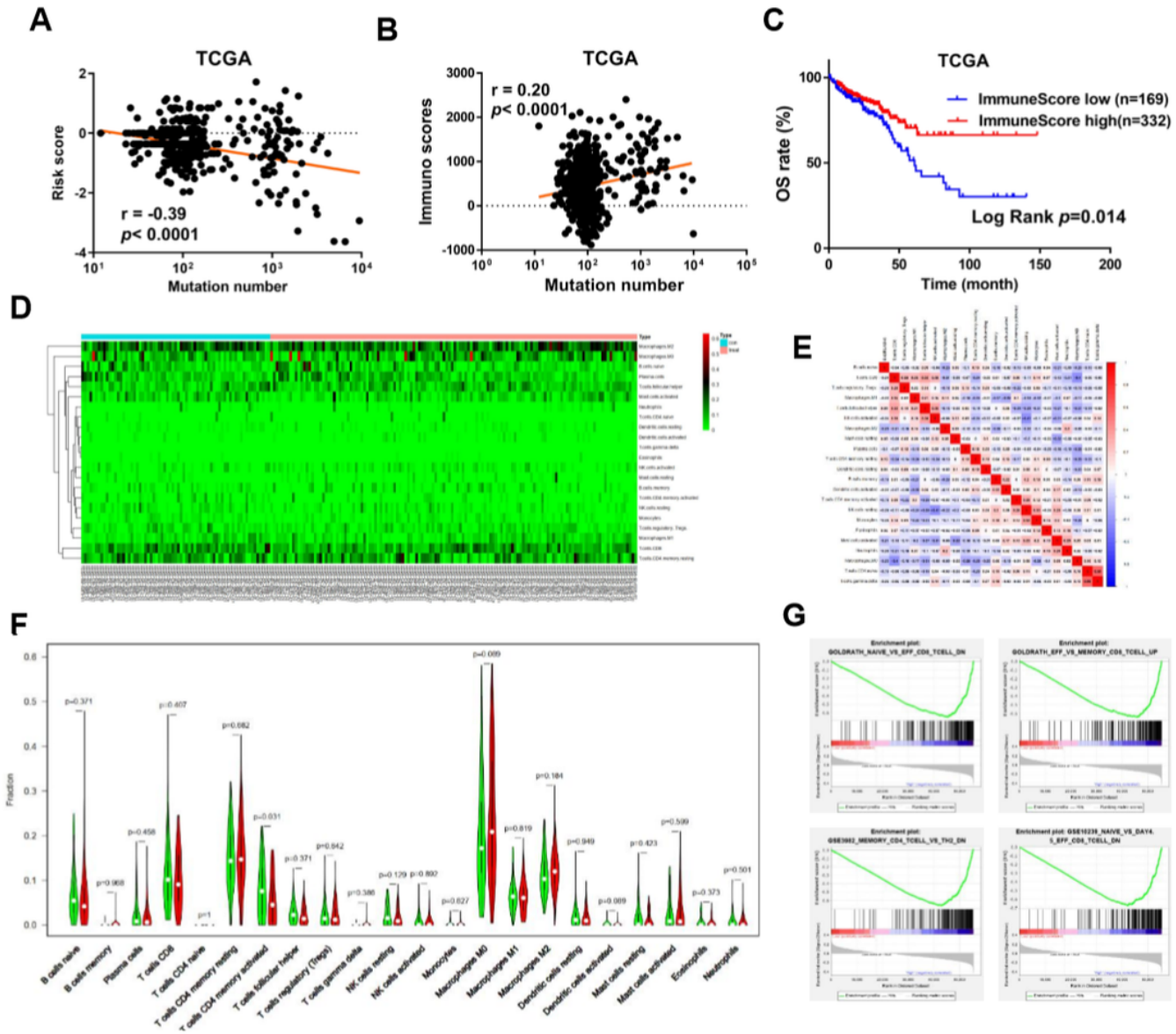

Figure 7

Composition of immune cells at low risk and high-risk tissues in the TCGA cohort. A. Correlation analysis between risk score and TMB in the TCGA dataset. B. Correlation analysis between immune score and TMB in the TCGA dataset. C. Kaplan-Meier analysis between low and high immune score. D. Fractions of immune cells in 63 high-risk and 63 low-risk groups in the TCGA dataset. E. Correlation of immune cells in the TCGA dataset. F. Comparison of immune cells between high- and low-risk groups in the TCGA dataset. G. GSEA analysis of high- and low-risk group.

Supplementary Files

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

- TableS1.png

- TableS1.png
- TableS1.png
- TableS1.png
- TableS2.png
- TableS2.png
- TableS2.png
- TableS2.png
- TableS3.png
- TableS3.png
- TableS3.png
- TableS3.png
- figS1.png
- figS1.png
- figS1.png
- figS1.png
- figS2.png
- figS2.png
- figS2.png
- figS2.png
- figS3.png
- figS3.png
- figS3.png
- figS3.png
- Table1.png
- Table1.png
- Table1.png
- Table1.png
- Table2.png
- Table2.png
- Table2.png
- Table2.png