Research Paper

Identification of candidate genes encoding tumor-specific neoantigens in early- and late-stage colon adenocarcinoma

Chong Wang[1], Wenhua Xue[2], Haohao Zhang[3], Yang Fu[4]

[1]Department of Hematology, The First Affiliated Hospital of Zhengzhou University, Henan, China
[2]Department of Pharmacy, The First Affiliated Hospital of Zhengzhou University, Henan, China
[3]Department of Endocrinology, The First Affiliated Hospital of Zhengzhou University, Henan, China
[4]Department of Gastrointestinal Surgery, The First Affiliated Hospital of Zhengzhou University, Henan, China

Correspondence to: Chong Wang; **email:** fccwangc@zzu.edu.cn
Keywords: neoantigens, colon adenocarcinoma, sequencing, recurrent mutations, machine learning
Received: June 24, 2020 **Accepted:** October 31, 2020 **Published:** January 10, 2021

ABSTRACT

Colon adenocarcinoma (COAD) is one of the most common gastrointestinal malignant tumors and is characterized by a high mortality rate. Here, we integrated whole-exome and RNA sequencing data from The Cancer Genome Atlas and investigated the mutational spectra of COAD-overexpressed genes to define clinically relevant diagnostic/prognostic signatures and to unmask functional relationships with both tumor-infiltrating immune cells and regulatory miRNAs. We identified 24 recurrently mutated genes (frequency > 5%) encoding putative COAD-specific neoantigens. Five of them (*NEB*, *DNAH2*, *ABCA12*, *CENPF* and *CELSR1*) had not been previously reported as COAD biomarkers. Through machine learning-based feature selection, four early-stage-related (*COL11A1*, *TG*, *SOX9*, and *DNAH2*) and four late-stage-related (*COL11A1*, *SOX9*, *TG* and *BRCA2*) candidate neoantigen-encoding genes were selected as diagnostic signatures. They respectively showed 100% and 97% accuracy in predicting early- and late-stage patients, and an 8-gene signature had excellent prognostic performance predicting disease-free survival (DFS) in COAD patients. We also found significant correlations between the 24 candidate neoantigen genes and the abundance and/or activation status of 22 tumor-infiltrating immune cell types and 56 regulatory miRNAs. Our novel neoantigen-based signatures may improve diagnostic and prognostic accuracy and help design targeted immunotherapies for COAD treatment.

INTRODUCTION

Colorectal cancer (CRC) is the third most common cancer and the second leading cause of cancer-related death in the United States [1]. Colon adenocarcinoma (COAD) is a CRC subtype associated with high mortality [2]. Even though survival prognosis has modestly improved over the last three decades, poor survival and high recurrence still entail a pressing need for novel diagnostic biomarkers and therapeutic targets for COAD [3]. Although chemotherapy shows significant therapeutic value, surgery is still the only curative form of treatment for this CRC form [4].

Immunotherapies that boost the ability of endogenous T cells to destroy cancer cells have shown therapeutic efficacy in various human malignancies [5]. The tumoricidal activity of T cells is basically determined by recognition of immunogenic peptides expressed by cancer cells, termed tumor-specific antigens (TSAs) or neoantigens. Comprehensive analysis of transcriptome and whole exome sequencing (WES) data allows identification of tumor-specific mutations giving rise to neoantigens, which can be eventually selected as diagnostic/prognostic biomarkers and therapeutic targets [6]. To screen candidate neoantigens derived from tumor-specific mutations, we evaluated the expression of the corresponding host genes in both

COAD and normal tissues. Only recurrent mutations that were highly expressed in tumor cells and lowly or not expressed in normal cells were selected as potential sources of candidate neoantigens [7].

WES analysis aims to uncover the most frequently mutated genes for a given condition or disease [8]. Nevertheless, multiple genetic variants may be present within individual genes, especially long ones. In tumor suppressor genes, multiple mutations, usually scattered across different loci, may lead to loss-of-function and drive tumorigenesis if both alleles become deficient or inactivated (i.e. the "two-hit" hypothesis). In contrast, mutations in oncogenes are often aggregated, triggering a specific pathogenic function [9, 10]. Therefore, canonical gene-level analysis is not fully adequate to mine disease-specific properties. Instead of selecting genes containing the most mutations, we assessed mutation recurrence among patients at the nucleotide resolution. Subsequently, we integrated RNAseq data to screen recurrent mutations within overexpressed genes in COAD tissues, compared to normal ones. In this manner, we separately compared early and late stage COAD data and contrasted these findings with normal controls to analyze differential gene expression, mutational profiles, and hypothetical functions therefore affected. This approach led us to identify several differentially expressed genes (DEGs) with potential to generate tumor-specific neoantigens. We then addressed the correlations between these DEGs and both regulatory miRNAs and infiltrating tumor cells, and applied a machine learning model to select, among the candidate neoantigen-forming DEGs, molecular signatures for COAD diagnosis and prognosis. Our findings may serve to improve diagnostic and prognostic accuracy in COAD, and help also design targeted immunotherapeutic approaches to increase patient survival.

RESULTS

COAD data selection and clinical information

We retrieved from TCGA a total of 459 COAD samples, including normal controls with clinical information. Of those, 329 and 399 samples underwent RNAseq and WES analysis, respectively. In addition, miRNA sequencing data was retrieved from 261 samples. A total of 20,529 mRNAs and 2,113 miRNAs were identified in the aggregated sequencing data. Clinical information, including gender, stage, vital status, and survival time are shown in Figure 1.

It can be seen that the number of male patients is slightly higher than the number of female patients at both early and late stages. The average age of first diagnosis shows no difference between genders, although late stage cases are more likely to be diagnosed in younger groups. As expected, the mortality ratio was significantly higher for late stage cases ($P = 1.639e-05$; Fisher's exact test). There were no differences in overall survival (OS) time between genders.

Differential gene expression analysis

We conducted differential analysis of mRNA expression profiles between normal vs total, early stage, and late stage tumors, as well as between early vs late stage tumors samples (Table 1 and Figure 2A–2C).

As shown in Table 1 and Figure 2A, 2B, a balanced distribution of up- and down-regulated genes was detected, regardless of tumor stage, upon comparison with normal control data. In contrast, most DEGs between early and late stage tumors were upregulated (Figure 2C). Since only 8 normal samples were included in the miRNA data, we only compared miRNA profiles between early and late stage tumors. As shown in Figure 2D, most differentially expressed miRNAs were downregulated. This finding seems to be consistent with the observed DEG pattern, considering that negative, rather than positive, regulation is usually exerted by miRNAs on protein-coding transcripts.

Gene clustering and functional analysis

The identified DEGs exhibited diverse expression patterns among normal, early-stage, and late-stage samples. As seen in Figure 3A, tumor and normal samples were separated into different groups based on DEG profiling. To some extent, early and late stage patients also presented some distinctions. Principal component analysis (PCA) was then used to visualize the distribution of all samples based on the first two principal components (Figure 3B). Consistent with the heatmap analysis, the result showed that tumor samples showed diverse patterns compared with normal ones.

To investigate the cellular functions regulated by the DEGs, we conducted Gene Ontology (GO) functional enrichment analysis (Figure 4).

Comparison between early stage and normal control samples revealed many interacting DEGs enriched mainly in homeostasis-related functions (Figure 4A, 4B). Between late stage patients and normal controls, the predominant DEG-enriched functional modules included 'homeostasis' and 'multiple system development' (Figure 4C, 4D). It implies that as the disease progresses, different biological functions are dynamically interfered.

Stage-specific co-expression network analysis

We next applied the Pearson's correlation algorithm to assess potential regulatory influences exerted by differentially co-expressed miRNAs on the identified DEGs. As shown in Figure 5A, a total of 4,656 edges and 362 nodes were identified in the network, and four significant modules were extracted using the MCODE plugin.

The principal biological GO terms within the network were then obtained using the BiNGO plugin [11]. The process of 'microtubule-based transport' was seemingly activated, since all the involved genes were upregulated in late stage patients. In contrast, 'negative regulation of angiogenesis' was apparently inhibited, as all the corresponding regulatory components were down-regulated. Lastly, significant enrichment in both upregulated and downregulated DEGs was detected for 'regulation of DNA repair' and 'microtubule-based process'.

Functional enrichment analysis was also conducted to assess the biological roles of the 19 differentially expressed miRNAs in relation to their target genes (Supplementary Table 1). Results showed that 86 DEGs were targeted by these 19 miRNAs, exerting a predominant regulatory influence on cell cycle dynamics (Figure 5B).

Recurrent somatic mutation selection

A mutation profile analysis of WES data from COAD patients revealed that missense mutations were the most dominant variants (Figure 6). In turn, a transition (Ti)–transversion (Tv) bias was one of the significant features in both whole-genome sequencing (WGS) and WES data. In COAD-WES data, the C>T transition was the most distinct feature, which suggests an essential role for oxidative DNA damage in COAD pathogenesis [12].

On gene-level analysis, the top 10 mutated genes included *TTN*, *APC*, *MUC16*, *SYNE1*, *TP53*, *FAT4*,

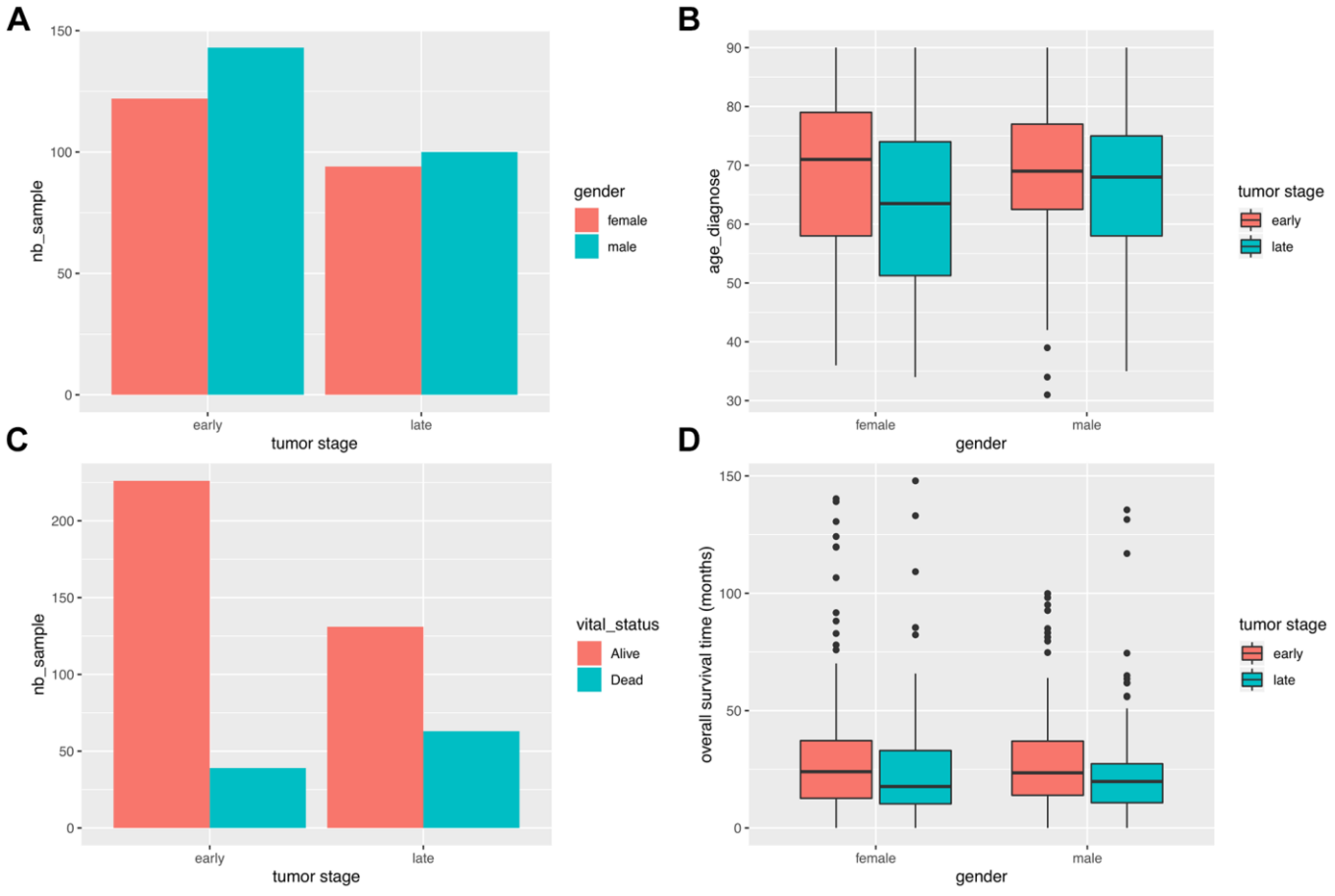


Figure 1. Clinical characteristics of COAD patients. (**A**) Stage distribution between male and female COAD patients. (**B**) Gender-based distribution of age of first diagnosis and stage. (**C**) Vital status distribution according to stage. (**D**) Survival time distribution for gender and stage.

Table 1. Number of differentially expressed genes among conditions and stages.

	N vs. T	N vs. E	N vs. L	E vs. L
Upregulated	1026	1027	1027	482
Downregulated	1027	1027	1027	369
Low cutoff	-2.21	-2.28	-2.12	-0.2
High cutoff	1.25	1.24	1.29	0.41

N, T, E, and L represent normal, tumor (any stage), early stage, and late stage, respectively. Low and high cutoffs were determined using 95% confidence interval limits across the logFC values of all genes.

KRAS, *RYR2*, *OBSCN*, and *PIK3CA*. Then we investigated the recurrence of each mutation across patients from each stage. All recurrent somatic mutations with a frequency larger than 5% were selected.

Identification of candidate neoantigen-coding genes

Under the assumption that genes with one or more recurrent mutations potentially leading to neoantigen production are exclusively overexpressed in tumor tissues and not in normal ones, we selected the recurrent mutations within differentially overexpressed genes between early and late stage patients. We eventually identified 24 genes harboring recurrent somatic mutations in at least 5% of patients from either stage (Supplementary Tables 1, 2). The host genes and their corresponding mutational frequency in either stage are shown in Figure 7A.

As seen in Figure 7A, the host genes of the candidate neoantigens were different from the top-mutated genes identified by the canonical protocol. A main reason for this is that the most mutated genes might be either silent in tumor tissues or expressed at similar levels than normal ones. Among the 24 genes identified by our

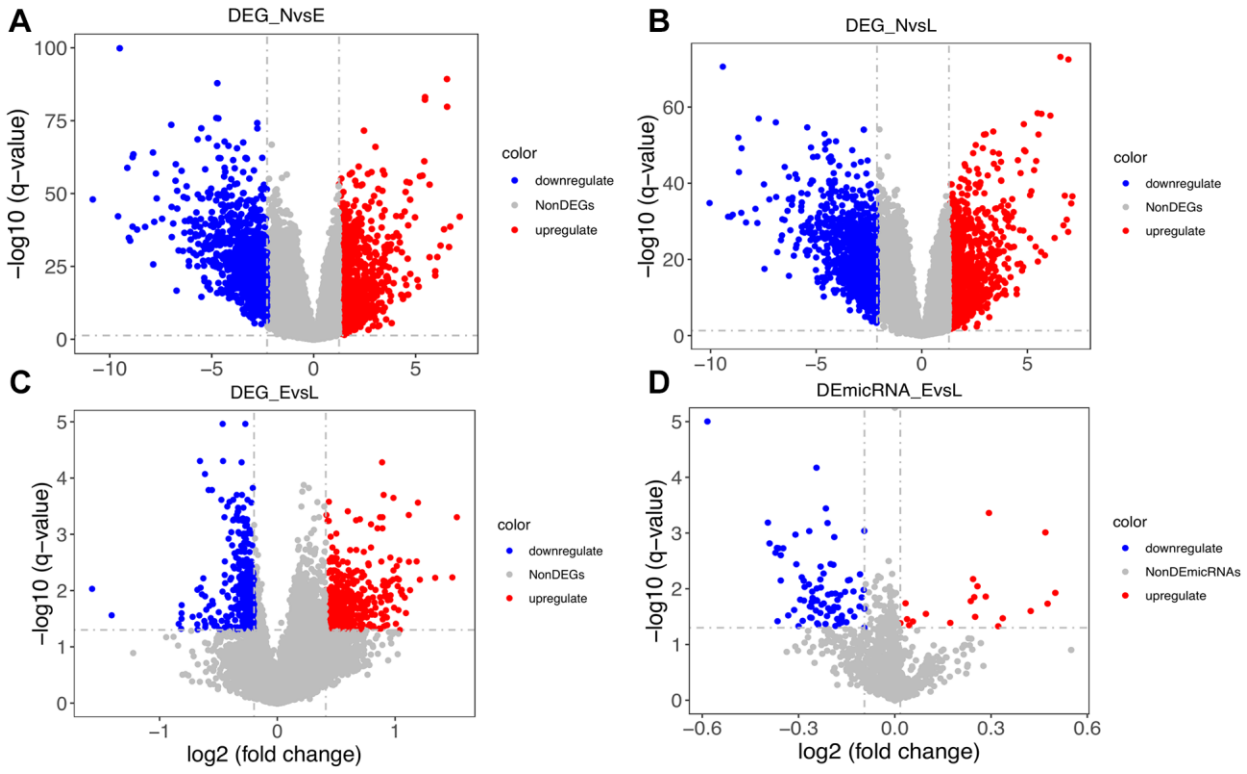


Figure 2. Distribution of differentially expressed genes. (**A**) Distribution of DEGs between normal (N) and early-stage (E) samples. (**B**) Distribution of DEGs between N and late-stage (L) samples. (**C**) Distribution of DEGs between E and L samples. (**D**) Distribution of differentially expressed microRNAs between E and L samples.

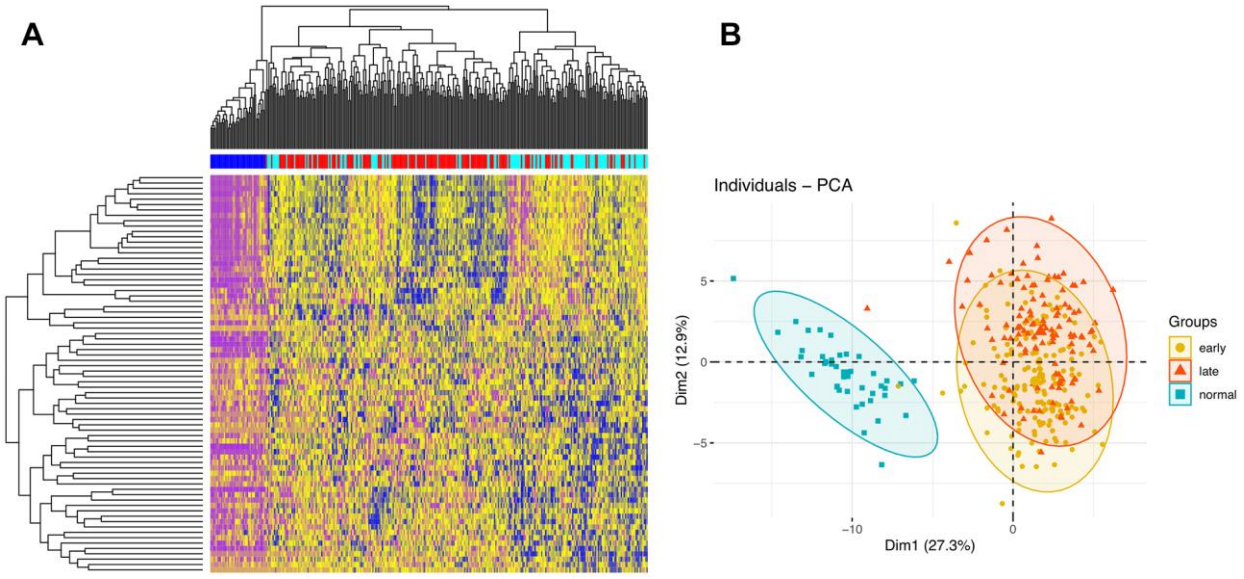


Figure 3. Clustering analyses. (**A**) Hierarchical clustering analysis of DEGs. The intersections of DEGs of early vs late stage are used to cluster samples. Normal, early stage, and late stage samples are marked by dark blue, red, and light blue, respectively. (**B**) PCA of samples.

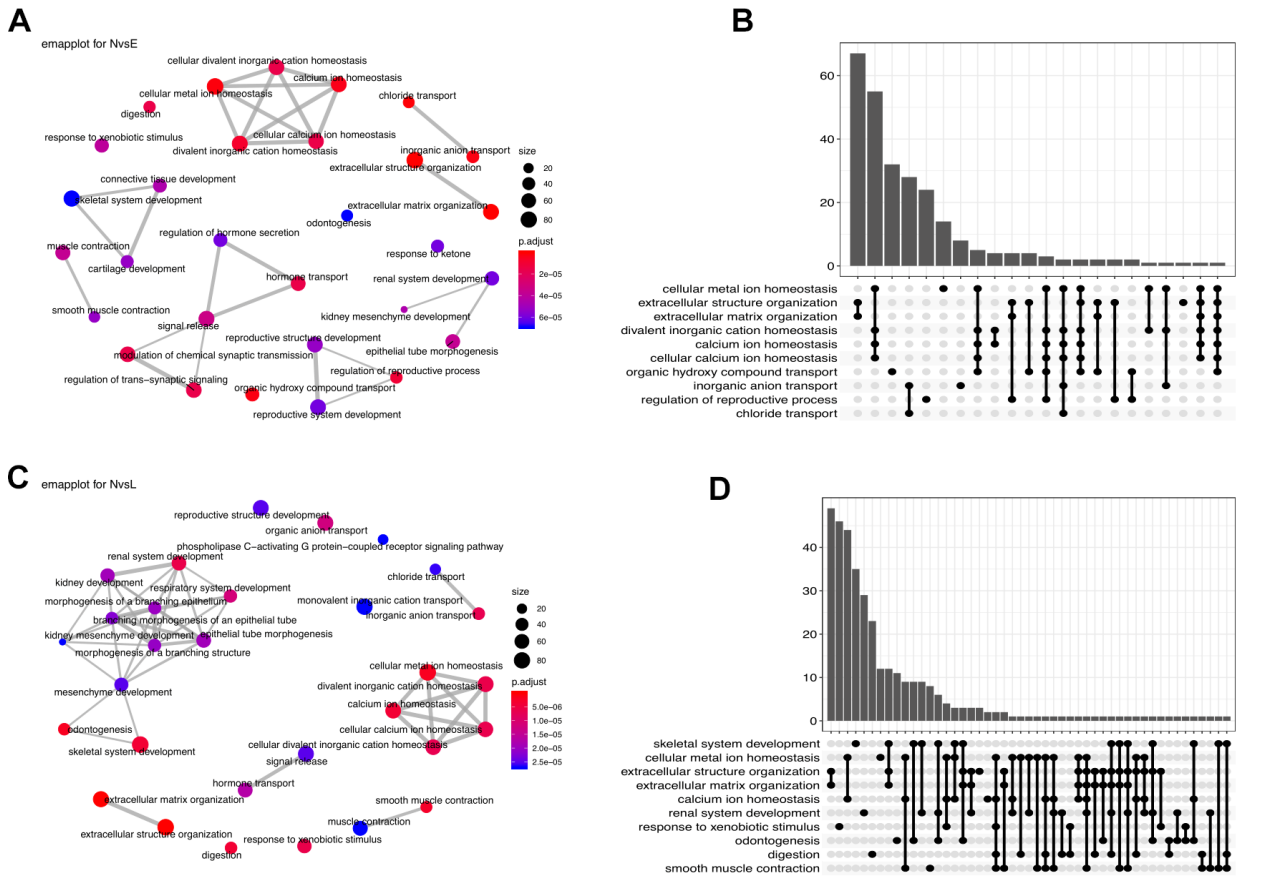


Figure 4. Functional enrichment analysis. (**A**) Enrichment network (emapplot)of functions regulated by the NvsE DEGs. (**B**) Upset plot of the top 10 NvsE-related functions. (**C**) Emapplot enrichment network of functions held by the NvsL DEGs. (**D**) Upset plot of the top 10 NvsL-associated functions.

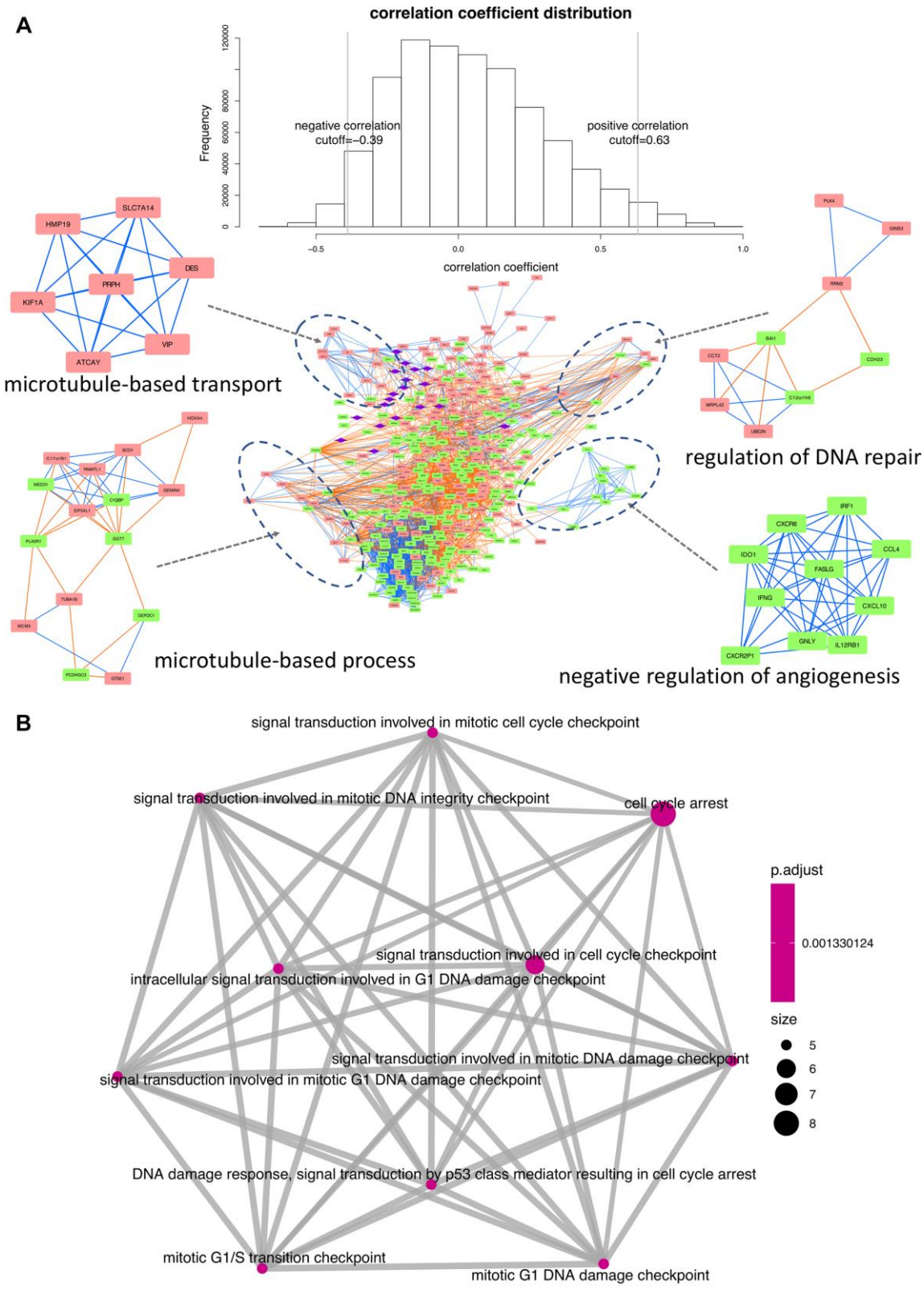


Figure 5. Interaction network and functional enrichment analysis of DEGs and miRNAs between early and late stage. (**A**) Co-expression network analysis. Purple diamonds represent miRNAs; red and green rectangles represent upregulated and downregulated DEGs, respectively. Orange and blue lines indicate, respectively, positive and negative correlations between nodes. Top-enriched functions are indicated under the corresponding modules. (**B**) Functional network depicting DEG-enriched processes regulated by the 19 differentially expressed miRNAs.

protocol, *NEB* and *DNAH5* were found to be mutually exclusive, especially in late-stage patients.

We next performed KEGG pathway enrichment analysis for the 24 genes to evaluate their functional properties (Figure 7B). Among them, 5 genes involved in collagen synthesis, i.e. *COL11A1*, *COL12A1*, *COL27A1*, *COL5A1*, and *COL7A1*, were enriched in the protein digestion and absorption pathway.

Correlation of COAD-associated neoantigen genes with tumor infiltrating immune cells and predicted miRNAs

The Pearson's correlation coefficients between the 24 neoantigen-associated DEGs and COAD-infiltrating immune cells (determined by RNA-seq data) are shown in a heatmap on Figure 8A. Intuitively, we observed two correlation patterns within the 22 immune cells analyzed. The most prominent, positively correlated immune cells consisted of neutrophils, macrophages, dendritic cells, activated mast cells, and naïve and activated T cells. In contrast, B cells, resting mast cells, resting T cells, monocytes, plasma cells, and eosinophils were mainly negatively correlated with the candidate neoantigen-forming DEGs.

We further investigated the correlation between the 24 selected neoantigen genes and miRNAs. We combined 47 validated miRNAs from three miRNA databases and 9 predicted miRNAs based on expressional correlation. These 9 predicted miRNAs

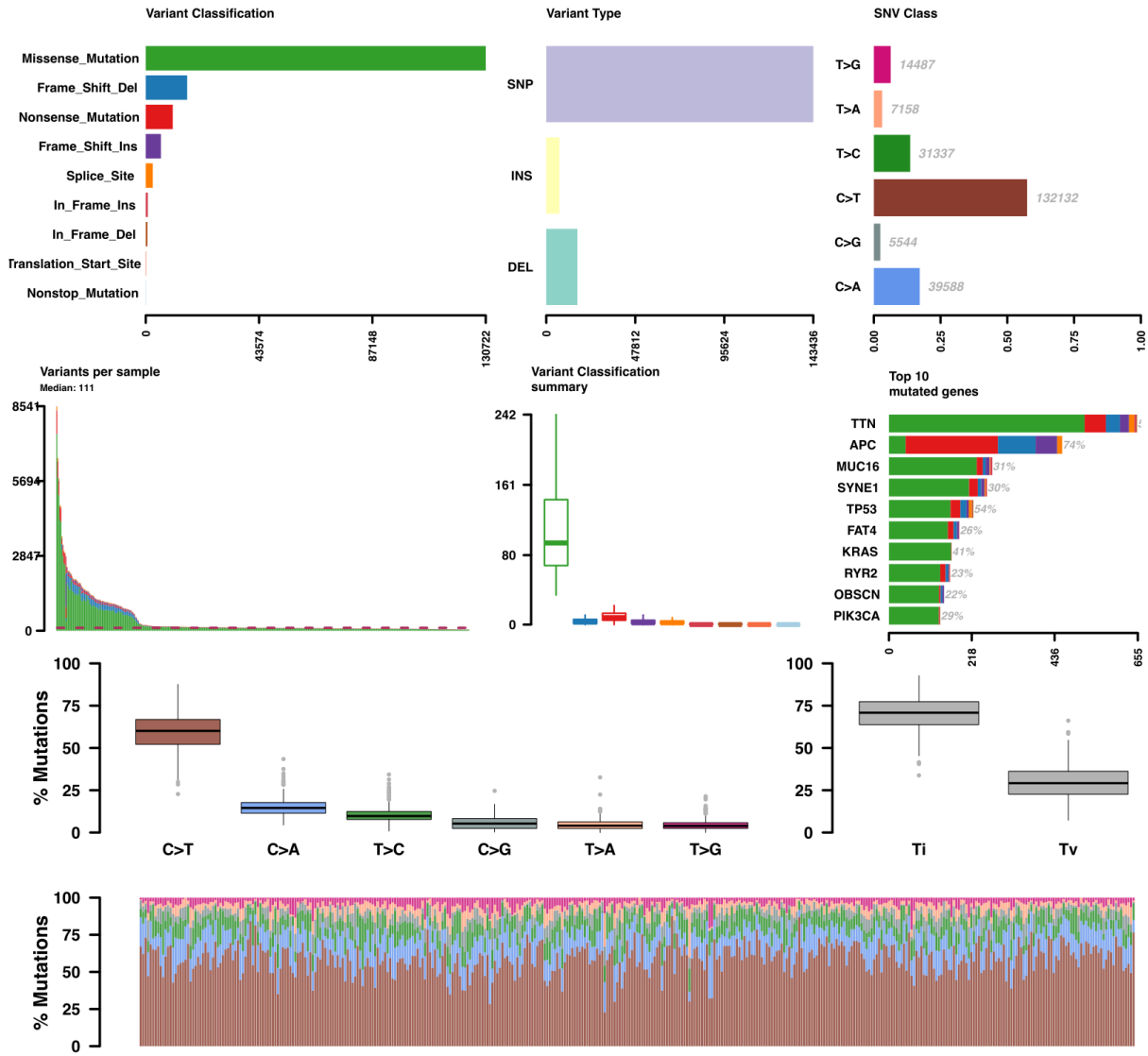
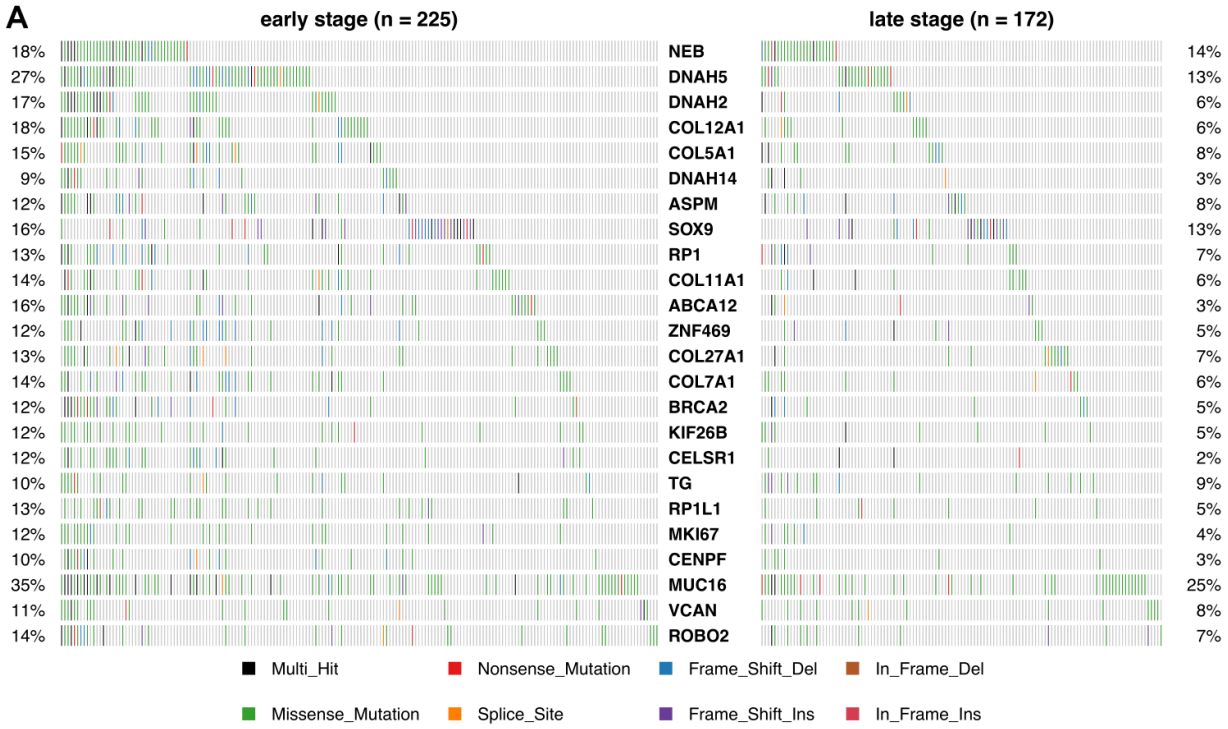


Figure 6. Somatic variant analysis of COAD-TCGA data. Variants per sample are shown as a stacked barplot and variant types as a boxplot summarized by variant classification.



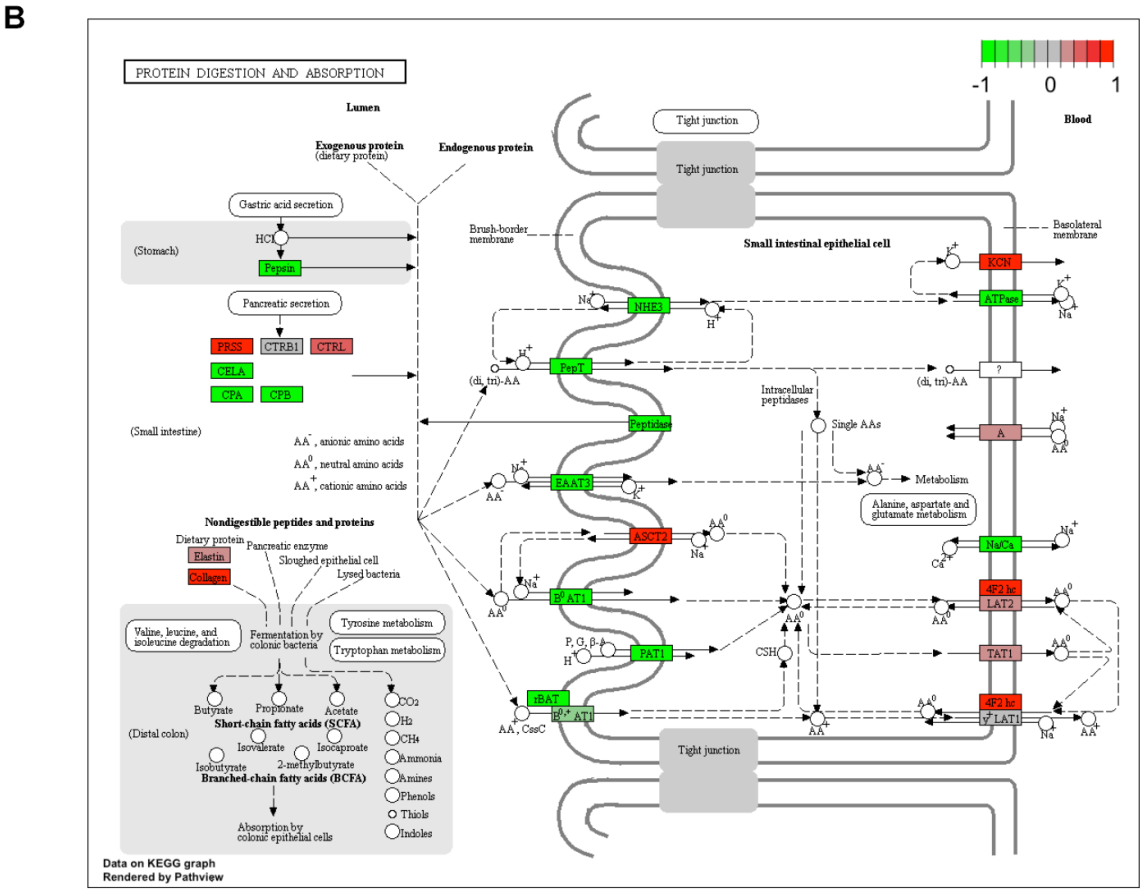


Figure 7. Mutational profile and KEGG pathway analysis of candidate COAD neoantigen-related DEGs. (**A**) Mutation frequency data. The percentage of patients harboring the color-coded variations listed at the bottom are indicated on the left and right sides. (**B**) KEGG pathways enriched in the 24 candidate neoantigen-related genes. Red and green boxes indicate up- and down-regulated genes, respectively.

were all correlated with at least one of the 24 neoantigen-related host genes. A correlation matrix between the 24 neoantigen genes and the 56 miRNAs is shown in Figure 8B. It can be seen that six host genes (*ZNF469*, *COL5A1*, *COL12A1*, *KIF26B*, *COL11A1*, and *VCAN*) were the most significantly regulated by the overall miRNA pool, especially by MIMAT0000440.

Although similar correlation patterns were observed for both predicted and validated miRNAs, stronger links were identified for the 9 predicted miRNAs. This suggests their potential involvement in the regulation of COAD progression.

Diagnostic model construction

Since the candidate neoantigen-forming DEGs were all overexpressed and tumor-specific, we evaluated their potential as diagnostic markers. Applying the random forest algorithm, we obtained four early stage-related signatures, represented by *COL11A1*, *TG*, *SOX9*, and *DNAH2*. Likewise, four late stage-related signatures, namely *COL11A1*, *SOX9*, *TG*, and *BRCA2*, were also selected (Figure 9). Except for *DNAH2* and *BRCA2*, all the other genes were shared by the two tumor stages. This suggests a relatively stable expression for these DEGs during COAD progression. However, the stage-specific signatures indicate that some differences exist between the two stages.

Using the selected genes as signatures, the trained predictor was tested against a validation dataset. The accuracy of predicting early and late stage cancer was 100% and 97%, respectively. These data suggest that the neoantigen-related DEGs identified by our protocol have the potential to be exploited as both therapeutic targets and diagnostic biomarkers.

Evaluation of neoantigen-related DEGs as survival indicators in COAD

The potential impact of the 24 neoantigen-associated DEGs on survival prognosis was investigated using stepwise regression. Eight genes, i.e. *ZNF469*, *VCAN*, *RP1*, *MUC16*, *KIF26B*, *COL5A1*, *COL12A1*, and *CENPF*, were chosen to build a signature for overall survival (OS). In turn, another 8-gene set, including *RP1L1*, *VCAN*, *SOX9*, *KIF26B*, *MUC16*, *COL5A1*, *COL11A1*, and *CELSR1*, was selected as a relevant prognostic signature for disease-free survival (DFS).

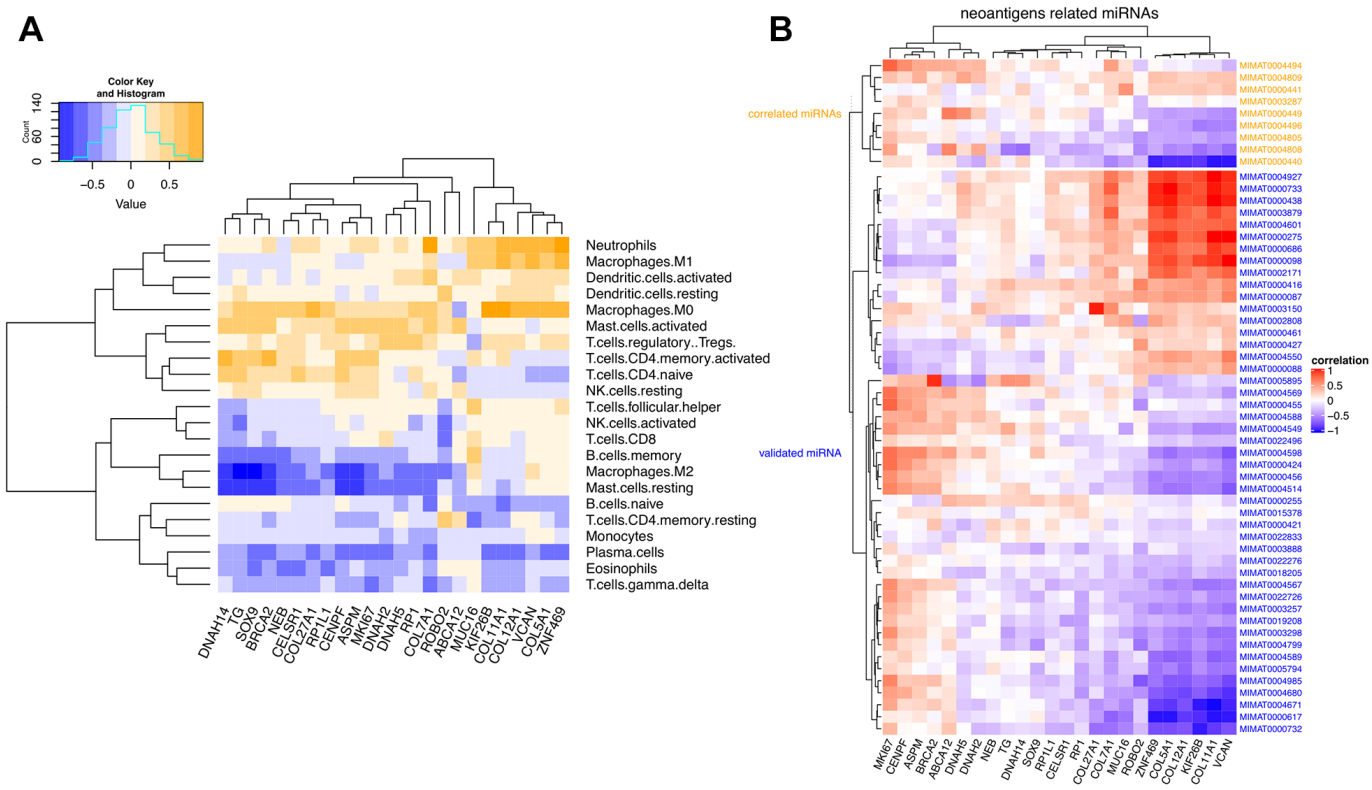


Figure 8. Correlation matrix between neoantigen-associated DEGs and immune cells and miRNAs. (**A**) Correlation between COAD-related neoantigen genes and immune infiltrating cells. (**B**) Correlation between COAD-related neoantigen genes and miRNAs. Validated and predicted miRNAs are marked in blue and orange, respectively.

Based on expression data for these signature genes, high- and low-risk patient groups were determined by the regression model prior to construction of Kaplan-Meier curves comparing OS and DFS (Figure 10A, 10B).

Despite a non-significant log-rank P value for OS (0.07), the 8-gene set still has valuable clinical implications. That the P-value did not reach statistical significance is attributable to the mixture of patients surviving no more than 3 years. However, the survival difference is obvious for patients surviving longer than 3 years. Moreover, another 8-gene set was significantly predictive of DFS (P-value=0.0025). These results show that differential expression of signature genes can be used to successfully predict OS and DFS among high- and low-risk patients.

Comparative analysis of neoantigen-related DEG expression between COAD and multiple cancers

To assess whether the 24 overexpressed host genes harboring recurrent mutations associated with candidate neoantigens identified herein were specific to COAD, we evaluated their expression in 10 other cancers (Figure 11).

Genes with FC > 2 and FDR > 0.05 were retained for analysis. Normalized RSEM values were used as expression values in GSCALite, but log2(x+1) transformed RSEM normalized count was used in our RNAseq analysis. Because of the different quantification methods, some slightly overexpressed genes, such as *NEB* and *DNAH2*, whose log2FC values were respectively 2.2 and 2.6 in NvsT analysis, were missed by the GSCALite analysis.

We found that most host genes were overexpressed in multiple cancers besides COAD, including *BRCA*, *LUAD*, *HNSC*, and *STAD*. Other genes, including *DNAH5*, *COL7A1*, *COL27A1*, *RP1L1*, and *ROBO2*, seemed instead to be COAD-specific, as they were silent or even downregulated in other cancers (Figure 11A). Most of these genes were however recurrently mutated in COAD and other digestive system cancers such as stomach adenocarcinoma (STAD) and esophageal carcinoma (ESCA). They were also highly mutated in lung adenocarcinoma (LUAD), but this may be attributed to the large number of mutations characteristic to this entity (Figure 11B).

Pathway analysis indicated that some biological functions were significantly impacted by mutations in the host genes analyzed. For instance, apoptosis, cell cycle, DNA damage response, EMT, and hormone_AR/ER were all simultaneously or alternatively activated and/or inhibited by the different mutations (Figure 12).

DISCUSSION

Immunotherapy approaches have gained prominence in the treatment of many cancers, especially leukemia, lymphoma and lung, kidney, and bladder cancer. Seven standard treatments, including surgery, radiofrequency ablation, cryosurgery, chemotherapy, radiation, target therapy, and immunotherapy, are currently used for COAD. Resection and anastomosis are the major strategies for early stage (stage I and II) colon cancer patients, while chemotherapy may be further indicated for stage III patients. Immunotherapy is currently used to treat stage IV and recurrent colon cancer patients [13], and its potential use in early stage patients remains controversial. This is largely due to uncertainty about both the mutational changes impacting early to late stage progression, as well as the immune influences that shape this transition.

Tumor neoantigens are modified proteins expressed on the surface of tumor cells and recognized as “non-self” or foreign by cells of the immune system [14]. Generally, these foreign proteins derive from tumor-specific mutations that change the original peptide sequence and/or structure. Tumor neoantigens have attracted a great deal of attention as potential targets for immunotherapy, including individualized or broad-spectrum cancer vaccines. To reduce the risk of potentially severe autoimmune reactions, proper screening protocols are required to identify clinically actionable tumor neoantigens. Ideally, the mutations that give rise to neoantigens should be recurrently observed in a significant fraction of patients so that the therapy can be broadly applied. High-throughput sequencing technologies provide the opportunity for large-scale screening of potential neoantigens in cancer patients. For instance, approaches using WES data alone or together with gene expression data have been used in clinical trials of checkpoint inhibitors [7].

In this study, we integrated WES and RNA-Seq data to screen candidate neoantigen-hosting genes in colon cancer. After selecting stage-specific DEGs through RNA-Seq analysis, recurrent mutations in the overexpressed genes were further selected through WES analysis. Based on mutational recurrence rates among COAD specimens, we then applied the random forest, a supervised machine learning model, to construct two gene signatures that showed high diagnostic accuracy to discriminate early and late tumor stages. In turn, survival analyses showed prognostic differences between patients with and without these recurrent mutations.

We identified *SOX9* and *COL11A1* as relevant biomarkers of diagnosis and survival prognosis in COAD. *COL11A1* expression is upregulated in many cancers, including colorectal, breast, and ovarian cancer, and head and neck squamous cell carcinoma [15–19]. Previous studies suggested that *COL11A1* regulates tumor progression through the APC/beta-catenin pathway, and inhibits apoptosis by modulating the NFkB pathway [20, 21].

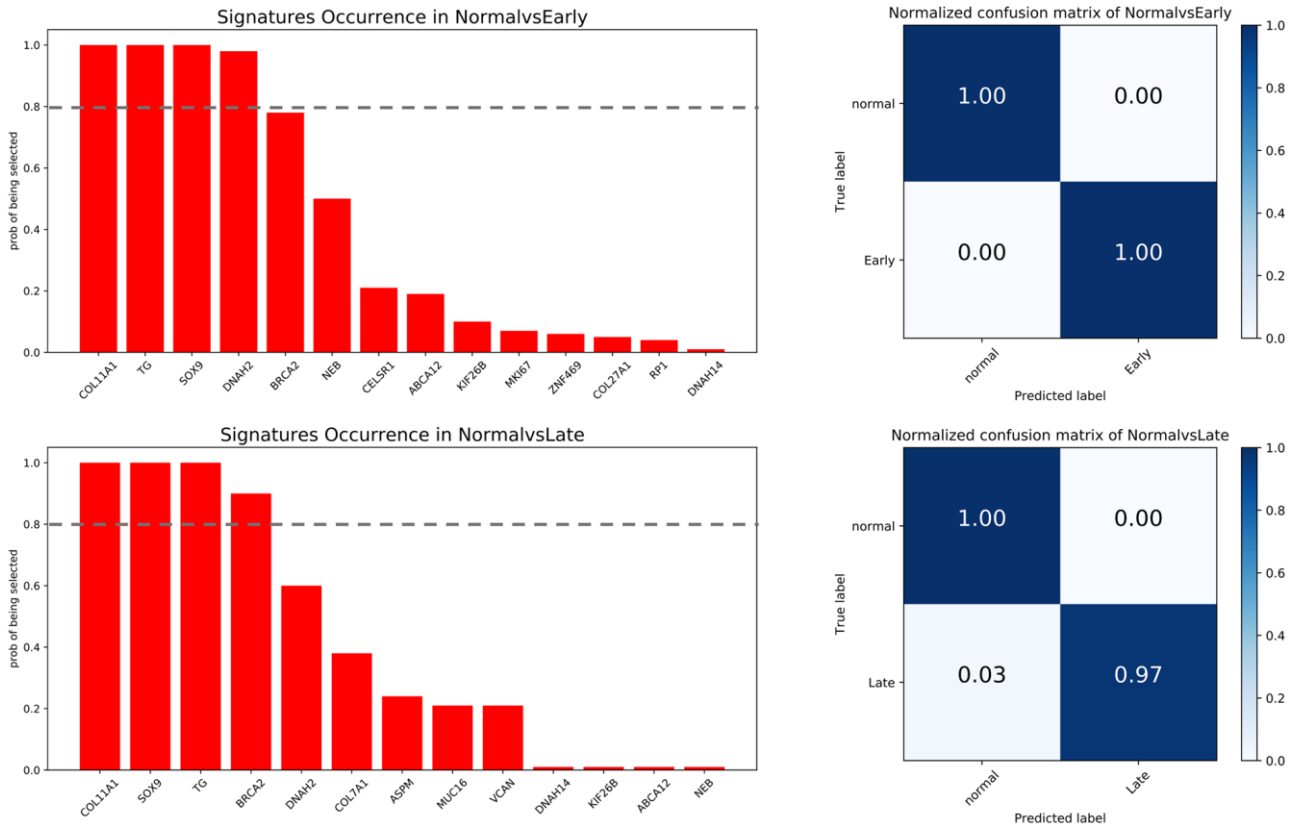


Figure 9. Feature selection and confusion matrices. Top: normal vs early stage. Bottom: normal vs late stage. The x-axis of barplot graphs lists featured genes and the y-axis indicates how many times each feature is selected over 100 permutations. The grey dash line represents the significance cutoff (0.8). The x-axis in the confusion matrices represents the predicted labels and the y-axis represents the true labels.

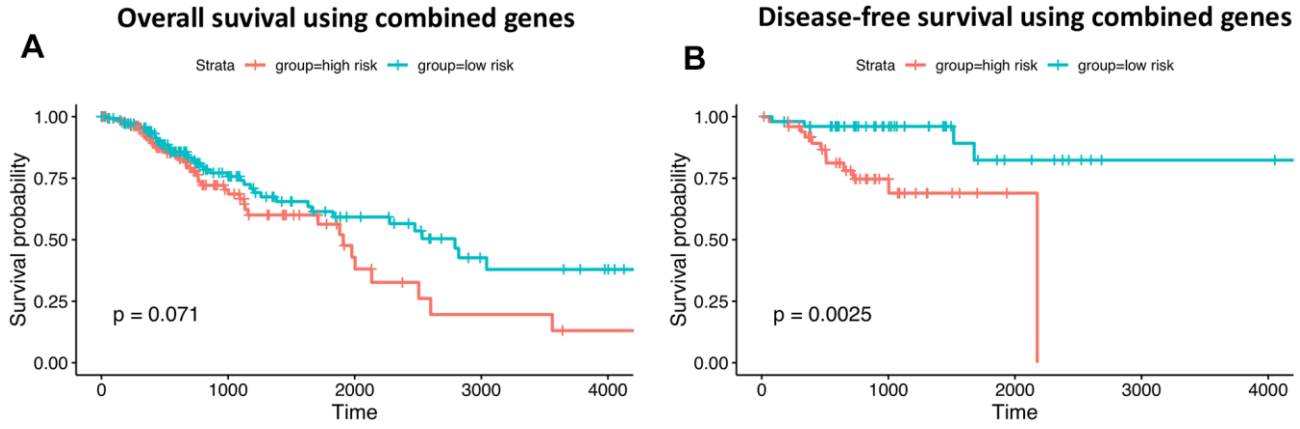


Figure 10. Prognostic ability of neoantigen-related gene signatures in COAD. (**A**) Overall survival. (**B**) Disease-free survival. Time is expressed as days in the graphs' x-axis.

Elevated *SOX9* expression was characteristic of COAD and this gene was thus included in the diagnostic signatures for early and late stage tumors. *SOX9* overexpression is associated with increased mortality in many cancers, including colon and rectal cancers [22–24]. *SOX9* is involved in multiple functions that promote cancer progression, such as proliferation and transformation, and resistance to apoptosis and chemotherapy [23, 25].

Among the 24 neoantigen-related genes identified herein, *COL11A1*, *COL12A1*, *COL27A1*, *COL5A1*, and *COL7A1* participate in the synthesis of various collagen types. Excessive, abnormal deposition of collagen chains may lead to enhanced activation of fermentation by colonic bacteria. This phenomenon has been linked to colon carcinogenesis; as colon cancer progresses, the activity of colonic microorganisms becomes more intense, which would intensify the digestion and absorption of proteins [26–29]. Other markers identified by us as candidate host genes for neoantigens have already been postulated as driving factors in COAD. These include *BRCA2* [30], *MKI67* [31], *MUC16* [32], *RP1* [33], and *VCAN* [34].

Novel findings of our study include the new diagnostic and prognostic signatures, the observed correlation between the neoantigen-associated genes and specific tumor-infiltrating immune cell populations, and the predicted regulatory influences exerted on the former by several miRNAs. Moreover, we identified several DEGs that had not so far been associated with COAD. These include *NEB*, which predicted high survival risk in COAD patients, and *DNAH2* and *ABCA12*, which are considered essential prognostic indicators for ESCA. Two other DEGs, *CENPF* and *CELSR1*, were in turn selected in this study as prognostic indicators of OS and DFS in COAD patients.

ABCA12 is a member of the ATP-binding cassette (*ABC*) family of transporters, which are essential mediators of chemoresistance [35]. *DNAH2* encodes for

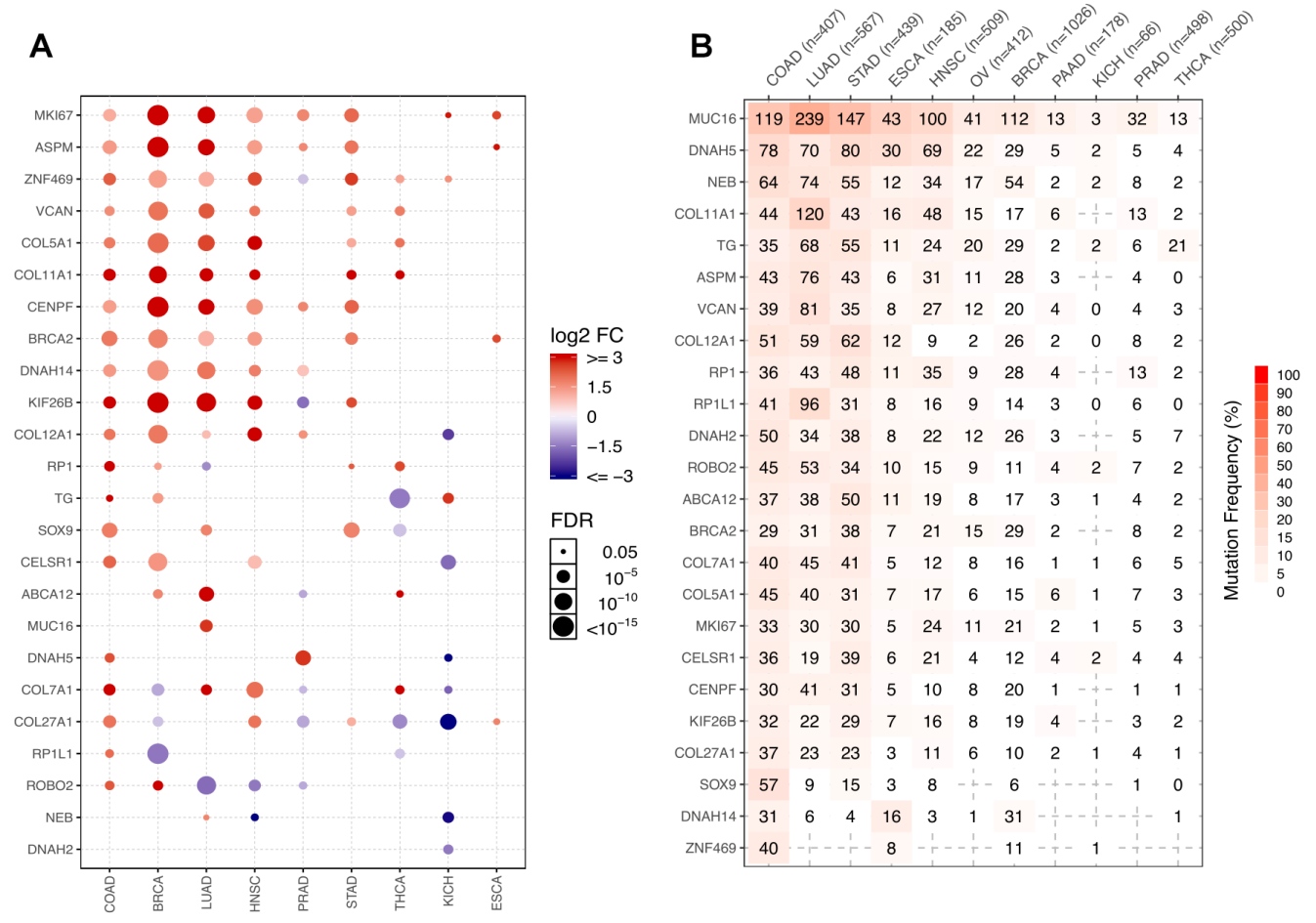


Figure 11. Representation of the 24 COAD-related neoantigen genes in other cancers. (**A**) Comparative expression analysis of the 24 host genes across 9 cancers. (**B**) Number of mutated samples across 11 cancers.

the dynein heavy chain, which also has ATPase activity [36]. *CENPF* is a widely studied driver gene in multiple cancers, such as gastric [37], prostate [38], breast [39], and bladder [40] cancer. Functional experiments revealed positive effects of *CENPF* on cellular proliferation, migration, and invasion [41]. *CELSR1* was shown to promote progression and paclitaxel resistance of ovarian cancer *in vitro* and *in vivo* [42]. Our data thus reveal novel associations between these genes and COAD, some of which present potential relevance as diagnostic/prognostic indicators.

Somatic mutations may lead to generation of neoantigens for T-cell recognition, leading in turn to increased recruitment of various kinds of immune cells [5]. Therefore, uncovering the relationship between tumor neoantigens and infiltrating immune cell populations will greatly boost the efficacy of immunotherapy [43–45]. Correlation analysis between COAD-associated neoantigen genes and tumor-infiltrating immune cells indicated that as neoantigen expression increases, the fractions of several immune cell types rise accordingly. Specifically, our data showed that some immune cells, such as dendritic cells, mast cells, and T cells were over-represented and synchronously activated in samples with high neoantigen expression. Meanwhile, other immune cells, such as B cells, monocytes, plasma cells, and eosinophils, were inhibited and downregulated. This suggests an obvious link between tumor neoantigen expression and differential representation and activity of tumor-associated immune cell populations. Although further research is warranted, this suggests the possibility of assessing neoantigen levels to estimate the immune status of tumors.

To gain insight on the regulatory landscape of the 24 DEGs harboring potential COAD neoantigens, we correlated their expression with that of miRNAs. Nine non-validated miRNAs were predicted to regulate these

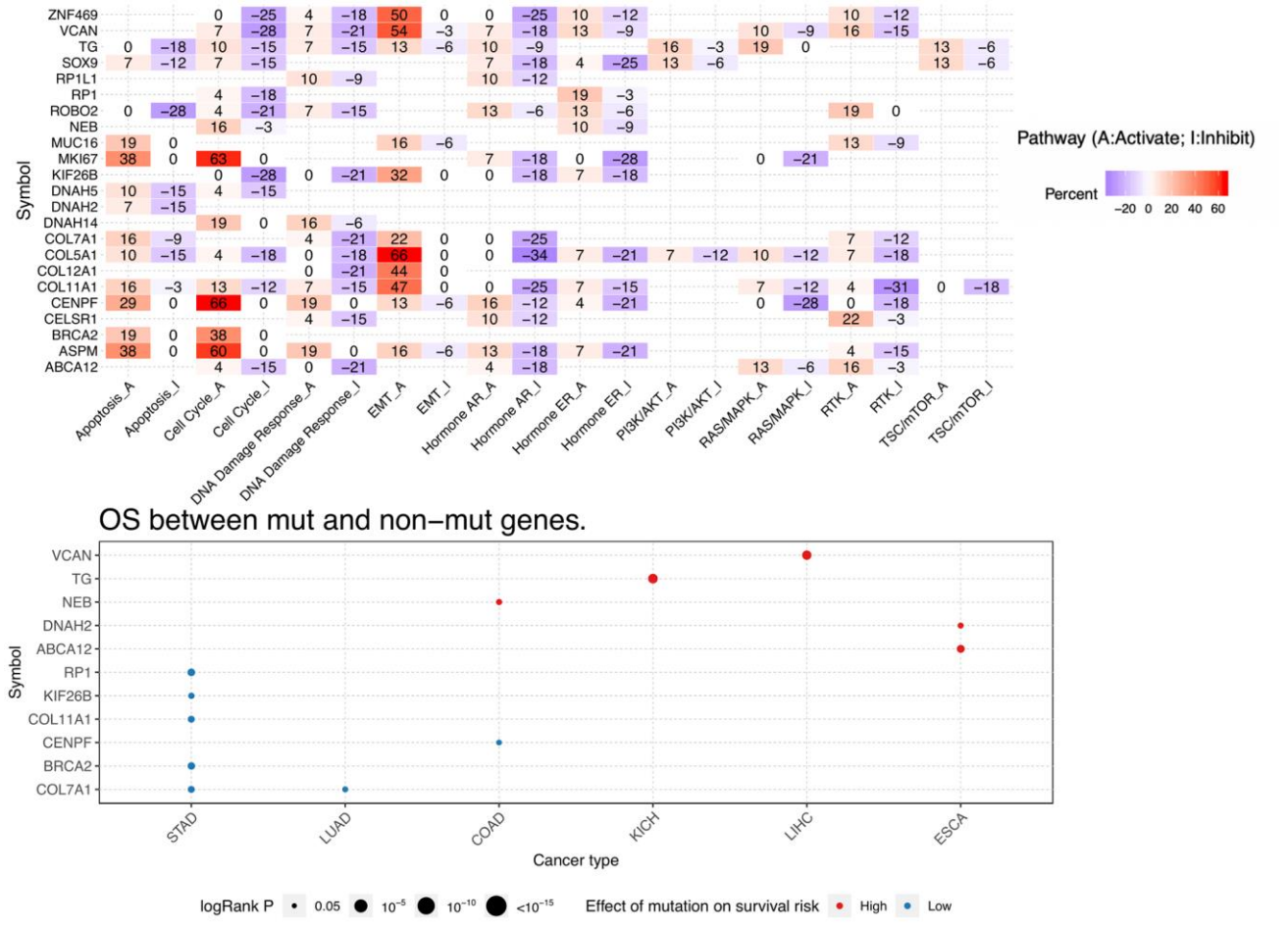


Figure 12. Pathway activity and mutation survival analysis. Top: Inferred activity of the identified host genes in biological pathways. Red and blue represent percent activation or inhibition. Bottom: Relationship between mutations in identified DEGs and survival prognosis for selected cancers.

24 neoantigens. The candidate neoantigen genes with the strongest negative correlation with these miRNAs were *ZNF469, COL5A1, COL12A1, KIF26B, COL11A1* and *VCAN*. Although experimental confirmation of these interactions is lacking, these miRNAs correlated with at least one of the genes more strongly than any of the validated miRNAs.

The recurrent mutations identified herein are predominantly observed in COAD, compared to other cancers. ESCA and STAD are the cancers most closely related to COAD in terms of recurrent mutation profiles. This is not unexpected, because these three cancers arise in the digestive system and share a similar tumor microenvironment [46]. However, the predicted impact of the identified mutations on the survival risk of patients was rather dissimilar among these entities. For instance, our analyses showed that mutations in *NEB* correlate with higher survival risk in COAD patients, but have no obvious impact on ESCA patients. Conversely, mutations in *DNAH2* and *ABCA12* were associated with higher risk in ESCA, but not COAD, patients. In turn, mutations in *RP1, KIF26B, COL11A1, BRCA2,* and *COL7A1* had no prognostic significance in COAD patients, but correlated with lower risk in STAD patients. Besides ESCA and STAD, we found that LUAD also shares a similar mutational profile as COAD. A possible reason for this is that LUAD patients tend to harbor more mutations due to exposures such as tobacco smoking [47].

In summary, our study integrated transcriptome and whole-exome sequencing data from COAD-TCGA and identified 24 DEGs harboring recurrent somatic mutations with neoantigen-forming potential in COAD patients. Among these candidate neoantigen genes, *NEB, DNAH2, ABCA12, CENPF,* and *CELSR1* were newly identified as COAD biomarkers, while *DNAH5, COL7A1, COL27A1, RP1L1,* and *ROBO2* had mutational profiles specific to COAD, compared to other solid tumors. We further constructed two diagnostic signatures, composed respectively of 4 early stage-related genes (*COL11A1, TG, SOX9,* and *DNAH2*) and 4 late stage-related genes (*COL11A1, SOX9, TG,* and *BRCA2*), which predicted COAD stage with high accuracy. Furthermore, several candidate neoantigen-yielding genes identified herein showed significant correlations with both miRNAs and diverse tumor-infiltrating immune cell types, and therefore represent promising therapeutic targets for immuno-therapy. Nevertheless, further research is warranted to experimentally validate the association between the recurrently mutated DEGS identified herein and the generation of tumor-specific neoantigens, and to explore the functional impact of the identified mutations on tumor biology.

MATERIALS AND METHODS

Data collection

We downloaded WES and RNAseq data of COAD from the TCGA database (https://www.cancer.gov/tcga). We also retrieved the clinical information for all patients, including MNT stage and survival data. A total of 459 COAD cases (329 with transcriptome and 399 with exome sequencing data, respectively) were thus obtained. We also retrieved the miRNA sequencing data for 261 samples, involving 2,113 microRNAs. Gene expression profiles for all patients were determined using the Illumina HiSeq 2000 RNA Sequencing platform. Level 3 data were downloaded from TCGA data coordination center. This dataset shows the gene-level transcription estimates as the log2(x+1) transformed RSEM-normalized count. Patients diagnosed with tumor stage I/II were assigned to the early stage group, and those with more advanced stages were assigned to the late stage group.

Differential gene expression analysis

We analyzed RNAseq data from 329 COAD patients, 41 of which had matching data for normal tissues. The limma algorithm [48] was used to identify differentially expressed genes (DEGs) in early and late stage COAD samples, compared with normal tissue specimens. As the microRNA data was only used to explore potential regulatory actions on protein-coding genes, differential microRNA expression was only assessed between the two tumor stages. All genes and microRNAs with $P < 0.05$ and logFC values over the 95% confidence limit were considered as differentially expressed.

Gene clustering and functional analysis

Overexpressed genes in either stage were selected to establish candidate neoantigen pools. These genes were either lowly expressed or silent in normal tissues, and therefore potentially good targets to avoid adverse effects if used to develop targeted therapies [49]. Hierarchical clustering [50] was used to visualize gene expression patterns in normal and tumor specimens. ClusterProfiler and enrichplot R packages [51] were used to conduct functional enrichment analyses.

Co-expression network analysis

DEGs and miRNAs identified at early and late tumor stages were used to construct the co-expression network [52]. Transcript (mRNA or microRNA) co-expression was determined by Pearson's correlation analysis [53], with a correlation coefficient cutoff

determined based on 95% CIs for all pairs. The network was constructed using Cytoscape 3.8.0 software [54]. Significant modules were mined from the network using the MCODE plugin with default parameters [55].

Recurrent somatic mutation selection

MAF files including somatic mutation information for exome sequencing data were retrieved from the TCGA database. We focused on nucleotide resolution and selected recurrent somatic mutations (i.e. those carried by at least 5% of patients), which represent potential therapeutic targets [56].

Selection of candidate neoantigen-associated genes

Candidate neoantigen-forming genes were initially selected based on high expression in COAD samples and low or no expression in normal ones [57]. Among those, we selected the early and late stage genes that harbored somatic mutations that were recurrent in at least 5% of COAD specimens. Putative neoantigen genes corresponding to each stage were compared to determine stage-specific differences with potential correlation with tumor progression.

Analysis of tumor-infiltrating immune cell populations

We applied CIBERSORT [58], a computational method for quantifying immune cell fractions from RNAseq data, to evaluate infiltration rates for 22 immune cell types. Pearson's correlation coefficients were subsequently computed to assess potential correlations between neoantigen genes and infiltrating immune cells. The correlation matrix was visualized by heatmap using the heatmap.2 R package.

Prediction of miRNAs targeting candidate neoantigen-associated transcripts

The co-expression network described above was used to extract all the miRNAs correlated with at least one of the selected host genes of the candidate neoantigens. The correlated miRNAs included both validated (retrieved from miRecords [59], miRTarBase [60], and TarBase [61] databases) as well as unannotated transcripts. Then, we inferred potential miRNAs targeting the candidate neoantigens by cross-assessment with the validated miRNAs.

Diagnostic model construction

To investigate whether the host genes encoding putative COAD neoantigens could serve as diagnostic signatures, we trained a random forest model as a diagnostic predictor [62]. The host genes corresponding to each stage were used as signatures. The whole data was randomly split into discovery (70%) and validation (30%) samples. For feature selection, we first randomly split the discovery data into train and test sets. A linear SVC model was used to select the most significant features [63]. This process was repeated 100 times, and the features selected in at least 80 iterations were considered as robust signatures. The predictor was trained with default parameters using the training set, and 10-fold cross-validation was used to evaluate the performance of the predictor [64]. Eventually, we applied the predictor to the test set and assessed the model's accuracy.

Survival analysis

The impact of candidate host genes harboring neoantigen-related mutations on COAD prognosis was initially assessed using stepwise regression [65]. Then, patients were separated into low- and high-risk cohorts based on individual expression frequencies. Survival analysis was conducted using Kaplan-Meier curves generated through the survival and ggsurvplot R packages [66, 67].

Comparative neoantigen expression analysis

To investigate whether the neoantigen-related DEGs identified herein are COAD-specific or are also shared by other types of cancers, we used the GSCALite tool [68] to compare the corresponding expression patterns in datasets from ten additional cancers retrieved from the TCGA database.

Abbreviations

COAD: colon adenocarcinoma; TSA: tumor-specific antigens; DEGs: differentially expressed genes; PCA: principal component analysis; GO: Gene Ontology; WGS: whole-genome sequencing; WES: whole-exome sequencing; STAD: Stomach adenocarcinoma; ESCA: Esophageal carcinoma; LUAD: Lung adenocarcinoma; HNSC: Head and Neck squamous cell carcinoma; OV: Ovarian serous cystadenocarcinoma; BRCA: Breast invasive carcinoma; PAAD: Pancreatic adenocarcinoma; KICH: Kidney Chromophobe; PRAD: Prostate adenocarcinoma; THCA: Thyroid carcinoma.

AUTHOR CONTRIBUTIONS

C.W. and W.X. contributed to the conception of the study, performed data analysis and wrote the manuscript. H.Z. and Y.F. collaborated with data analysis and interpretation.

ACKNOWLEDGMENTS

The results published here are based upon data generated by the TCGA.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China (U1804191, 81800137, 81871995, U1904137) and Key Scientific Research Projects of Colleges and Universities of Henan Provincial Department of Education (20A320061).

REFERENCES

1. Jemal A, Siegel R, Ward E, Murray T, Xu J, Thun MJ. Cancer statistics, 2007. CA Cancer J Clin. 2007; 57:43–66. https://doi.org/10.3322/canjclin.57.1.43 PMID:17237035

2. Hamilton SR, Aaltonen LA. World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of the Digestive System. International Agency for Research on Cancer (IARC). 2000.

3. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018; 68:394–424. https://doi.org/10.3322/caac.21492 PMID:30207593

4. Dean TM. Carcinoma of the colon and rectum. A perspective for practicing physicians, with recommendations for screening. West J Med. 1977; 126:431–40. PMID:878459

5. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. Science. 2015; 348:69–74. https://doi.org/10.1126/science.aaa4971 PMID:25838375

6. Craig DW, O'Shaughnessy JA, Kiefer JA, Aldrich J, Sinari S, Moses TM, Wong S, Dinh J, Christoforides A, Blum JL, Aitelli CL, Osborne CR, Izatt T, et al. Genome and transcriptome sequencing in prospective metastatic triple-negative breast cancer uncovers therapeutic vulnerabilities. Mol Cancer Ther. 2013; 12:104–16. https://doi.org/10.1158/1535-7163.MCT-12-0781 PMID:23171949

7. Karasaki T, Nagayama K, Kuwano H, Nitadori JI, Sato M, Anraku M, Hosoi A, Matsushita H, Takazawa M, Ohara O, Nakajima J, Kakimi K. Prediction and prioritization of neoantigens: integration of RNA sequencing data with whole-exome sequencing. Cancer Sci. 2017; 108:170–77. https://doi.org/10.1111/cas.13131 PMID:27960040

8. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018; 28:1747–56. https://doi.org/10.1101/gr.239244.118 PMID:30341162

9. Modrek B, Ge L, Pandita A, Lin E, Mohan S, Yue P, Guerrero S, Lin WM, Pham T, Modrusan Z, Seshagiri S, Stern HM, Waring P, et al. Oncogenic activating mutations are associated with local copy gain. Mol Cancer Res. 2009; 7:1244–52. https://doi.org/10.1158/1541-7786.MCR-08-0532 PMID:19671679

10. Grandér D. How do mutated oncogenes and tumor suppressor genes cause cancer? Med Oncol. 1998; 15:20–26. https://doi.org/10.1007/BF02787340 PMID:9643526

11. Maere S, Heymans K, Kuiper M. BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics. 2005; 21:3448–49. https://doi.org/10.1093/bioinformatics/bti551 PMID:15972284

12. Kreutzer DA, Essigmann JM. Oxidized, deaminated cytosines are a source of C → T transitions *in vivo*. Proc Natl Acad Sci USA. 1998; 95:3578–82. https://doi.org/10.1073/pnas.95.7.3578 PMID:9520408

13. Kalyan A, Kircher S, Shah H, Mulcahy M, Benson A. Updates on immunotherapy for colorectal cancer. J Gastrointest Oncol. 2018; 9:160–69. https://doi.org/10.21037/jgo.2018.01.17 PMID:29564182

14. Jiang T, Shi T, Zhang H, Hu J, Song Y, Wei J, Ren S, Zhou C. Tumor neoantigens: from basic research to clinical applications. J Hematol Oncol. 2019; 12:93. https://doi.org/10.1186/s13045-019-0787-5 PMID:31492199

15. Raglow Z, Thomas SM. Tumor matrix protein collagen XIα1 in cancer. Cancer Lett. 2015; 357:448–53. https://doi.org/10.1016/j.canlet.2014.12.011 PMID:25511741

16. Fischer H, Stenling R, Rubio C, Lindblom A. Colorectal carcinogenesis is associated with stromal expression

of COL11A1 and COL5A2. Carcinogenesis. 2001; 22:875–78.
https://doi.org/10.1093/carcin/22.6.875
PMID:11375892

17. Feng Y, Sun B, Li X, Zhang L, Niu Y, Xiao C, Ning L, Fang Z, Wang Y, Zhang L, Cheng J, Zhang W, Hao X. Differentially expressed genes between primary cancer and paired lymph node metastases predict clinical outcome of node-positive breast cancer patients. Breast Cancer Res Treat. 2007; 103:319–29.
https://doi.org/10.1007/s10549-006-9385-7
PMID:17123152

18. Sok JC, Lee JA, Dasari S, Joyce S, Contrucci SC, Egloff AM, Trevelline BK, Joshi R, Kumari N, Grandis JR, Thomas SM. Collagen type XI α1 facilitates head and neck squamous cell cancer growth and invasion. Br J Cancer. 2013; 109:3049–56.
https://doi.org/10.1038/bjc.2013.624
PMID:24231953

19. García-Pravia C, Galván JA, Gutiérrez-Corral N, Solar-García L, García-Pérez E, García-Ocaña M, Del Amo-Iribarren J, Menéndez-Rodríguez P, García-García J, de Los Toyos JR, Simón-Buela L, Barneo L. Overexpression of COL11A1 by cancer-associated fibroblasts: clinical relevance of a stromal marker in pancreatic cancer. PLoS One. 2013; 8:e78327.
https://doi.org/10.1371/journal.pone.0078327
PMID:24194920

20. Fischer H, Salahshor S, Stenling R, Björk J, Lindmark G, Iselius L, Rubio C, Lindblom A. COL11A1 in FAP polyps and in sporadic colorectal tumors. BMC Cancer. 2001; 1:17.
https://doi.org/10.1186/1471-2407-1-17
PMID:11707154

21. Wu YH, Huang YF, Chang TH, Chou CY. Activation of TWIST1 by COL11A1 promotes chemoresistance and inhibits apoptosis in ovarian cancer cells by modulating NF-κB-mediated IKKβ expression. Int J Cancer. 2017; 141:2305–17.
https://doi.org/10.1002/ijc.30932
PMID:28815582

22. Lü B, Fang Y, Xu J, Wang L, Xu F, Xu E, Huang Q, Lai M. Analysis of SOX9 expression in colorectal cancer. Am J Clin Pathol. 2008; 130:897–904.
https://doi.org/10.1309/AJCPW1W8GJBQGCNI
PMID:19019766

23. Matheu A, Collado M, Wise C, Manterola L, Cekaite L, Tye AJ, Canamero M, Bujanda L, Schedl A, Cheah KS, Skotheim RI, Lothe RA, López de Munain A, et al. Oncogenicity of the developmental transcription factor Sox9. Cancer Res. 2012; 72:1301–15.
https://doi.org/10.1158/0008-5472.CAN-11-3660
PMID:22246670

24. Mazur PK, Riener MO, Jochum W, Kristiansen G, Weber A, Schmid RM, Siveke JT. Expression and clinicopathological significance of notch signaling and cell-fate genes in biliary tract cancer. Am J Gastroenterol. 2012; 107:126–35.
https://doi.org/10.1038/ajg.2011.305 PMID:21931375

25. Bastide P, Darido C, Pannequin J, Kist R, Robine S, Marty-Double C, Bibeau F, Scherer G, Joubert D, Hollande F, Blache P, Jay P. Sox9 regulates cell proliferation and is required for paneth cell differentiation in the intestinal epithelium. J Cell Biol. 2007; 178:635–48.
https://doi.org/10.1083/jcb.200704152
PMID:17698607

26. Corpet DE, Yin Y, Zhang XM, Rémésy C, Stamp D, Medline A, Thompson L, Bruce WR, Archer MC. Colonic protein fermentation and promotion of colon carcinogenesis by thermolyzed casein. Nutr Cancer. 1995; 23:271–81.
https://doi.org/10.1080/01635589509514381
PMID:7603887

27. Wan Y, Xin Y, Zhang C, Wu D, Ding D, Tang L, Owusu L, Bai J, Li W. Fermentation supernatants of Lactobacillus delbrueckii inhibit growth of human colon cancer cells and induce apoptosis through a caspase 3-dependent pathway. Oncol Lett. 2014; 7:1738–42.
https://doi.org/10.3892/ol.2014.1959
PMID:24765211

28. Thangaraju M, Cresci GA, Liu K, Ananth S, Gnanaprakasam JP, Browning DD, Mellinger JD, Smith SB, Digby GJ, Lambert NA, Prasad PD, Ganapathy V. GPR109A is a G-protein-coupled receptor for the bacterial fermentation product butyrate and functions as a tumor suppressor in colon. Cancer Res. 2009; 69:2826–32.
https://doi.org/10.1158/0008-5472.CAN-08-4466
PMID:19276343

29. Arun KB, Madhavan A, Reshmitha TR, Thomas S, Nisha P. Short chain fatty acids enriched fermentation metabolites of soluble dietary fibre from musa paradisiaca drives HT29 colon cancer cells to apoptosis. PLoS One. 2019; 14:e0216604.
https://doi.org/10.1371/journal.pone.0216604
PMID:31095579

30. Pompili L, Maresca C, Dello Stritto A, Biroccio A, Salvati E. BRCA2 deletion induces alternative lengthening of telomeres in telomerase positive colon cancer cells. Genes (Basel). 2019; 10:697.
https://doi.org/10.3390/genes10090697
PMID:31510074

31. Zhang X, Zhang H, Shen B, Sun XF. Chromogranin-a expression as a novel biomarker for early diagnosis of colon cancer patients. Int J Mol Sci. 2019; 20:2919.

https://doi.org/10.3390/ijms20122919
PMID:31207989

32. Björkman K, Mustonen H, Kaprio T, Haglund C, Böckelman C. Mucin 16 and kallikrein 13 as potential prognostic factors in colon cancer: results of an oncological 92-multiplex immunoassay. Tumour Biol. 2019; 41:1010428319860728.
https://doi.org/10.1177/1010428319860728
PMID:31264534

33. Ruan W, Zhu S, Wang H, Xu F, Deng H, Ma Y, Lai M. IGFBP-rP1, a potential molecule associated with colon cancer differentiation. Mol Cancer. 2010; 9:281.
https://doi.org/10.1186/1476-4598-9-281
PMID:20977730

34. Long X, Deng Z, Li G, Wang Z. Identification of critical genes to predict recurrence and death in colon cancer: integrating gene expression and bioinformatics analysis. Cancer Cell Int. 2018; 18:139.
https://doi.org/10.1186/s12935-018-0640-x
PMID:30237752

35. Fletcher JI, Haber M, Henderson MJ, Norris MD. ABC transporters in cancer: more than just drug efflux pumps. Nat Rev Cancer. 2010; 10:147–56.
https://doi.org/10.1038/nrc2789 PMID:20075923

36. Whitfield M, Thomas L, Bequignon E, Schmitt A, Stouvenel L, Montantin G, Tissier S, Duquesnoy P, Copin B, Chantot S, Dastot F, Faucon C, Barbotin AL, et al. Mutations in DNAH17, encoding a sperm-specific axonemal outer dynein arm heavy chain, cause isolated male infertility due to asthenozoospermia. Am J Hum Genet. 2019; 105:198–212.
https://doi.org/10.1016/j.ajhg.2019.04.015
PMID:31178125

37. Li L, Jiang X, Zhang Q, Dong X, Gao Y, He Y, Qiao H, Xie F, Xie X, Sun X. Neuropilin-1 is associated with clinicopathology of gastric cancer and contributes to cell proliferation and migration as multifunctional co-receptors. J Exp Clin Cancer Res. 2016; 35:16.
https://doi.org/10.1186/s13046-016-0291-5
PMID:26795388

38. Aytes A, Mitrofanova A, Lefebvre C, Alvarez MJ, Castillo-Martin M, Zheng T, Eastham JA, Gopalan A, Pienta KJ, Shen MM, Califano A, Abate-Shen C. Cross-species regulatory network analysis identifies a synergistic interaction between FOXM1 and CENPF that drives prostate cancer Malignancy. Cancer Cell. 2014; 25:638–51.
https://doi.org/10.1016/j.ccr.2014.03.017
PMID:24823640

39. Kang Y, Siegel PM, Shu W, Drobnjak M, Kakonen SM, Cordón-Cardo C, Guise TA, Massagué J. A multigenic program mediating breast cancer metastasis to bone. Cancer Cell. 2003; 3:537–49.
https://doi.org/10.1016/s1535-6108(03)00132-6
PMID:12842083

40. Du E, Lu C, Sheng F, Li C, Li H, Ding N, Chen Y, Zhang T, Yang K, Xu Y. Analysis of potential genes associated with primary cilia in bladder cancer. Cancer Manag Res. 2018; 10:3047–56.
https://doi.org/10.2147/CMAR.S175419
PMID:30214299

41. Shahid M, Lee MY, Piplani H, Andres AM, Zhou B, Yeon A, Kim M, Kim HL, Kim J. Centromere protein F (CENPF), a microtubule binding protein, modulates cancer metabolism by regulating pyruvate kinase M2 phosphorylation signaling. Cell Cycle. 2018; 17:2802–18.
https://doi.org/10.1080/15384101.2018.1557496
PMID:30526248

42. Zhang S, Cheng J, Quan C, Wen H, Feng Z, Hu Q, Zhu J, Huang Y, Wu X. circCELSR1 (hsa_circ_0063809) contributes to paclitaxel resistance of ovarian cancer cells by regulating FOXR2 expression via miR-1252. Mol Ther Nucleic Acids. 2020; 19:718–30.
https://doi.org/10.1016/j.omtn.2019.12.005
PMID:31945729

43. Tumeh PC, Harview CL, Yearley JH, Shintaku IP, Taylor EJ, Robert L, Chmielowski B, Spasic M, Henry G, Ciobanu V, West AN, Carmona M, Kivork C, et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. Nature. 2014; 515:568–71.
https://doi.org/10.1038/nature13954
PMID:25428505

44. Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, Hackl H, Trajanoski Z. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. Cell Rep. 2017; 18:248–62.
https://doi.org/10.1016/j.celrep.2016.12.019
PMID:28052254

45. Van Allen EM, Miao D, Schilling B, Shukla SA, Blank C, Zimmer L, Sucker A, Hillen U, Foppen MHG, Goldinger SM, Utikal J, Hassel JC, Weide B, et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science. 2015; 350:207–211.
https://doi.org/10.1126/science.aad0095
PMID:26359337

46. Percesepe A, Ponz De Leon M. [Hereditary factors in tumors of the digestive system]. Ann Ist Super Sanita. 1996; 32:629–42.
PMID:9382432

47. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012; 489:519–25. https://doi.org/10.1038/nature11404 PMID:22960745

48. Smyth GK, Ritchie M, Thorne N, Wettenhall J, Shi W, Hu Y. Linear Models for Microarray and RNA-Seq Data User's Guide. Walter Eliza Hall Inst Med Res Aust. 2018.

49. Gold P, Freedman SO. Specific carcinoembryonic antigens of the human digestive system. J Exp Med. 1965; 122:467–81. https://doi.org/10.1084/jem.122.3.467 PMID:4953873

50. Kimes PK, Liu Y, Neil Hayes D, Marron JS. Statistical significance for hierarchical clustering. Biometrics. 2017; 73:811–21. https://doi.org/10.1111/biom.12647 PMID:28099990

51. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012; 16:284–87. https://doi.org/10.1089/omi.2011.0118 PMID:22455463

52. Chen PF, Wang F, Nie JY, Feng JR, Liu J, Zhou R, Wang HL, Zhao Q. Co-expression network analysis identified CDH11 in association with progression and prognosis in gastric cancer. Onco Targets Ther. 2018; 11:6425–36. https://doi.org/10.2147/OTT.S176511 PMID:30323620

53. Månsson R, Tsapogas P, Akerlund M, Lagergren A, Gisler R, Sigvardsson M. Pearson correlation analysis of microarray data allows for the identification of genetic targets for early B-cell factor. J Biol Chem. 2004; 279:17905–13. https://doi.org/10.1074/jbc.M400589200 PMID:14960572

54. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13:2498–504. https://doi.org/10.1101/gr.1239303 PMID:14597658

55. Zaki N, Efimov D, Berengueres J. Protein complex detection using interaction reliability assessment and weighted clustering coefficient. BMC Bioinformatics. 2013; 14:163. https://doi.org/10.1186/1471-2105-14-163 PMID:23688127

56. Sun Z, Wang L, Eckloff BW, Deng B, Wang Y, Wampfler JA, Jang J, Wieben ED, Jen J, You M, Yang P. Conserved recurrent gene mutations correlate with pathway deregulation and clinical outcomes of lung adenocarcinoma in never-smokers. BMC Med Genomics. 2014; 7:32. https://doi.org/10.1186/1755-8794-7-32 PMID:24894543

57. Boegel S, Castle JC, Kodysh J, O'Donnell T, Rubinsteyn A. Bioinformatic methods for cancer neoantigen prediction. Prog Mol Biol Transl Sci. 2019; 164:25–60. https://doi.org/10.1016/bs.pmbts.2019.06.016 PMID:31383407

58. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling tumor infiltrating immune cells with CIBERSORT. Methods Mol Biol. 2018; 1711:243–59. https://doi.org/10.1007/978-1-4939-7493-1_12 PMID:29344893

59. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res. 2009; 37:D105–10. https://doi.org/10.1093/nar/gkn851 PMID:18996891

60. Chou CH, Shrestha S, Yang CD, Chang NW, Lin YL, Liao KW, Huang WC, Sun TH, Tu SJ, Lee WH, Chiew MY, Tai CS, Wei TY, et al. miRTarBase update 2018: A resource for experimentally validated microRNA-target interactions. Nucleic Acids Res. 2018; 46:D296–302. https://doi.org/10.1093/nar/gkx1067 PMID:29126174

61. Vergoulis T, Vlachos IS, Alexiou P, Georgakilas G, Maragkakis M, Reczko M, Gerangelos S, Koziris N, Dalamagas T, Hatzigeorgiou AG. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. Nucleic Acids Res. 2012; 40:D222–29. https://doi.org/10.1093/nar/gkr1161 PMID:22135297

62. Athey S, Tibshirani J, Wager S. Generalized random forests. Ann Stat. 2019; 47:1148–1178. https://doi.org/10.1214/18-AOS1709

63. Pandis N. Multiple linear regression analysis. Am J Orthod Dentofacial Orthop. 2016; 149:581. https://doi.org/10.1016/j.ajodo.2016.01.012 PMID:27021463

64. Gianola D, Schön CC. Cross-validation without doing cross-validation in genome-enabled prediction. G3 (Bethesda). 2016; 6:3107–28. https://doi.org/10.1534/g3.116.033381 PMID:27489209

65. Burkholder TJ, Lieber RL. Stepwise regression is an alternative to splines for fitting noisy data. J Biomech. 1996; 29:235–38. https://doi.org/10.1016/0021-9290(95)00044-5 PMID:8849817

66. Chiba Y. Kaplan-meier curves for survivor causal effects with time-to-event outcomes. Clin Trials. 2013; 10:515–21. https://doi.org/10.1177/1740774513483601 PMID:23610455

67. Zhou M, Zhao H, Wang Z, Cheng L, Yang L, Shi H, Yang H, Sun J. Identification and validation of potential prognostic lncRNA biomarkers for predicting survival in patients with multiple myeloma. J Exp Clin Cancer Res. 2015; 34:102. https://doi.org/10.1186/s13046-015-0219-5 PMID:26362431

68. Liu CJ, Hu FF, Xia MX, Han L, Zhang Q, Guo AY. GSCALite: a web server for gene set cancer analysis. Bioinformatics. 2018; 34:3771–72. https://doi.org/10.1093/bioinformatics/bty411 PMID:29790900

SUPPLEMENTARY MATERIALS

Supplementary Table

Please browse Full Text version to see the data of Supplementary Tables 1 and 2.

Supplementary Table 1. Differentially expressed miRNAs and target genes.

Supplementary Table 2. Variant information of recurrent somatic mutations.