Molecular Therapy
Nucleic Acids
Original Article

Multi-omics Data Analyses Construct TME and Identify the Immune-Related Prognosis Signatures in Human LUAD

Yuwei Zhang,[1,2,3] Minglei Yang,[1,3] Derry Minyao Ng,[2] Maria Haleem,[2] Tianfei Yi,[2] Shiyun Hu,[2] Huangkai Zhu,[1,3] Guofang Zhao,[1,3] and Qi Liao[1,2,3]

[1]Hwa Mei Hospital, University of Chinese Academy of Science, Ningbo, Zhejiang, China; [2]Department of Preventative Medicine, Zhejiang Provincial Key Laboratory of Pathophysiology Technology, Medical School of Ningbo University, Ningbo, China; [3]Ningbo Institute of Life and Health Industry, University of Chinese Academy of Sciences

Lung cancer has been the focus of attention for many researchers in recent years for the leading contribution to cancer-related death worldwide, in which lung adenocarcinoma (LUAD) is the most common histological type. However, the potential mechanism behind LUAD initiation and progression remains unclear. Aiming to dissect the tumor microenvironment of LUAD and to discover more informative prognosis signatures, we investigated the immune-related differences in three types of genetic or epigenetic characteristics (expression status, somatic mutation, and DNA methylation) and considered the potential roles that these alterations have in the immune response and both the immune-related metabolic and neural systems by analyzing the multi-omics data from The Cancer Genome Atlas (TCGA) portal. Additionally, a four-step strategy based on lasso regression and Cox regression was used to construct the prognostic prediction model. For the prognostic predictions on the independent test set, the performance of the trained models (average concordance index [C-index] = 0.839) is satisfied, with average 1-year, 3-year, and 5-year areas under the curve (AUCs) equal to 0.796, 0.786, and 0.777. Finally, the overall model was constructed based on all samples, which comprised 27 variables and achieved a high degree of accuracy on the 1-year (AUC = 0.861), 3-year (AUC = 0.850), and 5-year (AUC = 0.916) survival predictions.

INTRODUCTION

Lung cancer represents the peaks of both incidence and mortality among all cancer types, with 2,093,876 (11.6%) new worldwide cases and 1,761,007 (18.4%) deaths out of all cancer types.[1] Lung adenocarcinoma (LUAD) is the most common histological type of lung cancer. Many epidemiological investigations and experimental studies have attributed the occurrence and progression of LUAD largely to both environmental factors and genetic alterations.[2–4] Despite the increasing prevalence of health education and the decrease in smoking rates, the incidence trends have not changed considerably. A large number of never-smokers suffering from LUAD overturned the previous theory based only on environmental factors and refocused the researchers' attention on the in-depth alterations of genetic content. Until now, there are two major types of genetically related therapeutic strategies, i.e., targeted therapy and immunotherapy. Targeted therapy requires a specific gene mutation in cancer patients, and the epidermal growth factor receptor (*EGFR*) is the most representative mutation in LUAD patients, especially in Asians (51.4%), and can therefore act as a sensitive therapeutic target for the tyrosine kinase inhibitor (TKI).[5,6] Unfortunately, there remains a large proportion of LUAD patients trapped in the lack of targetable mutations. Current progress on immunotherapy paves an efficient and safe way to bail them out. Immunotherapy is always realized by the introductions of immune checkpoint blockers (ICBs) such as antibodies targeting programmed death 1 (*PD-1*) and cytotoxic T lymphocyte antigen-4 (*CTLA-4*), but this method is also largely affected by the tumor microenvironment (TME) such as T cell abundance, tumor mutation burden, and the regulatory relationships with other biomolecules,[6–8] which highlights the importance and the urgent need of deeply understanding how TME orchestrates the therapeutic and prognostic outcomes.

TME is commonly defined as the environment around a tumor, which comprises the extracellular matrix, blood vessels, and cellular players such as immune cells and neurons, all of which have a strong association with tumor progression and therapeutic outcome.[9] Increasing studies have experimentally elucidated the contributive roles of TME infiltrates in the immunotherapy response and resistance in various cancer types, and explored their influence on patient prognosis.[10–12] For example, B7-CD28 family-based signatures were found to have significant prognostic values and demonstrated that the tumor immune landscape plays a role in LUAD.[13] Additionally,

Received 14 March 2020; accepted 20 July 2020;
https://doi.org/10.1016/j.omtn.2020.07.024.

Correspondence: Qi Liao, HwaMei Hospital, University of Chinese Academy of Science, Ningbo, Zhejiang, China.
E-mail: liaoqi@nbu.edu.cn
Correspondence: Guofang Zhao, HwaMei Hospital, University of Chinese Academy of Science, Ningbo, Zhejiang, China.
E-mail: guofzhao@hotmail.com

 © 2020 The Author(s).
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

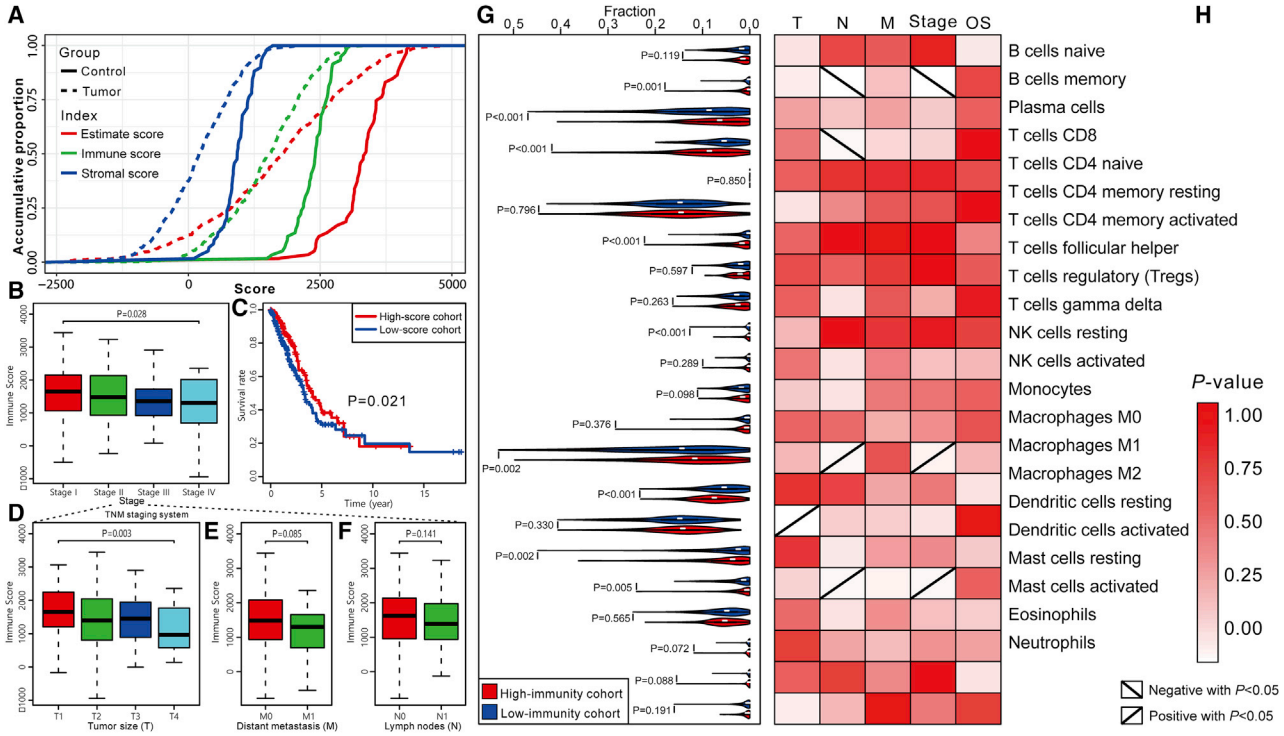


Figure 1. Construction of TME in LUAD

(A) Comparison of the distributions of estimate scores, immune scores, and stromal scores between tumor and control cohorts. (B) Comparison of immune scores on the TNM stage. (C) Kaplan-Meier curves show the independent relevance between overall survival time and immune scores. (D–F) Comparisons of immune scores on (D) tumor size, (E) distance metastasis, and (F) lymph nodes. (G) Comparisons of the immune cell members between high-immunity and low-immunity groups. (H) Relevance between clinical factors and immune cell members. The red to white gradient represents the significance level. The positive and negative diagonals in the boxes respectively indicate the significantly positive and negative relationships between cell fractions and the level of clinical factors ($p < 0.05$). (G) and (H) share the same y axis.

Song et al.[14] established a model comprising of 30 immune-related genes as predictive signatures and attempted to determine their relationship with clinicopathological factors. However, most of the proposed models for LUAD prognosis prediction[14–16] only incorporate expression changes and are therefore not comprehensive enough to reach a satisfying performance.

In the present study, we aim to estimate the TME infiltration patterns, especially the tumor-associated immune system in LUAD based on the expression profiles from The Cancer Genome Atlas (TCGA) portal, and then correlate the immune status with genetic or epigenetic characteristics through analyzing the multi-omics data (RNA sequencing [RNA-seq], whole-exome sequencing [WES], and DNA methylation array), and finally establish a prognosis prediction model from the significant alterations. It is expected that the results from this study could provide a more comprehensive map of the immunogenomic landscape and potentially find a better prognosis predictor for human LUAD.

RESULTS

Construction of the TME in LUAD

Tumor tissues are not simply composed of only tumor cells but also the heterogeneous microenvironment constituents such as fibroblasts, blood vessels, immune cells, stromal cells, and so on, which can be infiltrated by tumor cells and thus equipped with tumor-associated influence. The interactions between the tumor cells and the surrounding infiltrate, especially two major non-tumor constituents (stromal cells and immune cells), are able to orchestrate either tumor progression or inhibition.[17]

To evaluate the tumor-associated effects of the infiltrating stromal and immune cells, the TME was preliminarily established based on the expression profiles from TCGA portal by using the ESTIMATE (estimation of stromal and immune cells in malignant tumor tissues using expression data) algorithm.[18] ESTIMATE generates a stromal score that measures the presence of tumor-associated stroma and an immune score that represents the infiltration level of the immune cells and combines them to produce an index termed "estimate score" that comprehensively infers tumor purity. As shown in Figure 1A, the estimate scores of the LUAD samples were distributed at the significantly lower side, compared with those of the normal samples (Mann-Whitney U test, $p < 0.05$), as were the stromal scores and immune scores (Mann-Whitney U test, $p < 0.05$). The difference in the distributions between the tumor and normal cohorts supports the higher tumor purity in the selected LUAD samples.

Next, we investigated the relationship between tumor purity and clinical factors. As shown in the results (Figure S1A), the estimate scores

were significantly different in tumor size (Kruskal-Wallis test, $p < 0.05$), distant metastasis (Mann-Whitney U test, $p < 0.05$), and tumor stage (Kruskal-Wallis test, $p < 0.05$). Pairwise comparisons (Table S1) show that the samples in stage I have significantly higher estimate scores than do the samples in stages III and IV (Mann-Whitney U test, false discover rate [FDR] < 0.05). Furthermore, patients in the cohort with low estimate scores (below the median) have a poorer prognosis compared with those in the high-score (above the median) cohort (Figure S1A). To separately elaborate the roles of the stromal and immune cells, the relationships between the two types of scores and clinical factors were further studied. From the aspect of the TNM (tumor size, lymph node, and distant metastasis) staging system, the immune scores were significantly different in different tumor sizes (Kruskal-Wallis test, $p < 0.05$), but there was no difference in lymph nodes and distant metastasis (Figures 1D–1F). As for the integrative TNM stage classification, the immune scores were significantly different between the early stage and advanced stage as shown in Figure 1B (Kruskal-Wallis test, $p < 0.05$), of which those in stage I were significantly higher than those in stages III and IV (Mann-Whitney U test, FDR < 0.05) (Table S1). In contrast, stromal scores were only significantly related with distant metastasis (Mann-Whitney U test, $p < 0.05$) rather than tumor size, lymph nodes, and stage (Figure S1B), which conforms to the favorable functions of tumor-associated stroma in tumor metastasis.[19,20] A more important finding is that higher immune scores were significantly associated with longer overall survival time (Figure 1C), whereas stromal scores have no significant relevance with patient prognosis (Figure S1B). Extrapolating from all of the above results, we speculated that tumor-infiltrating immune cells have a stronger evidential clinical relevance compared to stromal cells in LUAD. Therefore, the following analyses mainly focused on the tumor immune microenvironment (TIME) and immunity-related genes.

Dissecting the Infiltrating Immune Content in LUAD

Since immune infiltration level and cell composition are strongly related to tumor progression and patient outcome,[21] we divided the LUAD samples into a high-immunity cohort and a low-immunity cohort using their median immune score, and further characterized cell composition using CIBERSORT[22] to explore the relationship between the immune cell subsets and clinical features. The samples with CIBERSORT[22]-generated p values greater than 0.05 were removed. A total of 468 samples were reserved, of which 203 belong to the low-immunity cohort and 265 belong to the high-immunity cohort.

Then, the immune content of each sample was dissected into 22 types of immune cell members. The high-immunity cohort had significantly larger fractions of memory B cells, CD8 T cells, activated memory CD4 T cells, M1 macrophages, resting dendritic cells, activated mast cells, and gamma delta T cells and smaller fractions of plasma cells, M0 macrophages, and activated dendritic cells (Mann-Whitney U test, $p < 0.05$) (Figure 1G), which is consistent with the proposed anti-immune or pro-immune functions of these cell members.[23–28] Of note, some of these immune cell members have potential correlations, such as the positive correlation of activated dendritic cells on CD8 T cell infiltration in lung cancer.[29] While considering the relationships between the immune score and the clinical factors detected above, we speculated that different immune cell members may have different contributions to patient outcome, and, as we expected, memory B cells, CD8 T cells, M0 macrophages, M2 macrophages, and activated dendritic cells were detected to have significant differences in TNM stage, tumor size, or lymph nodes (Mann-Whitney U test or Kruskal-Wallis test, $p < 0.05$), while the other cell subsets appear to be statistically insignificant in all clinical factors (Figure 1H). Furthermore, the single type of cell members contributed little to patient overall survival in LUAD (Figure 1H), although the combination shows discriminative power (Figure 1C).

Identification of the Immune Infiltration-Dependent Differentially Expressed Genes

The expression profiles of the LUAD samples from TCGA portal were used to identify the expression change between the high-immunity and low-immunity cohorts. The genes meeting the standard of $\log_2$(fold change) >1 and FDR <0.05 (Student t test) were considered to be differentially expressed, of which 611 and 164 genes were, respectively, upregulated and downregulated in the high-immunity cohort (Figure 2A). Note that 29 chemotactic factors such as *CXCR4* and *CCL8* were found to be significantly upregulated (Figure 2B), which are capable of modulating the recruitment of a wide variety of immune cells to tumor,[30]

Then, functional enrichment analyses using clusterProfiler[31] were performed to infer the potential functions of these two parts of genes. As we expected, the upregulated genes were found to be enriched in immune-related biological processes such as T cell activation and leukocyte proliferation (Figure 2C), indicating that they have a positive role in the enhancement of tumor-associated immunity. Furthermore, some of the upregulated genes such as *LILRB4*,[32] *RUNX3*,[33,34] and *CXCR3*[35] have been experimentally verified in prior studies to regulate T cell activation and support tumor infiltration. Alternatively, the downregulated genes were mainly enriched in the metabolic process (Figure 2D). The metabolic communication between the tumor cells and infiltrating immune cells is a battleground for energy competition and can therefore influence tumor progression.[36] Hence, we speculated that some downregulated genes modulate the activities of immune and tumor cells using the metabolic switch; for example, the role of AKR1C1/2/3 in the tumor infiltration level of immune cells and the metabolic process has been partially verified.[37]

Comparisons of Somatic Mutations under Different Immune Infiltration Levels

After detecting the transcriptional alterations in the above section, we further investigated whether there is evidence of the disparity in the genomic layer between the high-immunity and low-immunity cohorts. Somatic mutations including the single-nucleotide variant (SNV), single-nucleotide polymorphism (SNP), insertion (INS), and deletion (DEL) were analyzed and visualized using the R package maftools[38] based on the WES data from TCGA portal in which the

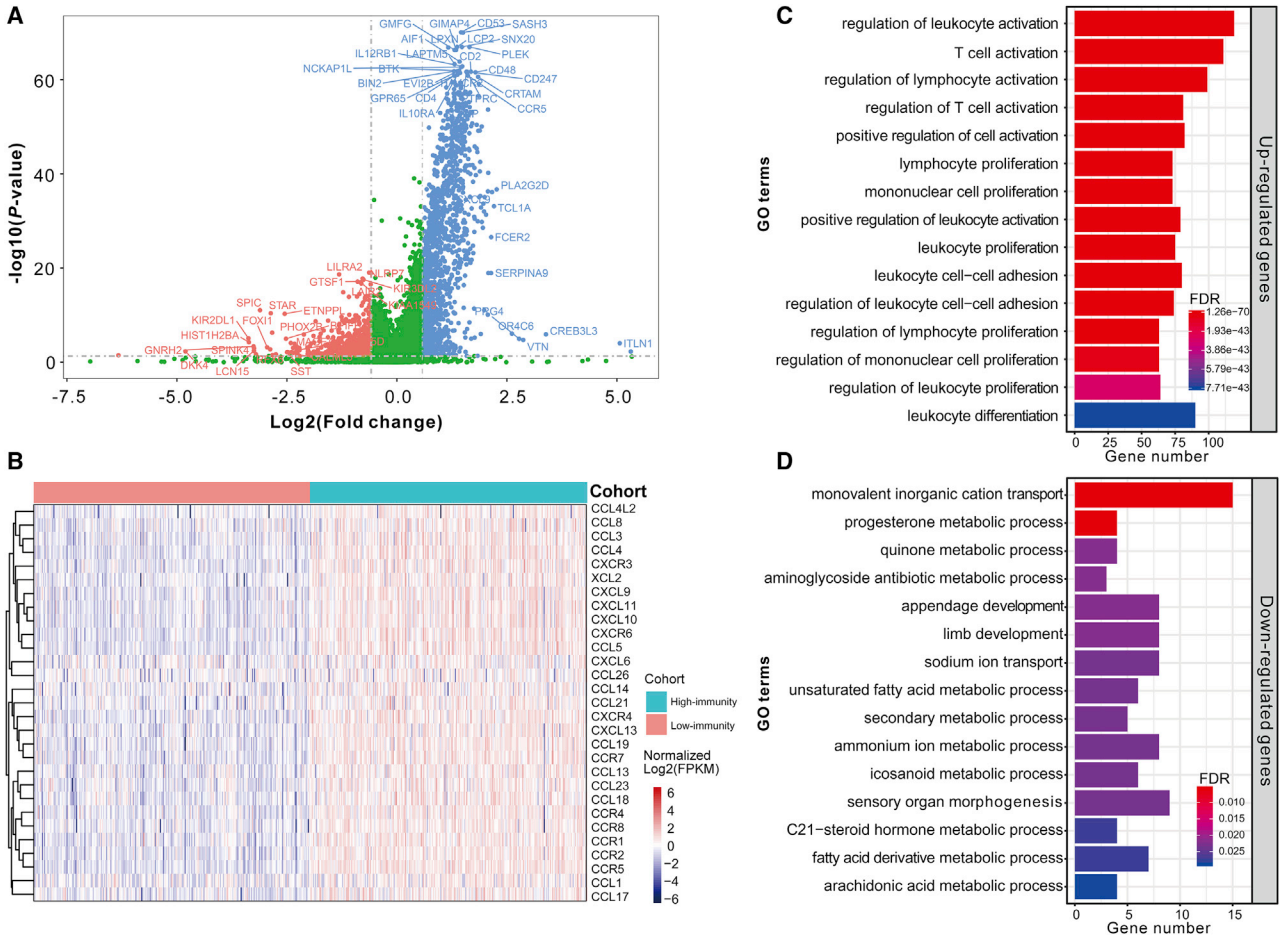


Figure 2. Investigation of the Immune Infiltration-Dependent Expression Change

(A) Volcano plot showing the differentially upregulated (blue nodes) and downregulated genes (red nodes). (B) Heatmap displaying the expression changes of 29 chemotactic factors. The colors of the top bar and central plot represent the grouping information and gradient of normalized log_2FPKM (fragments per kilobase of transcript per million mapped reads), respectively. (C) Bar plot showing the biological processes enriched by the upregulated genes. (D) Bar plot showing the biological processes enriched by the downregulated genes. The y axis in (C) and (D) reflects the name of GO terms. The x axis reflects the overlapped gene numbers between each GO term and query gene set. The color of the bars represents the gradient of adjusted p values (FDR correction).

mutations had been called by VarScan2.[39] As shown in Figure 3A and Figure S3A, most genomic variants were missense mutations in both the high-immunity and low-immunity cohorts (around 60%). Therefore, it was necessary to measure the mutation changes, classify the types of variants, and uncover their potential meanings.

From a global perspective, the samples in the low-immunity cohort hold a significantly larger number of variants (median = 235.00, quartile range = 398.00) than those in the high-immunity cohort (median = 172.50, quartile range = 278.75) (Mann-Whitney U test, p = 0.001) (Figure S2). As for SNVs, a total of 64,344 and 88,708 SNVs across all samples in both high-immunity and low-immunity cohorts were detected, of which C>A was the most common type in both the high-immunity cohort (median frequency = 45, median proportion = 42.65%) and low-immunity cohort (median frequency = 65, median proportion = 42.15%). No matter the type of SNV, the mutation numbers in the low-immunity cohort were significantly higher than those in the high-immunity cohort, as shown in Figure 3B (Mann-Whitney U test, $p < 0.05$). The ratios between transversion (Tv) and transition (Ti) in all SNVs were approximate 2:1 and remained stable in both cohorts (Figure S3D). Furthermore, Figure 3C shows that the SNPs, INSs, and DELs in the high-immunity cohort (SNP: median = 167.50, quartile range = 270.75; INS: median = 1, quartile range = 3; DEL: median = 5, quartile range = 8) were also outnumbered by those in the low-immunity cohort (SNP: median = 226, quartile range = 386; INS: median = 2, quartile range = 3; DEL: median = 6, quartile range = 9). Note that the samples in the low-immunity cohort have a significantly higher level of variant allele fractions (VAFs) than do those in the high-immunity cohort (Figure 3D), which had been proposed to be associated with tumor progression and a poorer prognosis,[40,41] which supports the finding of relatively higher tumor purity and lower heterogeneity in the low-immunity

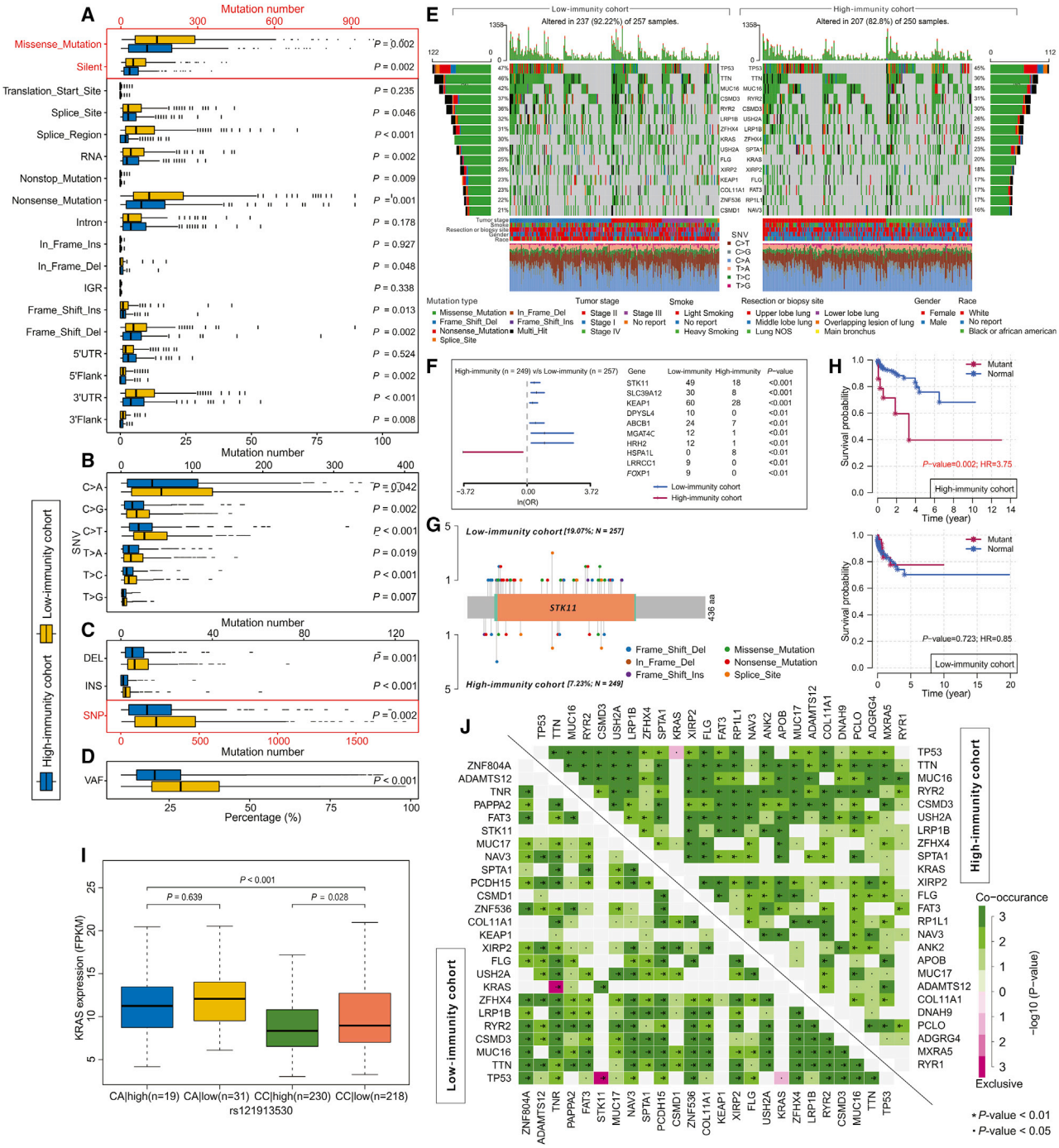


Figure 3. Landscape of Somatic Mutation in High-Immunity and Low-Immunity Cohorts

(A–D) Boxplots showing the comparisons of mutation frequencies of (A) every mutation type classified by effects, (B) SNV, (C) INDEL and SNP, and (D) the percentage of VAF between high-immunity and low-immunity cohorts. (E) Waterfall plot shows the mutation distribution of the top 15 most frequently mutated genes. The central panel shows the types of mutations in each LUAD sample. The upper panel shows the mutation frequency of each LUAD sample. The bar plots on the left and right side show the frequency and mutation type of genes mutated in the low-immunity and high-immunity cohort, respectively. The lower part shows the clinical features (tumor stage, smoke, resection or biopsy site, sex, and race) and SNV types of each sample. The bottom panel is the legend for mutation types and clinical features. (F) Forest plot displays the top 10 most significantly differentially mutated genes between two cohorts. (G) The lollipop plot illustrates the differential distribution of variants for *STK11*. (H) Kaplan-Meier curves show the independent relevance between overall survival time and *STK11* mutation in high-immunity and low-immunity cohorts. (I) The boxplot shows the expression

(legend continued on next page)

samples. Despite the significant differences in variant number of the four types of somatic mutations between the two immune cohorts, the inner constituent ratio of each mutation type occupied in all variants stayed almost unchanged (Figures S3A–3C), suggesting that the differences observed in mutation number are not caused by type shift.

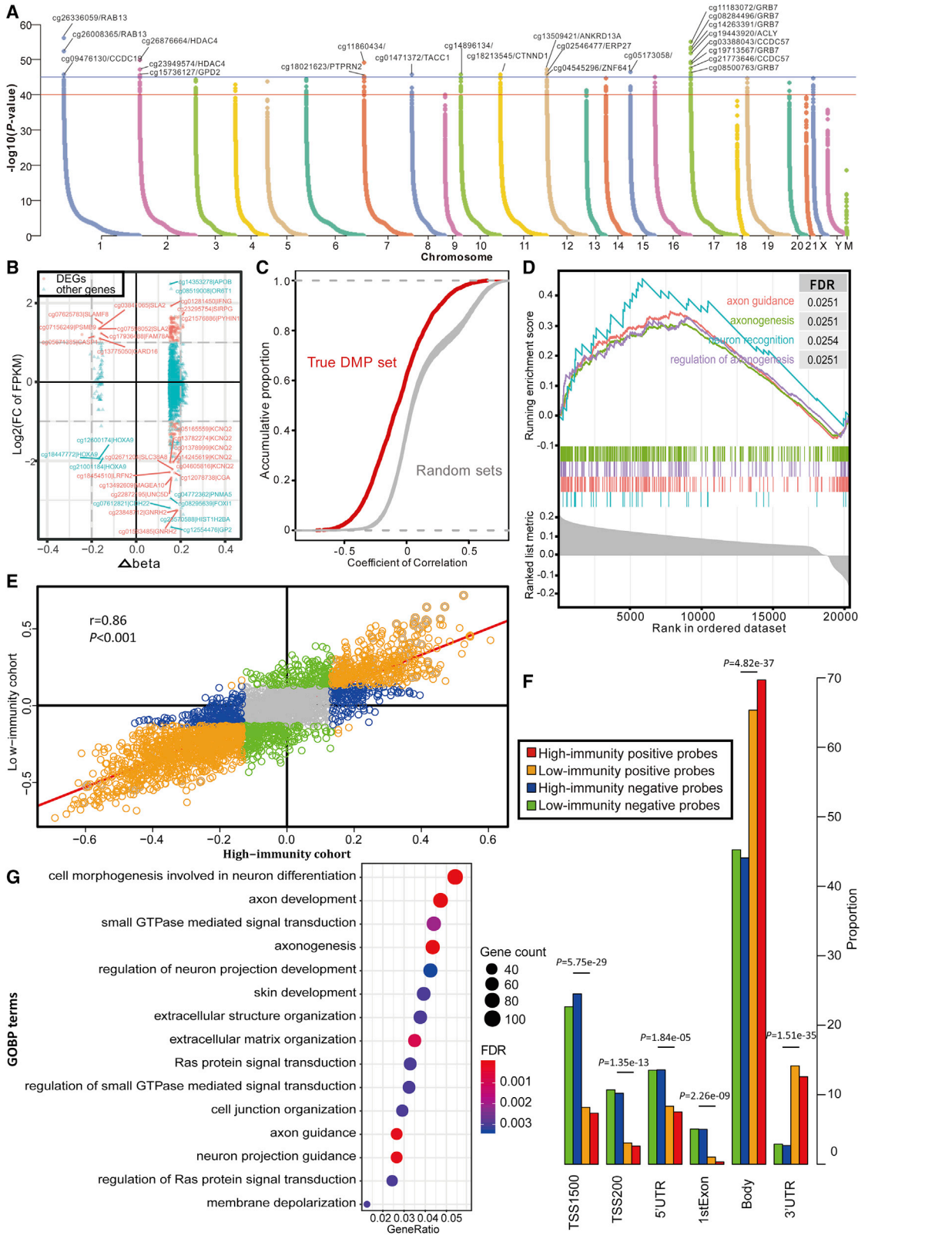
In the low-immunity cohort, 129 genes were mutated in more than 10% of the samples while only 62 genes met this criterion in the high-immunity cohort, of which there was an overlap of 56 genes. The top 15 most frequently mutated genes in the corresponding cohorts are illustrated in Figure 3E. Interestingly, *TP53*, *TTN*, and *MUC16* occupy the top three positions in both cohorts, and there were interactions among them that were regulating various tumor-associated biological processes in LUAD,[42–44] which indicates that they may be less involved in the immune infiltration process but are mainly involved in tumor progression. Next, we investigated the co-occurring and exclusive mutations of the top 25 most frequently mutated genes by using the CoMEt algorithm.[45] Compared with the pervasive co-occurrence landscape, there are three unique cases in two cohorts (*KRAS-TP53*, *KRAS-TNR*, and *STK11-TP53*) that were exhibiting mutually exclusive mutations (Figure 3J), which suggests their probably redundant effect in the same pathway and the selective advantages between them to keep more than one copy of the mutations. More interestingly, some genes had differential mutation frequencies between the two cohorts. From the results, 268 differentially mutated genes were detected using Fisher's exact test, which were sorted in ascending order of p value (Table S2), and the top 10 are shown in Figure 3F. Notably, different variants may exert distinct impacts on other genetic alterations and even the clinical outcome of patients. For example, *KRAS* was detected to hold a differential SNP pattern between high-immunity and low-immunity cohorts (Fisher's exact test, $p < 0.05$). The expression levels of *KRAS* without SNPs were significantly different between the high-immunity and low-immunity cohorts (Mann-Whitney U test, $p < 0.05$), but the reverse was observed when SNP rs121913530 (C>A) exists (Figure 3I). Furthermore, *STK11* is another typical example to demonstrate the different mutation spots between two cohorts (Figure 3G) and the plausible chain reaction of the differences in prognostic impact (Figure 3H).

Depicting DNA Methylation Pattern in TIME of LUAD

Failure to maintain normal DNA methylation, which includes both hypermethylation in CpG islands and hypomethylation in CpG-poor regions, increases the susceptibility to triggering tumor formation and deterioration.[46,47] Therefore, we aimed to detect and compare the effects of DNA methylation patterns in different immune cohorts using Illumina Infinium 450k DNA methylation data from TCGA portal. In this section, 451 samples in which no more than 20% genes have missing beta values were used to detect the differential methylation probes (DMPs) by using ChAMP.[48] A total of 5,764 immune-related DMPs were identified according to the standard of Δbeta >0.15 and FDR <0.05 (Figure 4A; Table S3). Compared with the low-immunity cohort, 5,647 (97.97%) hypermethylated positions involving 2,386 genes were detected in the high-immunity cohort, of which 2,221 loci were located on 1,687 CpG islands. In contrast, the hypomethylated positions were drastically outnumbered with only 117 (2.03%) loci related to 68 genes and were located on 56 CpG islands. Therefore, the high-immunity cohort tends to have hypomethylated positions overall, but hypomethylation only occurs in a few genes. In summary, 2,441 DMP-associated genes were composed of 2,393 genes containing either hypermethylated or hypomethylated positions and 13 genes containing both hypermethylated and hypomethylated positions. Additionally, many DMP-associated genes were found to be differentially expressed between the two cohorts ($\log_2FC > 1$, $p < 0.05$). From the 2,386 hypermethylated genes in the high-immunity cohort, there were 63 upregulated and 32 downregulated DEGs (Figure 4B; Table S4). However, from the hypomethylated gene set, only seven upregulated DEGs were detected (Table S5). The functions of the DMP-associated genes were investigated based on the Gene Ontology (GO) analysis. The top 15 enriched GO terms of biological processes with the lowest FDRs indicated their potential roles in cell differentiation and development (Figure 4G). Meanwhile, they are also significantly enriched in various neuron-associated biological processes. In particular, gene set enrichment analysis (GSEA) of the DMP-associated genes shows that the hypermethylated genes with highly positive beta differences have more essential contributions to the tumor-associated neural biological processes such as axon guidance and axonogenesis (Figure 4D), indicating that the aggressive behaviors induced by the aberrant methylations on tumor immunity are through the recognition and participation in the neural pathways. More interestingly is that the hypermethylation-associated upregulated DEGs were found to participate in immune system activation (Figure S4), which suggests that the probable effect of hypermethylation on gene overexpression is to trigger an elevated level of immune cell infiltration.

While considering the previous finding about the correlations between DNA methylations and gene expression levels,[49] we speculated whether there exists a similar phenomenon in our study and whether the trend is stable across different immune levels. The results showed that out of 2,441 DMP-associated genes, there were 329 positively correlated and 926 negatively correlated genes (Table S6) in the high-immunity cohort, and 346 positively correlated and 939 negatively correlated genes (Table S7) in the low-immunity cohort ($p < 0.05$, Pearson correlation). Therefore, the preferred negative correlations observed in the DMP-associated genes were as postulated. As expected, the probe signals associated with the DMP-associated genes were prone to be negatively correlated with expression level compared with the relatively balanced distributions of correlation coefficients in 100 random sets, which were constructed using the randomly selected

changes of *KRAS* with or without SNP rs121913530 between the high-immunity and low-immunity cohorts. (J) The heatmap illustrates the mutually co-occurring and exclusive mutations of the top 25 frequently mutated genes. The color and symbol in each cell represent the statistical significance of the exclusivity or co-occurrence for each pair of genes.



(legend on next page)

probes (Figure 4C). For example, *S100P*, a well-studied driver in LUAD,[50] was hypomethylated in the low-immunity cohort and had significant correlations between its low methylation signals and high gene expressions (Figure S5). Furthermore, the immunity level did not affect the correlations between methylation levels and expression levels overall, which is supported by the high consistency observed in the correlation coefficients between the two cohorts as shown in Figure 4E ($r = 0.86$, $p < 0.05$, Pearson correlation) and the large overlaps in negatively (Venn diagram in Figure 5B) or positively (Venn diagram in Figure 5A) correlated gene sets between two cohorts. These consistently positively and negatively correlated genes in the low-immunity and high-immunity cohorts were respectively enriched in the immune systems and cell proliferation, while the genes with inconsistent trends had some distinctive functions. For example, the uniquely positively correlated genes in low-immunity cohorts could participate in synapse-associated functions (Figure 5A). A previous study that immune synapse proteins produced by tumor cells will suppress the immune system through epigenetic mechanisms, especially DNA methylation,[51] provides evidence for this finding. Furthermore, the probes of the positively correlated genes were more frequently located in the gene body and 3′ UTR region ($p < 0.001$, chi-square test), while the probes of the negatively correlated genes preferred the regions adjacent to a promoter such as transcription start sites (TSSs), the 5′ UTR, and the first exon (Figure 4F), indicating that there are regional differences in the effect of DNA methylation on expression.

Multi-omics Characteristics Provide an Accurate Prognostic Prediction

From the results obtained in the above sections, some significant immune-related alterations in the multi-omics characteristics, which include expression change, somatic mutation, and DNA methylation, were identified. For expression change, a total of 611 upregulated and 164 downregulated genes were detected in the high-immunity cohort. For somatic mutation, 129 and 62 frequently mutated genes were detected in the low-immunity and high-immunity cohorts, respectively. For DNA methylation, 3,858 out of 5,764 DMPs were located at the regions of 2,442 annotated genes and were differentially methylated.

To identify the immune-related prognostic signatures from the numerous genetic alterations, a four-step strategy (see Materials and Methods) based on lasso regression and Cox proportional hazards regression was adopted. Additionally, the joint and separate effects of the three types of alterations were investigated separately to determine which model had the best performance. On the one hand, for joint effect, all genetic alterations were combined, of which 337 items composed of DEGs, 5 mutations, and 217 DMPs were identified to have a significantly independent effect on the overall survival time of patients using a univariate Cox proportional hazards model. Due to the large number of significant alterations and the possible interactions among them, a lasso regression model was adopted to crop the less contributive variables. Under the optimal parameter $\ln(\text{lambda}) = -3.7$ (Figure 6A), 52 variables were reserved (Table S8) and used to establish the multivariate Cox proportional hazards regression model (stepwise method). For the lack of the matching multi-omics data from other sources, we randomly divided TCGA samples into a training and independent test set. Each third of the samples ($n = 142$) took turns to be the independent test set, and the other two-thirds served as the training set, thus constructing three pairs of sample sets. In the results (Table 1; Figure S6), the performance of the trained models is satisfied with the average concordance index (C-index) equal to 0.839. Next, the risk score for each sample was calculated based on the established models, which has great discriminative power on survival status. The average AUC values of 1-year, 3-year, and 5-year prognosis predictions on training sets reached 0.871, 0.875, and 0.928. With regard to the predictions on the test sets, the performance exhibited a slight decline with the average AUC values of 1-year, 3-year, and 5-year survival equal to 0.796, 0.786, and 0.777. Moreover, the samples were classified into high-risk and low-risk cohorts by median risk score (Table S9; Figure 6F). Kaplan-Meier survival analyses (Figure 6G) show that the high-risk cohort had a poorer overall survival compared with the low-risk cohort ($p < 0.001$).

Considering the great robustness and validity of the above-constructed models, we then combined all TCGA samples and generated an overall prediction model comprising of 27 variables (Figure 6D), of which some such as the DMPs cg10122865 and cg13208438 seemed to be statistically insignificant initially ($p > 0.05$) but are likely associated with other signatures and outcome. The contributions of 27 variables on the overall model are shown in Figure S7. In brief, the expression levels of *PROZ*, *FOXN4*, *LCN15*, *UNC5D*, *CD70*, and *BIRC3*, the methylation levels of cg08780166, cg04240491, cg01090026, cg25407540, and cg26904049 probes, and the mutation of *COL22A1* and *PTPRT* have significantly positive contributions to a poorer prognosis, while the expression levels of *CLEC12A* and *CLEC17A* and the methylation levels of cg22706883, cg03237845, cg07379581, cg15149938, and cg13261536 probes played opposite roles. Moreover, in line with our above findings, the prognosis of

Figure 4. DNA Methylation Pattern in TIME

(A) Manhattan plot of the genome-wide DNA differential methylation in high-immunity and low-immunity cohorts. (B) The relationships between expression change and DNA methylation level. The nodes in red represent the DEGs with $\log_2$FC of FPKM >2 and Δbeta >0.15. (C) The distributions of Pearson correlation coefficients between gene expression level and beta value in true DMP set and 100 random sets. (D) Gene set enrichment analysis (GSEA) shows the significant enrichment in four neuron-associated biological processes. Genes are ranked by Δbeta. (E) Scatterplot showing the correlations of methylation-expression correlation coefficients between high-immunity and low-immunity cohorts. (F) The histogram shows the regional difference between the positive and negative probes. (G) Bubble plot showing the results of GO biological process (GOBP) enrichment analyses on DMP-associated genes. The x axis (fold enrichment) represents the ratio of two proportions, of which one is the proportion of input genes involved in the given GO term and another is the proportion of genes in the given GO term involved in the background.

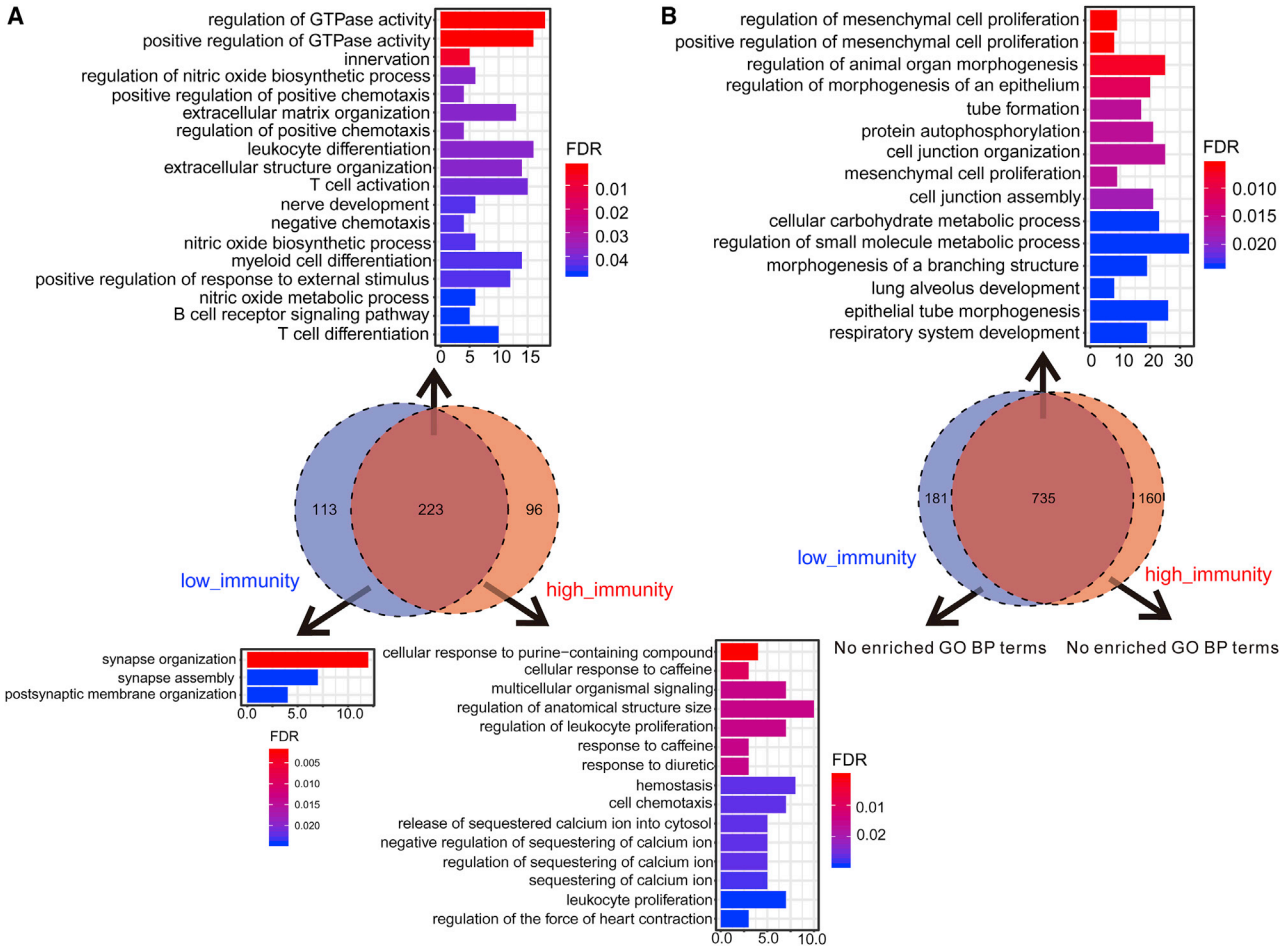


Figure 5. Functional Enrichment Analyses of the Positively and Negatively Correlated Genes

(A) Positively correlated genes. (B) Negatively correlated genes. The results of the consistently correlated genes between the high-immunity and low-immunity cohorts are shown at the top of the Venn diagram, and the results of the uniquely correlated genes are shown on the bottom. The x axis reflects the overlapped gene numbers between each GO term and query gene set. The color of the bars represents the gradient of adjusted p values (FDR correction).

the high-risk cohort is poorer compared to the low-risk cohort according to the results of the survival analysis (Figures 5F and 5G). Furthermore, no matter whether it is the 1-year, 3-year, or 5-year survival rates, the risk score has high discriminative power, with respective AUC values equal to 0.861, 0.850, and 0.916 (Figure 6B; Figure S7). On the other hand, we adopted the same strategy as mentioned above to determine whether the separate effect of each type of genetic alterations has equal or even superior power and whether the differential mutations between the high-immunity and low-immunity cohorts can replace the role of frequently mutated genes in the prediction model. From the results shown in Figure S9, no matter the expression change, somatic mutations, and differential DNA methylation, no single characteristic could provide a powerful enough prognostic prediction compared to the joint model. Additionally, when the frequent somatic mutations were replaced by differential mutations, a 23 variable-based prediction model was obtained, but the performance had no improvement (Figure S10). Also, given the relatively low frequency of these differential mutations, this model was eventually not reserved.

In addition to the genetic alterations, some clinical factors may also have predictive values for the overall survival time. Therefore, the risk score of the joint model was further combined with three clinical and demographic factors (stage, sex, and age) to construct another Cox model. The contributions of the risk score and clinical factors are as shown in Figure S8. Although the stage was significantly associated with overall survival time, the discrimination ability of the new model that incorporated the clinical factor was not improved as shown in Figure 6E (C-index = 0.73), compared with the above overall model constructed only based on multi-omics alterations (C-index = 0.82). Moreover, this new model did not achieve a better performance on its 1-year (AUC = 0.861), 3-year (AUC = 0.848), and 5-year (AUC = 0.914) survival predictions (Figure 6C). We speculated that these clinical factors were highly correlated with and reflected by

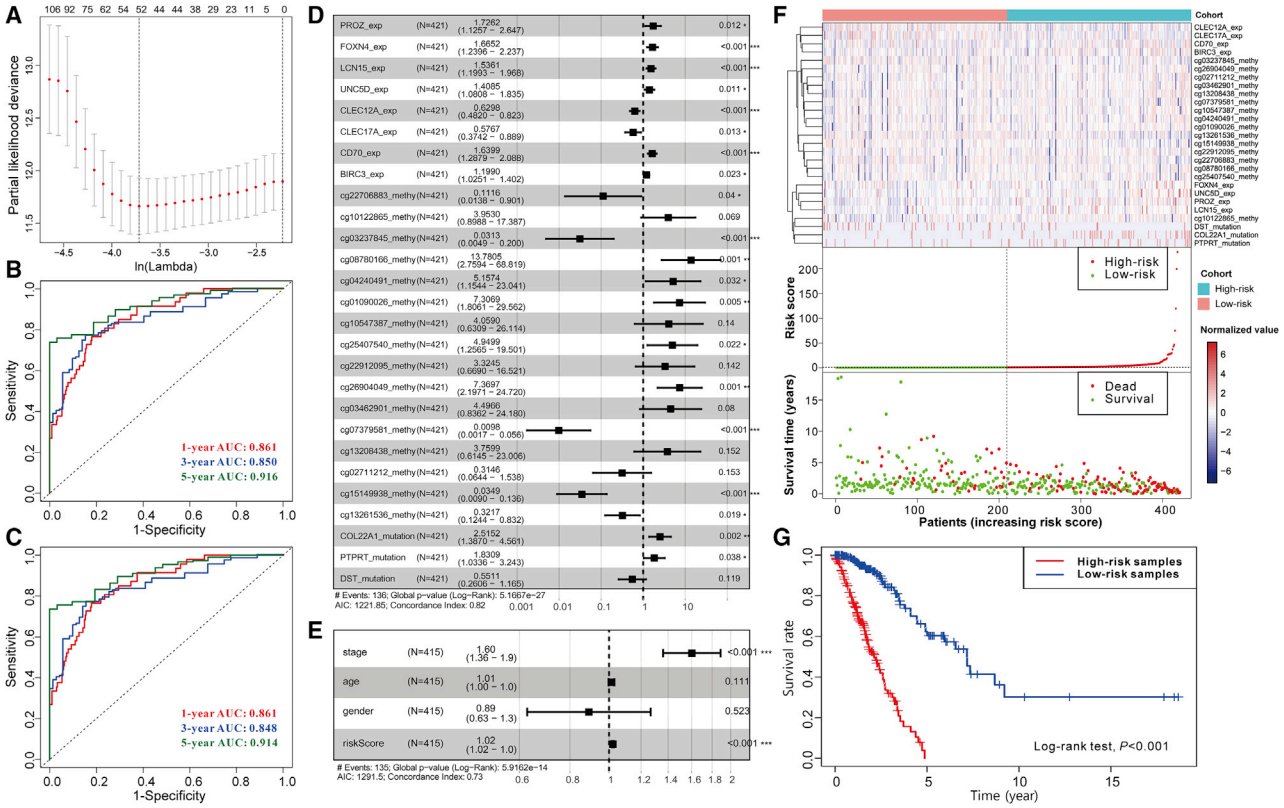


Figure 6. Establishing a Prognostic Model for LUAD

(A) Identification of the optimal penalization coefficient lambda in the Lasso regression model. (B) ROC curves of the risk score for predicting 1-year, 3-year, and 5-year survival. (C) ROC curves of the risk score combined with clinical factors for predicting 1-year, 3-year, and 5-year survival. (D) Forest plot of the prognostic impact of 27 variables. (E) Forest plot of the prognostic impact of risk score and clinical factors. (F) Heatmap in the top panel shows the gradient of 27 variables. The color in the heatmap represents the normalized values of log_2FPKM, log_2Δbeta, and mutation status (0 for non-mutated and 1 for mutated). The scatterplots in the middle and bottom panels respectively illustrate the distribution of risk score and survival status of LUAD patients. (G) Kaplan-Meier curves show the independent relevance between overall survival time and risk scores.

the genetic characteristics. Hence, only multi-omics characteristics comprising the above 27 alterations already could produce an accurate prognostic prediction.

DISCUSSION

Significant effort and attention have been invested to explore the sophisticated mechanism of LUAD, but the current understanding, especially on TME, therapeutic target, and prognostic factors, remains far from satisfactory. In the present study, we first used gene expression profiles to construct the TME of LUAD and conducted a further study on the immune infiltration landscape. As expected, some immune member cells had distinct fractions between the high-immunity and low-immunity cohorts, which had been partially reported before. For instance, memory B cells, a kind of antigen-presenting cell, can drive the expansion and memory formation of T cells, thus exerting their anti-tumor functions in tumors,[52] which conforms to our findings that memory B cells hold a higher fraction in the high-immunity study and are negatively correlated with tumor stage and lymph nodes in LUAD. Then, we turned to investigate the genetic or epigenetic alterations under different immunity infiltration levels and to infer their potential functions. Apart from the canonical immune biological processes such as T cell activation and lymphocyte proliferation, some altered genes are likely able to modulate the immune-related metabolism and neural systems. The genes with downregulated expression levels may have a role in the metabolic communication between tumor cells and infiltrating immune cells. Additionally, the differentially methylated genes may be associated with axon guidance and axonogenesis, thus modulating cell migration and tumor angiogenesis and further promoting LUAD progression. Several studies had reported similar findings in various cancer types.[53–56] Finally, according to the detected genetic or epigenetic alterations, a prognosis model was constructed and was found to have a superior performance with a higher C-index or AUC value compared to the previous models.[14–16] To our surprise, the trained models showed higher accuracy of 5-year survival prediction on the training sets compared with the result from the 1-year survival prediction, whereas the superiority was lost when the prediction was done using the test set. This is likely because the immune-related genetic

Table 1. The AUC Values and C-indexes of Three Trained Models

	Model 1		Model 2		Model 3		Average	
Index	Training Set 1	Test Set 1	Training Set 2	Test Set 2	Training Set 3	Test Set 3	Training Set	Test Set
1-Year AUC	0.881	0.775	0.869	0.862	0.863	0.752	0.871	0.796
3-Year AUC	0.854	0.787	0.873	0.811	0.897	0.761	0.875	0.786
5-Year AUC	0.943	0.723	0.909	0.793	0.933	0.814	0.928	0.777
C-index	0.841		0.839		0.836		0.839	

characteristics have increasingly amplified but uncertain alterations during cancer progression. The uncertainty may be the increasingly different compositions of tumor-infiltrating immune cells for long-term cancer patients. For example, high levels of tumor-infiltrating T cells and macrophages are respectively correlated with better and worse prognosis,[57] which caused higher individual variation, thus making the model unstable when projecting long-term patient survival.

Although this study provides a more comprehensive view into the TIME of LUAD and establishes a powerful model for prognostic prediction, there are still two major drawbacks that require further study. The first drawback, that is, because matching multi-omics data and clinical information are required, we were limited to data solely from TCGA portal and cannot cover other data sources. This hindered us from testing the robustness of the model when used for other data. The second drawback was that the application of the prognostic prediction model required three kinds of omics data, including RNA-seq, WES, and DNA methylation array data, which is cost-intensive and not easy to realize in practical use. Even so, the rapid development of biological technologies will hopefully produce a three-in-one toolkit and thereby blaze the way for its realization and popularization. Yet, despite having such limitations, there is no denying that our study provides significant clues for the elucidation of TIME in LUAD. Moreover, our prognostic prediction model can potentially exhibit compelling clinical value that may lead to the improvement of overall survival and even for the development of new therapeutic strategies for LUAD patients.

MATERIALS AND METHODS

Data Source

The data of 535 LUAD samples and 59 normal control samples of RNA-seq profiles, 561 LUAD samples of WES data, and 504 profiles of the Illumina 450k DNA methylation array were collected from TCGA portal.

TME Construction

First, the ESTIMATE[18] algorithm was used to construct the TME and generate an estimate score, stromal score, and immune score based on the expression profiles. A lower estimate score, stromal score, and immune score respectively represent the higher tumor purity and lower infiltration levels of stromal and immune cells in tumor tissue. Next, the LUAD samples of expression profiles were divided into a high-immunity cohort (n = 268) and low-immunity (n = 267) cohort by the median immune score. The infiltrating immune content for each sample was dissected using the software CIBERSORT.[22] Samples with a p value greater than 0.05 were filtered out, which means that there is no cell type in the signature matrix that can reflect on the given gene expression profile mixture. The fraction levels of all 22 immune cell members between the high-immunity and low-immunity cohorts were compared using the Mann-Whitney U test. For the LUAD samples of the WES and methylation profiles, we also constructed a high-immunity cohort and a low-immunity cohort by mapping the sample IDs of the RNA-seq profiles.

Multi-omics Data Analyses

In this study, we aimed to investigate the differences in gene expression, somatic mutations, and DNA methylation between the high-immunity cohort and low-immunity cohort, respectively, based on the data obtained from TCGA portal. For gene expression change, the differentially expressed genes (DEGs) between the high-immunity cohort (n = 268) and low-immunity cohort (n = 267) were identified based on the RNA-seq data, which should meet the standard of expression fold change greater than 2 and p value adjusted by FDR correction lower than 0.05 (Student's t test). For somatic mutations, WES data of both high-immunity (n = 250) and low-immunity cohorts (n = 257) were used to detect the SNVs, SNPs, and INDELs using the software VarScan2.[39] Fisher's exact test was used to identify the differential mutation pattern, and genes with a p value lower than 0.05 were defined as differentially mutated genes. The co-occurrence and mutually exclusive mutations were identified using the CoMEt algorithm.[45] R package maftools[38] was used to create the visualization of the somatic mutations. For DNA methylation, the R package ChAMP was used to process the Illumina Infinium 450k DNA methylation array data. Samples with more than 20% missing values were filtered out and 451 samples were used, which were further divided into the high-immunity cohort with 219 samples and the low-immunity cohort with 232 samples. The remaining missing values were statistically imputed using ChAMP's imputation function. The beta values were normalized using peak-based correction (PBC). Furthermore, the differential methylation probes and regions were respectively identified using the limma package[58] and Bumphunter algorithm. The correlation between the probe signal and gene expression level was investigated using the Pearson correlation, and the same numbers of probes as in the true DMP set were randomly selected from all probes to construct 100 random sets.

Functional Enrichment Analysis

GO analyses of the DEGs and DMP-associated genes and GSEA were performed using the R package clusterProfiler.[31]

Clinical Relevance Investigation

Clinical factors include the TNM staging system, integrative stage classification, and overall survival. The differences of the scores generated by ESTIMATE and the fractions of immune cell members in distant metastasis and lymph nodes were evaluated using the Mann-Whitney U test, while those in integrative stages and tumor size were evaluated using the Kruskal-Wallis test. Moreover, the contributions to overall survival were investigated by Kaplan-Meier survival analysis (log-rank test).

Establishing and Evaluating the Prognosis Prediction Model

Survival time and status were used to evaluate the prognosis of LUAD patients. We only considered cases that included their overall survival. Next, 494, 490, and 430 cases were reserved to construct the prognosis prediction models based on gene expression values, gene mutation, and methylation probe signals, respectively. For the model with integrating characteristics from multi-omics data, we only considered cases with all of the characteristics being studied, and 421 same cases were obtained. The workflow to explore the prognostic signatures from multiple immunity-dependent genetic or epigenetic alterations in LUAD consists of four steps: (1) univariate Cox proportional hazards regression was used to assess the individual effect of every alteration using the "survival" R package, and then the features with a p value less than 0.05 were selected for further analysis; (2) a lasso regression model was adopted to filter out some less informative variables using the R package glmnet; (3) multivariate Cox proportional hazards regression with a stepwise procedure was used to obtain the genetic variables-based prediction model and to generate the risk scores for all samples using the R package survival. In model construction, TCGA datasets (n = 423) were randomly divided into three equal parts, of which each part took turns to serve as the independent testing set and the remaining two parts to serve as the training set to obtain the optimal variable combination. (4) The risk score of each patient was predicted based on the training model by using the "predict" function included in the survival R package. The risk scores were combined with three other clinical and demographic characteristics (stage, sex, and age) and further introduced into the multivariate Cox proportional hazards regression model to assess the overall effect. Then, the patients were divided into a high-risk group (above the median) and low-risk group (below the median) by the median risk score. We conducted the survival analyses of LUAD patients with respect to the risk level and drew a 1-year, 3-year, and 5-year receiver operating characteristic (ROC) curve using the R package timeROC.[59] The AUC value of the survival ROC curve and C-index were calculated to evaluate the performance of the prognosis prediction model.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.omtn.2020.07.024.

AUTHOR CONTRIBUTIONS

Y.Z. and Q.L. performed most of the analyses and wrote the manuscript. Q.L. and G.Z. conceived and designed the study. G.Z., M.Y., and H.Z. performed some of the analyses and edited the manuscript. D.M.N, M.H., S.H., and T.Y. contributed to additional analyses in the revision procedure.

CONFLICTS OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (grant no. 31970630); the Fundamental Research Funds for the Provincial Universities of Zhejiang (grant no. SJLZ2021001); Ningbo Health Branding Subject Fund (grant no. PPXK2018-05), Natural Science Foundation of Ningbo (grant no. 2017A610154); the Zhejiang Key Laboratory of Pathophysiology (grant no. 201812); The Scientific Innovation Team Project of Ningbo (grant nos. 2016C51001 and 2017C110019); and by the K.C. Wong Magna Fund in Ningbo University.

REFERENCES

1. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. *68*, 394–424.

2. Vaz, M., Hwang, S.Y., Kagiampakis, I., Phallen, J., Patil, A., O'Hagan, H.M., Murphy, L., Zahnow, C.A., Gabrielson, E., Velculescu, V.E., et al. (2017). Chronic cigarette smoke-induced epigenomic changes precede sensitization of bronchial epithelial cells to single-step transformation by KRAS mutations. Cancer Cell *32*, 360–376.e6.

3. Ortega-Gómez, A., Rangel-Escareño, C., Molina-Romero, C., Macedo-Pérez, E.O., Avilés-Salas, A., Lara-García, A., Alanis-Funes, G., Rodríguez-Bautista, R., Hidalgo-Miranda, A., and Arrieta, O. (2016). Gene-expression profiles in lung adenocarcinomas related to chronic wood smoke or tobacco exposure. Respir. Res. *17*, 42.

4. Song, M.A., Benowitz, N.L., Berman, M., Brasky, T.M., Cummings, K.M., Hatsukami, D.K., Marian, C., O'Connor, R., Rees, V.W., Woroszylo, C., and Shields, P.G. (2017). Cigarette filter ventilation and its relationship to increasing rates of lung adenocarcinoma. J. Natl. Cancer Inst. *109*, djx075.

5. Shi, Y., Au, J.S., Thongprasert, S., Srinivasan, S., Tsai, C.M., Khoa, M.T., Heeroma, K., Itoh, Y., Cornelio, G., and Yang, P.-C. (2014). A prospective, molecular epidemiology study of EGFR mutations in Asian patients with advanced non-small-cell lung cancer of adenocarcinoma histology (PIONEER). J. Thorac. Oncol. *9*, 154–162.

6. Herbst, R.S., Morgensztern, D., and Boshoff, C. (2018). The biology and management of non-small cell lung cancer. Nature *553*, 446–454.

7. Sharma, P., Hu-Lieskovan, S., Wargo, J.A., and Ribas, A. (2017). Primary, adaptive, and acquired resistance to cancer immunotherapy. Cell *168*, 707–723.

8. He, J., Hu, Y., Hu, M., and Li, B. (2015). Development of PD-1/PD-L1 pathway in tumor immune microenvironment and treatment for non-small cell lung cancer. Sci. Rep. *5*, 13110.

9. Binnewies, M., Roberts, E.W., Kersten, K., Chan, V., Fearon, D.F., Merad, M., Coussens, L.M., Gabrilovich, D.I., Ostrand-Rosenberg, S., Hedrick, C.C., et al. (2018). Understanding the tumor immune microenvironment (TIME) for effective therapy. Nat. Med. *24*, 541–550.

10. Mariathasan, S., Turley, S.J., Nickles, D., Castiglioni, A., Yuen, K., Wang, Y., Kadel, E.E., III, Koeppen, H., Astarita, J.L., Cubas, R., et al. (2018). TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. Nature *554*, 544–548.

11. Lee, K., Hwang, H., and Nam, K.T. (2014). Immune response and the tumor microenvironment: how they communicate to regulate gastric cancer. Gut Liver *8*, 131–139.

12. Gentles, A.J., Bratman, S.V., Lee, L.J., Harris, J.P., Feng, W., Nair, R.V., Shultz, D.B., Nair, V.S., Hoang, C.D., West, R.B., et al. (2015). Integrating tumor and stromal gene expression signatures with clinical indices for survival stratification of early-stage non-small cell lung cancer. J. Natl. Cancer Inst. *107*, djv211.

13. Zheng, S., Luo, X., Dong, C., Zheng, D., Xie, J., Zhuge, L., Sun, Y., and Chen, H. (2018). A B7-CD28 family based signature demonstrates significantly different prognoses and tumor immune landscapes in lung adenocarcinoma. Int. J. Cancer *143*, 2592–2601.

14. Song, Q., Shang, J., Yang, Z., Zhang, L., Zhang, C., Chen, J., and Wu, X. (2019). Identification of an immune signature predicting prognosis risk of patients in lung adenocarcinoma. J. Transl. Med. *17*, 70.

15. Yue, C., Ma, H., and Zhou, Y. (2019). Identification of prognostic gene signature associated with microenvironment of lung adenocarcinoma. PeerJ *7*, e8128.

16. Tong, Y., Whitford, C.M., Robertsen, H.L., Blin, K., Jørgensen, T.S., Klitgaard, A.K., Gren, T., Jiang, X., Weber, T., and Lee, S.Y. (2019). Highly efficient DSB-free base editing for streptomycetes with CRISPR-BEST. Proc. Natl. Acad. Sci. USA *116*, 20366–20375.

17. Higareda-Almaraz, J.C., Ruiz-Moreno, J.S., Klimentova, J., Barbieri, D., Salvador-Gallego, R., Ly, R., Valtierra-Gutierrez, I.A., Dinsart, C., Rabinovich, G.A., Stulik, J., et al. (2016). Systems-level effects of ectopic galectin-7 reconstitution in cervical cancer and its microenvironment. BMC Cancer *16*, 680.

18. Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Treviño, V., Shen, H., Laird, P.W., Levine, D.A., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. Nat. Commun. *4*, 2612.

19. Morales, M., Planet, E., Arnal-Estape, A., Pavlovic, M., Tarragona, M., and Gomis, R.R. (2011). Tumor-stroma interactions a trademark for metastasis. Breast *20* (*Suppl 3*), S50–S55.

20. Guo, S., and Deng, C.X. (2018). Effect of stromal cells in tumor microenvironment on metastasis initiation. Int. J. Biol. Sci. *14*, 2083–2093.

21. Coussens, L.M., Zitvogel, L., and Palucka, A.K. (2013). Neutralizing tumor-promoting chronic inflammation: a magic bullet? Science *339*, 286–291.

22. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457.

23. Selitsky, S.R., Mose, L.E., Smith, C.C., Chai, S., Hoadley, K.A., Dittmer, D.P., Moschos, S.J., Parker, J.S., and Vincent, B.G. (2019). Prognostic value of B cells in cutaneous melanoma. Genome Med. *11*, 36.

24. Savelyeva, N., Allen, A., Chotprakaikiat, W., Harden, E., Jobsri, J., Godeseth, R., Wang, Y., Stevenson, F., and Ottensmeier, C. (2017). Linked CD4 T cell help: broadening immune attack against cancer by vaccination. Curr. Top. Microbiol. Immunol. *405*, 123–143.

25. Yang, L., and Carbone, D.P. (2004). Tumor-host immune interactions and dendritic cell dysfunction. Adv. Cancer Res. *92*, 13–27.

26. Navegantes, K.C., de Souza Gomes, R., Pereira, P.A.T., Czaikoski, P.G., Azevedo, C.H.M., and Monteiro, M.C. (2017). Immune modulation of some autoimmune diseases: the critical role of macrophages and neutrophils in the innate and adaptive immunity. J. Transl. Med. *15*, 36.

27. Rosenberg, A.S., Pariser, A.R., Diamond, B., Yao, L., Turka, L.A., Lacana, E., and Kishnani, P.S. (2016). A role for plasma cell targeting agents in immune tolerance induction in autoimmune disease and antibody responses to therapeutic proteins. Clin. Immunol. *165*, 55–59.

28. Ryzhov, S., Goldstein, A.E., Matafonov, A., Zeng, D., Biaggioni, I., and Feoktistov, I. (2004). Adenosine-activated mast cells induce IgE synthesis by B lymphocytes: an A2B-mediated process involving Th2 cytokines IL-4 and IL-13 with implications for asthma. J. Immunol. *172*, 7726–7733.

29. Goc, J., Germain, C., Vo-Bourgais, T.K., Lupo, A., Klein, C., Knockaert, S., de Chaisemartin, L., Ouakrim, H., Becht, E., Alifano, M., et al. (2014). Dendritic cells in tumor-associated tertiary lymphoid structures signal a Th1 cytotoxic immune contexture and license the positive prognostic value of infiltrating $CD8^+$ T cells. Cancer Res. *74*, 705–715.

30. Zheng, X., Hu, Y., and Yao, C. (2017). The paradoxical role of tumor-infiltrating immune cells in lung cancer. Intractable Rare Dis. Res. *6*, 234–241.

31. Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287.

32. Deng, M., Gui, X., Kim, J., Xie, L., Chen, W., Li, Z., He, L., Chen, Y., Chen, H., Luo, W., et al. (2018). LILRB4 signalling in leukaemia cells mediates T cell suppression and tumour infiltration. Nature *562*, 605–609.

33. Manandhar, S., and Lee, Y.M. (2018). Emerging role of RUNX3 in the regulation of tumor microenvironment. BMB Rep. *51*, 174–181.

34. Zhou, W.N., Du, Y.F., Bai, J., Song, X.M., Zheng, Y., Yuan, H., Zhang, W., Zhang, Z.D., and Wu, Y.N. (2017). RUNX3 plays a tumor suppressor role by inhibiting cell migration, invasion and angiogenesis in oral squamous cell carcinoma. Oncol. Rep. *38*, 2378–2386.

35. Li, K., Zhu, Z., Luo, J., Fang, J., Zhou, H., Hu, M., Maskey, N., and Yang, G. (2015). Impact of chemokine receptor CXCR3 on tumor-infiltrating lymphocyte recruitment associated with favorable prognosis in advanced gastric cancer. Int. J. Clin. Exp. Pathol. *8*, 14725–14732.

36. Ho, P.C., and Liu, P.S. (2016). Metabolic communication in tumors: a new layer of immunoregulation for immune evasion. J. Immunother. Cancer *4*, 4.

37. Xu, M., Zhong, Y., Yang, F., Liu, K., and Zhuang, B. (2020). Identification and validation of AKR1C1/2/3 as hepatocellular carcinoma risk modules via an integrated strategy. Research Square. https://doi.org/10.21203/rs.2.24543/v1.

38. Mayakonda, A., Lin, D.C., Assenov, Y., Plass, C., and Koeffler, H.P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. *28*, 1747–1756.

39. Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. *22*, 568–576.

40. Sallman, D.A., and Padron, E. (2016). Integrating mutation variant allele frequency into clinical practice in myeloid malignancies. Hematol. Oncol. Stem Cell Ther. *9*, 89–95.

41. Shin, H.T., Choi, Y.L., Yun, J.W., Kim, N.K.D., Kim, S.Y., Jeon, H.J., Nam, J.Y., Lee, C., Ryu, D., Kim, S.C., et al. (2017). Prevalence and detection of low-allele-fraction variants in clinical cancer samples. Nat. Commun. *8*, 1377.

42. Qi, W., Qiao, D., and Martinez, J.D. (2002). Caffeine induces TP53-independent G_1-phase arrest and apoptosis in human lung tumor cells in a dose-dependent manner. Radiat. Res. *157*, 166–174.

43. Cheng, X., Yin, H., Fu, J., Chen, C., An, J., Guan, J., Duan, R., Li, H., and Shen, H. (2019). Aggregate analysis based on TCGA: TTN missense mutation correlates with favorable prognosis in lung squamous cell carcinoma. J. Cancer Res. Clin. Oncol. *145*, 1027–1035.

44. Lakshmanan, I., Salfity, S., Seshacharyulu, P., Rachagani, S., Thomas, A., Das, S., Duan, R., Li, H., and Shen, H. (2017). MUC16 regulates TSPYL5 for lung cancer cell growth and chemoresistance by suppressing p53. J. Cancer Res. Clin. Oncol. *145*, 1027–1035.

45. Leiserson, M.D., Wu, H.T., Vandin, F., and Raphael, B.J. (2015). CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. Genome Biol. *16*, 160.

46. Soozangar, N., Sadeghi, M.R., Jeddi, F., Somi, M.H., Shirmohamadi, M., and Samadi, N. (2018). Comparison of genome-wide analysis techniques to DNA methylation analysis in human cancer. J. Cell. Physiol. *233*, 3968–3981.

47. Koch, A., Joosten, S.C., Feng, Z., de Ruijter, T.C., Draht, M.X., Melotte, V., Smits, K.M., Veeck, J., Herman, J.G., Van Neste, L., et al. (2018). Analysis of DNA methylation in cancer: location revisited. Nat. Rev. Clin. Oncol. *15*, 459–466.

48. Morris, T.J., Butcher, L.M., Feber, A., Teschendorff, A.E., Chakravarthy, A.R., Wojdacz, T.K., and Beck, S. (2014). ChAMP: 450k Chip Analysis Methylation Pipeline. Bioinformatics *30*, 428–430.

49. Greenberg, M.V.C., and Bourc'his, D. (2019). The diverse roles of DNA methylation in mammalian development and disease. Nat. Rev. Mol. Cell Biol. *20*, 590–607.

50. Chien, M.H., Lee, W.J., Hsieh, F.K., Li, C.F., Cheng, T.Y., Wang, M.Y., Chen, J.-S., Chow, J.-M., Jan, Y.-H., Hsiao, M., et al. (2015). Keap1-Nrf2 interaction suppresses

cell motility in lung adenocarcinomas by targeting the S100P protein. Clin. Cancer Res. *21*, 4719–4732.

51. Berglund, A., Mills, M., Putney, R.M., Hamaidi, I., Mulé, J., and Kim, S. (2020). Methylation of immune synapse genes modulates tumor immunogenicity. J. Clin. Invest. *130*, 974–980.

52. Yuen, G.J., Demissie, E., and Pillai, S. (2016). B lymphocytes and cancer: a love-hate relationship. Trends Cancer *2*, 747–757.

53. Tang, H., Wei, P., Duell, E.J., Risch, H.A., Olson, S.H., Bueno-de-Mesquita, H.B., Gallinger, S., Holly, E.A., Petersen, G., Bracci, P.M., et al. (2014). Axonal guidance signaling pathway interacting with smoking in modifying the risk of pancreatic cancer: a gene- and pathway-based interaction analysis of GWAS data. Carcinogenesis *35*, 1039–1045.

54. Yang, D., Qu, F., Cai, H., Chuang, C.H., Lim, J.S., Jahchan, N., Grüner, B.M., S Kuo, C., Kong, C., Oudin, M.J., et al. (2019). Axon-like protrusions promote small cell lung cancer migration and metastasis. eLife *8*, e50616.

55. Nasarre, P., Potiron, V., Drabkin, H., and Roche, J. (2010). Guidance molecules in lung cancer. Cell Adhes. Migr. *4*, 130–145.

56. Yang, X., Deng, Y., He, R.Q., Li, X.J., Ma, J., Chen, G., and Hu, X.H. (2018). Upregulation of HOXA11 during the progression of lung adenocarcinoma detected via multiple approaches. Int. J. Mol. Med. *42*, 2650–2664.

57. Gonzalez, H., Hagerling, C., and Werb, Z. (2018). Roles of the immune system in cancer: from tumor initiation to metastatic progression. Genes Dev. *32*, 1267–1284.

58. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

59. Blanche, P., Dartigues, J.F., and Jacqmin-Gadda, H. (2013). Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat. Med. *32*, 5381–5397.

OMTN, Volume *21*

Supplemental Information

Multi-omics Data Analyses Construct TME and Identify the Immune-Related Prognosis Signatures in Human LUAD

Yuwei Zhang, Minglei Yang, Derry Minyao Ng, Maria Haleem, Tianfei Yi, Shiyun Hu, Huangkai Zhu, Guofang Zhao, and Qi Liao

Supplemental Figures

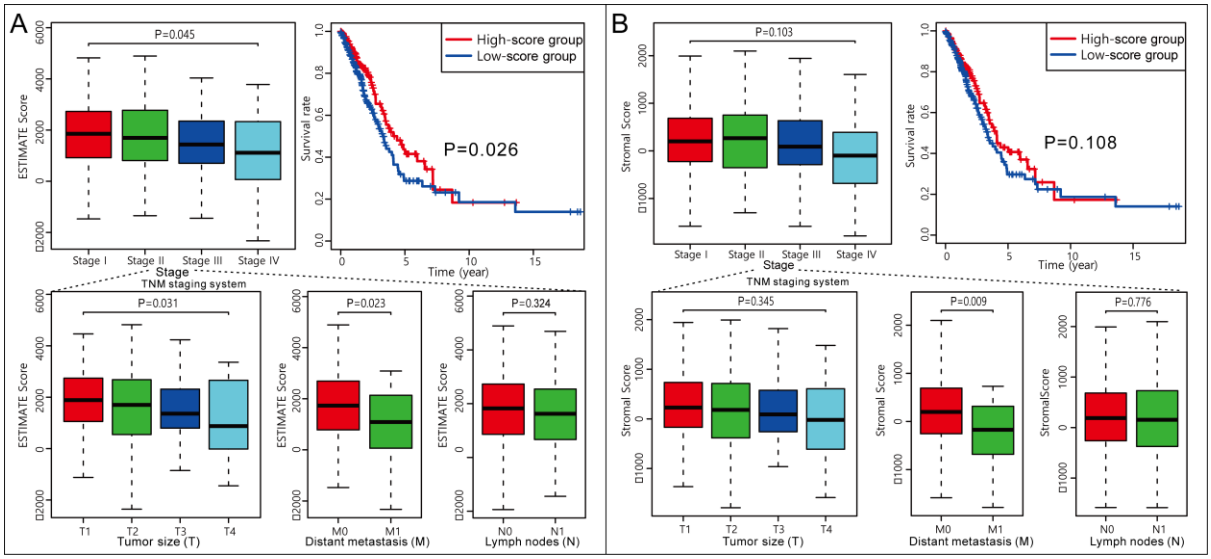


Figure S1 The comparisons of the distributions of **(A)** estimate scores and **(B)** stromal scores on TNM stage, tumor size, distant metastasis, lymph nodes and overall survival time.

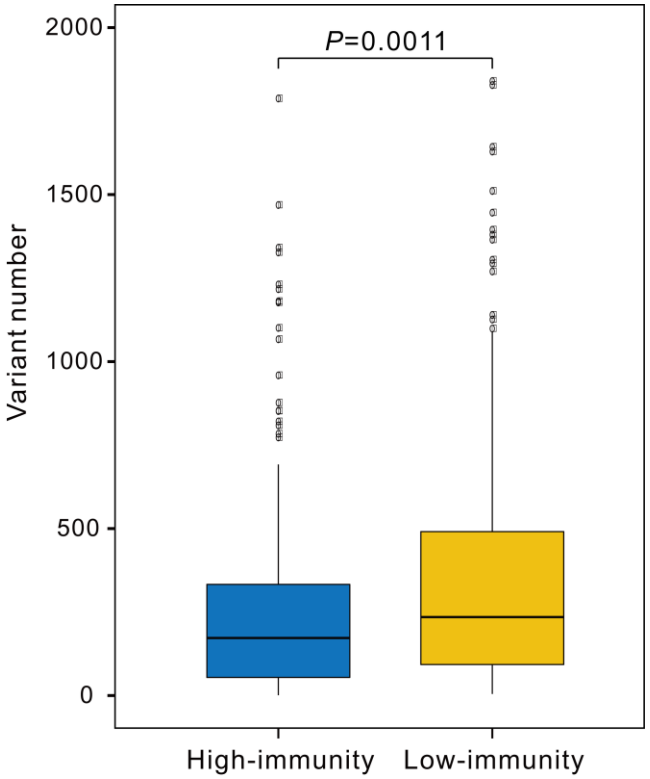


Figure S2 The comparison of the variant number between high-immunity and low-immunity cohorts.

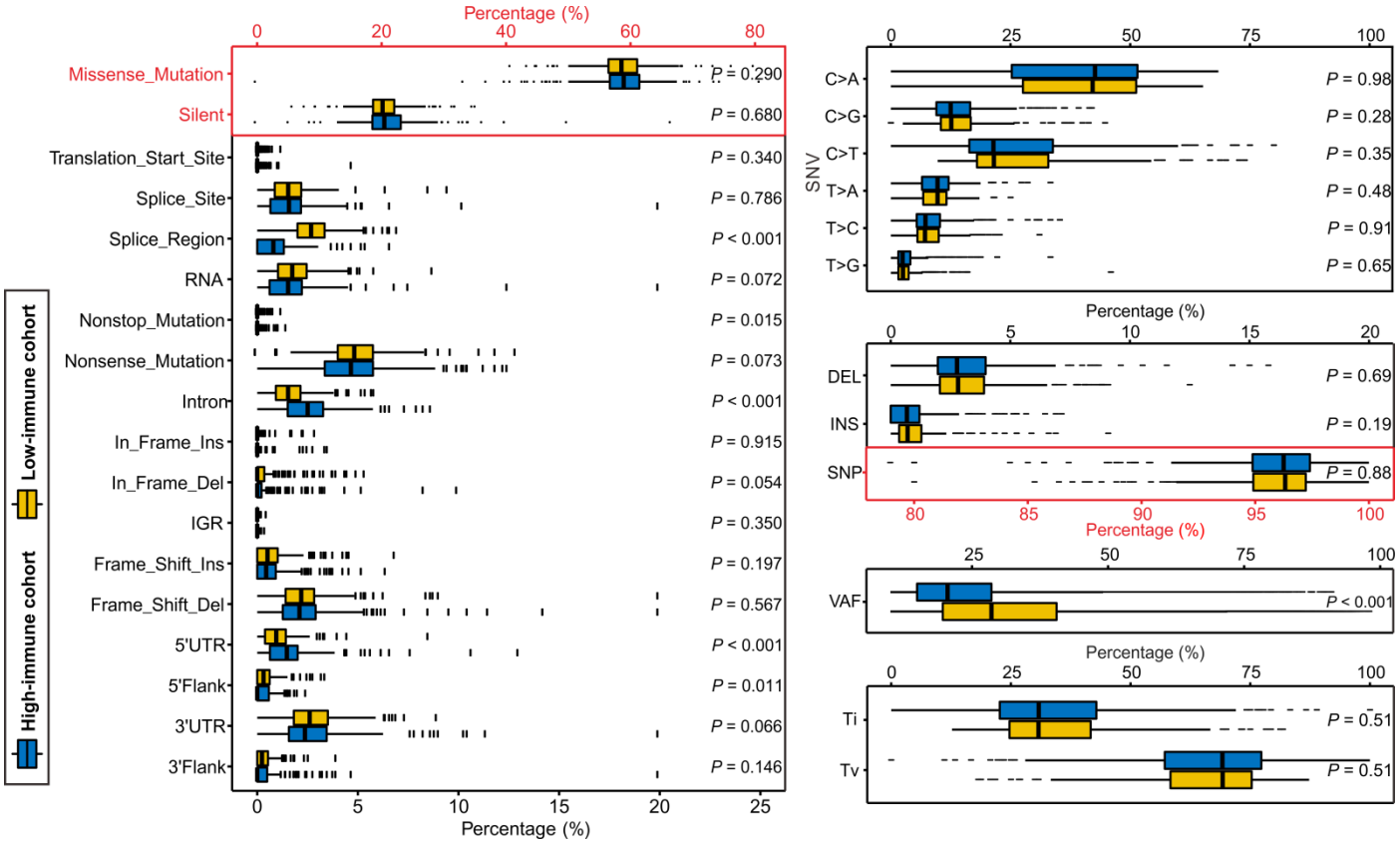


Figure S3 The percentages of various mutation types in high-immunity and low-immunity cohorts. Boxplots respectively display the comparisons of the percentages of **(A)** every mutation type classified by effects, **(B)** SNV, **(C)** INDEL and SNP, and **(D)** DNA substitution types including transition (Ti) and transversion (Tv).

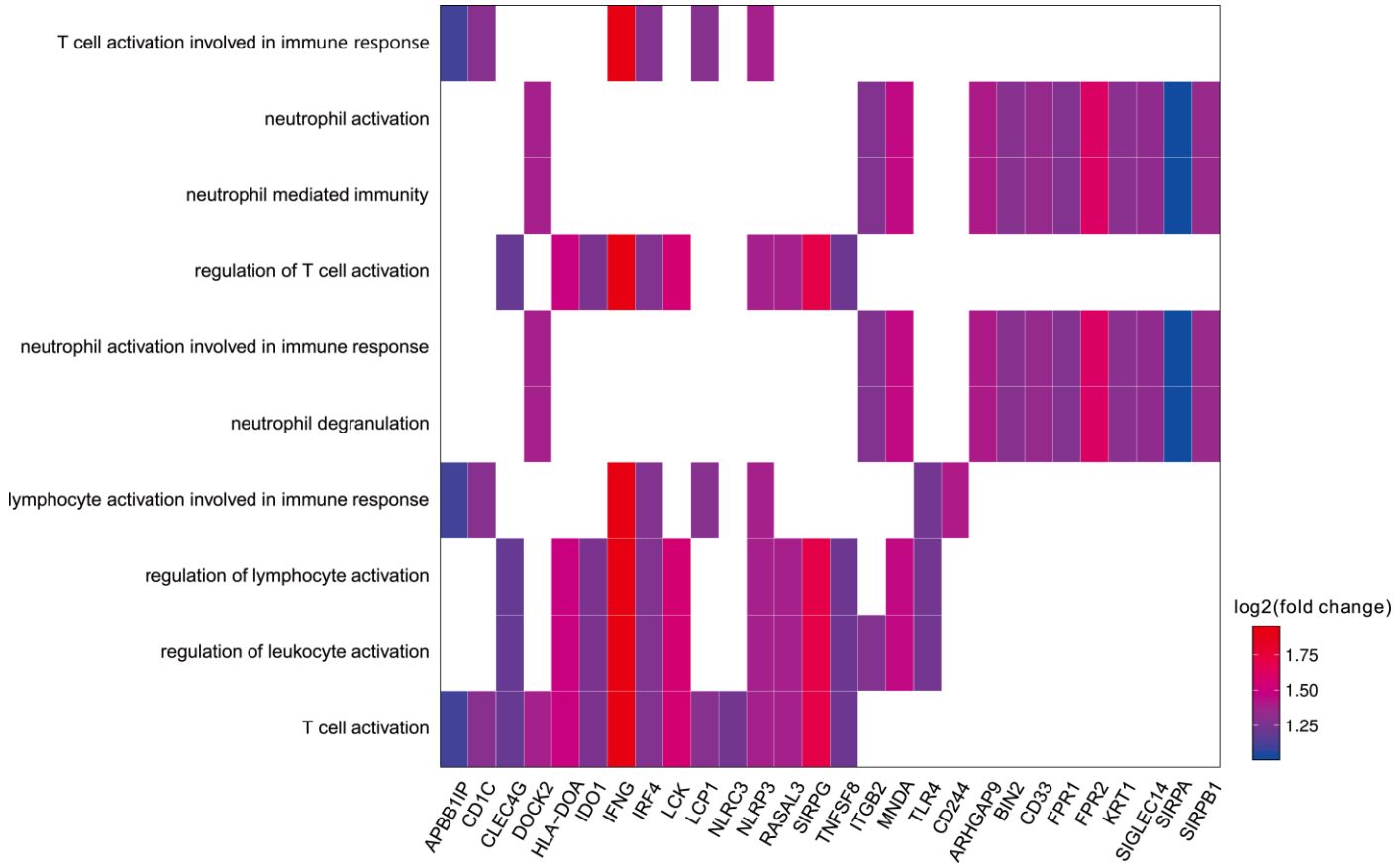


Figure S4 Functional enrichment analysis of the hypermethylation-associated up-regulated DEGs in high –immunity cohort. The y-axis and x-axis respectively represent the significant GOBP terms and the genes in terms. The color of the cells represents the gradient of expression fold change.

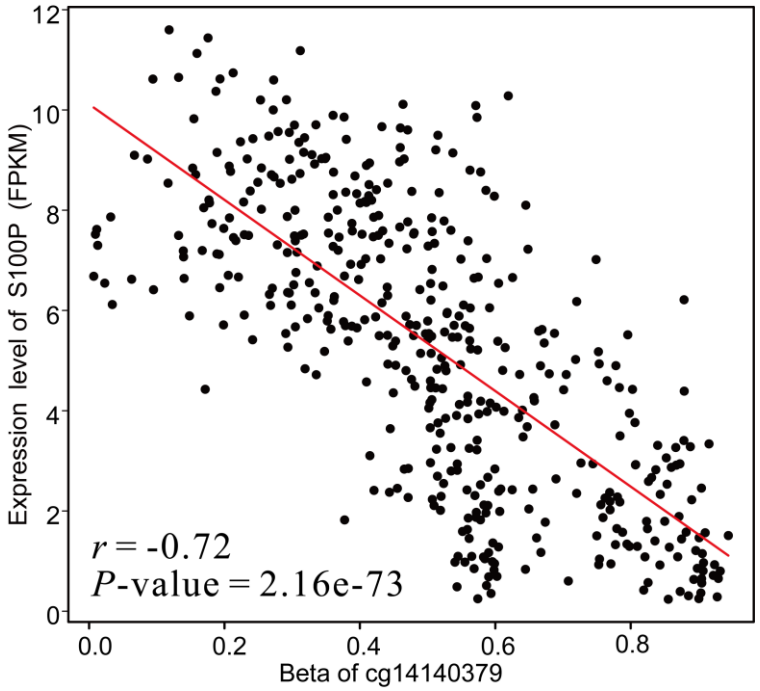


Figure S5 The correlation between the expression level and the beta value of *S100P*.

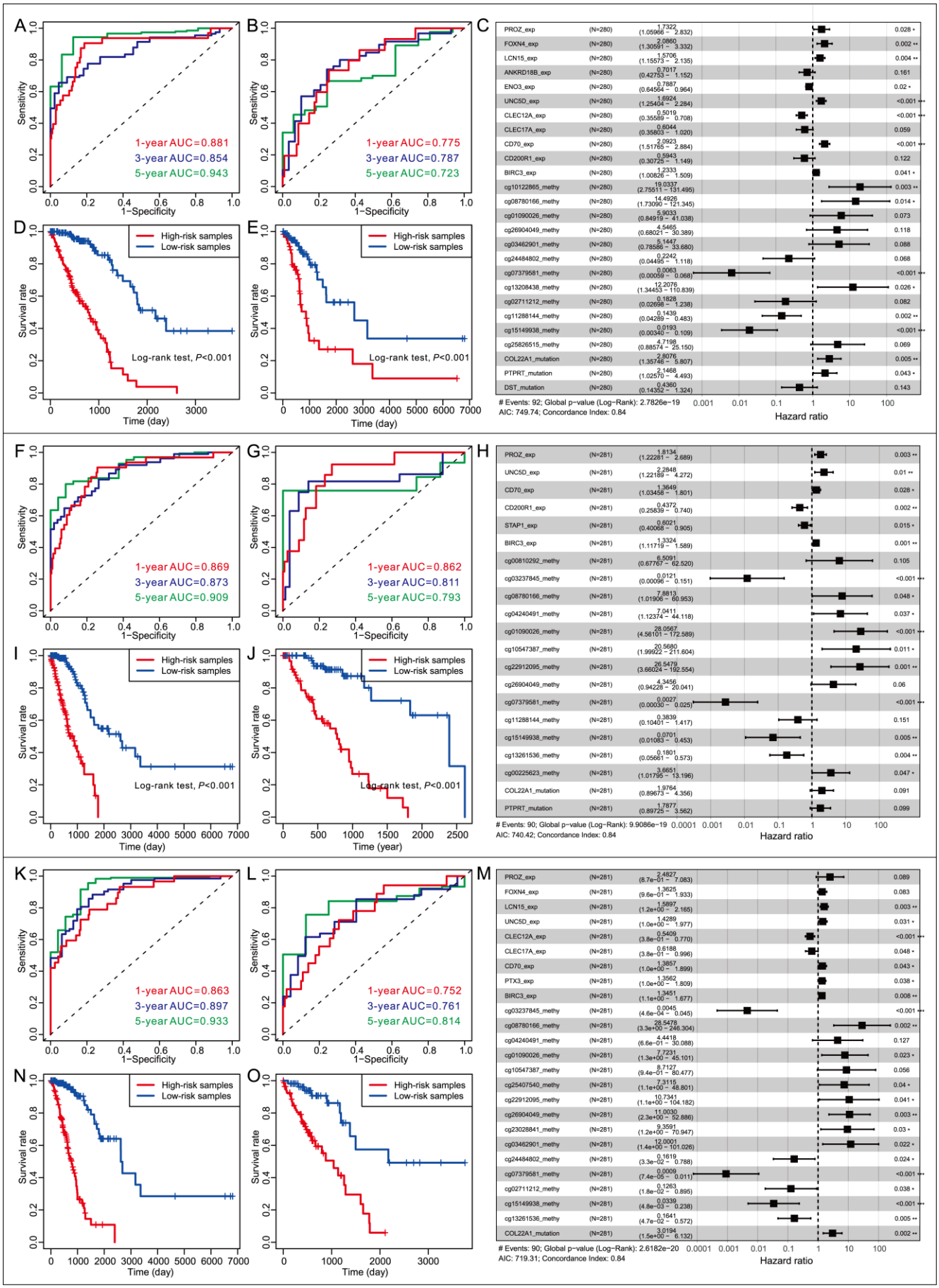


Figure S6 The performance of the trained model 1 (A-E), model 2 (F-J) and model 3 (K-O). Subplot (A/F/K) and (B/G/L) show the roc curves of the risk score for 1-year, 3-year and 5-year survival prediction on the training set and test set, respectively. Subplot (D/I/N) and (E/J/O) are the Kaplan-Meier curves showing the independent relevance between overall survival time and risk scores when training sets and test sets were respectively used. Subplot C, H, and M are the forest plots showing the prognostic impact of variables.

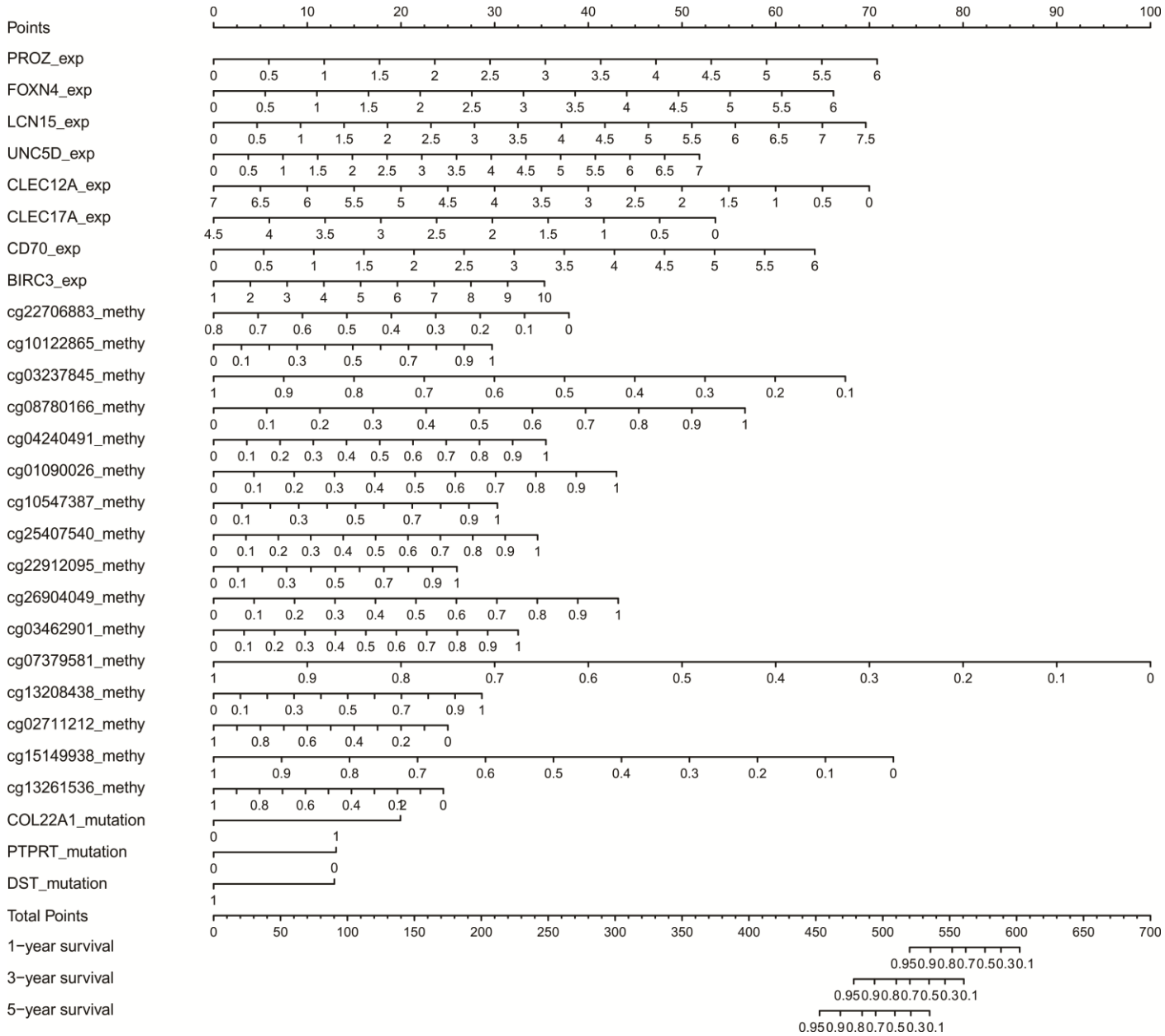


Figure S7 Nomogram for prediction of overall survival time based on multi-Omics characteristics in LUAD. Each variable is assigned a point value. The total points in the bottom panel can be used to estimate the probability of 1-year, 3-year and 5-year survival.

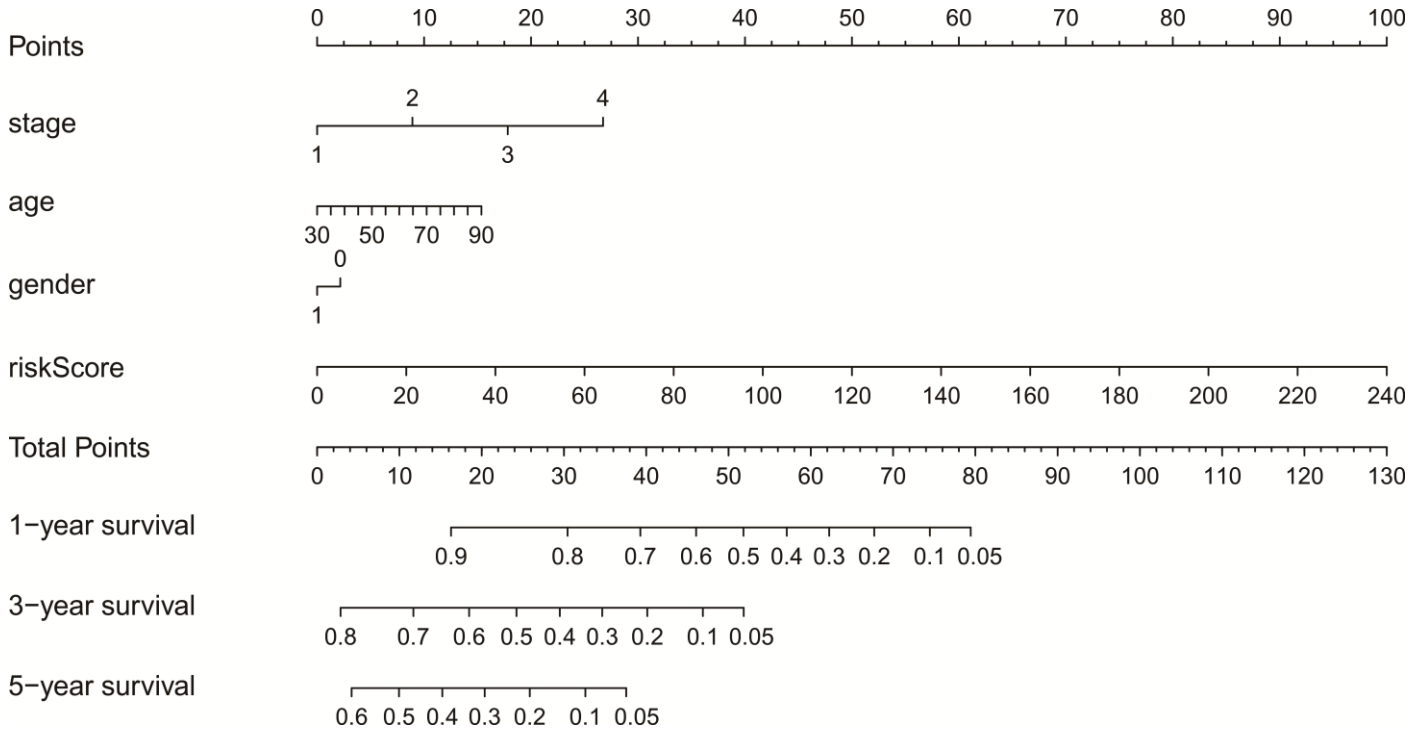


Figure S8 Nomogram for prediction of overall survival time based on the combination of multi-Omics characteristics and clinical factors in LUAD. Each variable is assigned a point value. The total points in the bottom panel can be used to estimate the probability of 1-year, 3-year and 5-year survival.

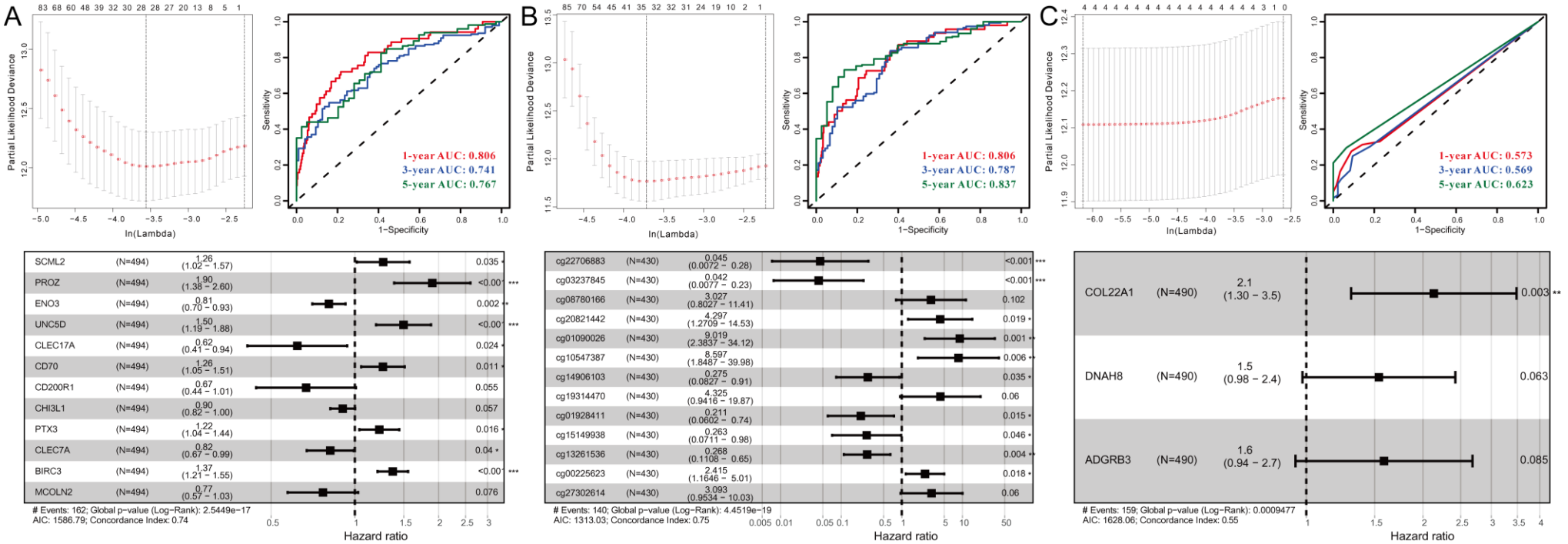


Figure S9 The prognosis prediction models based on **(A)** gene expression change, **(B)** DNA methylation and **(C)** somatic mutation**.**

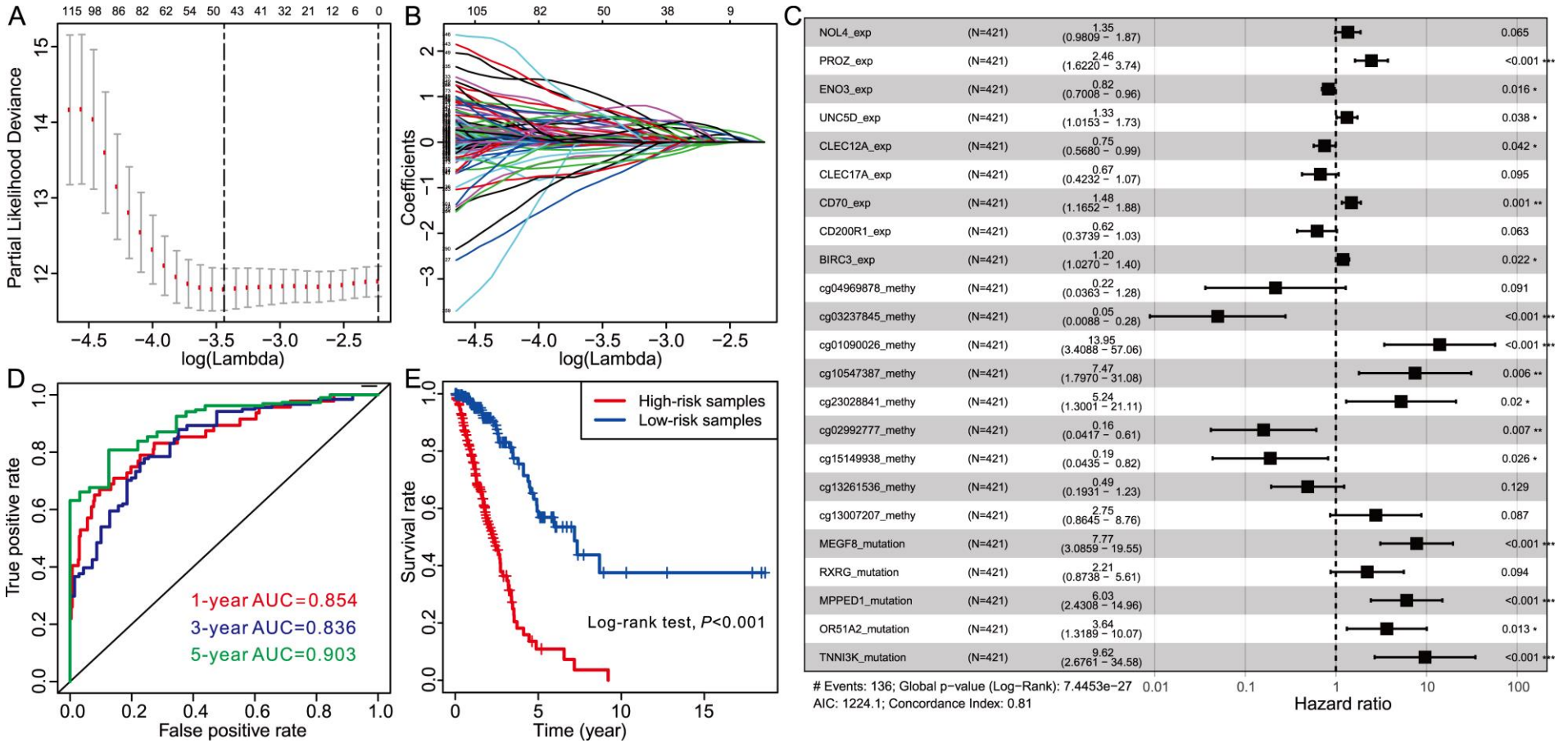


Figure S10 The prognosis prediction model based on differential mutations, gene expression change and DNA methylation. (A-B) Identification of the optimal penalization coefficient lambda in Lasso regression model. **(C)** Forest plot of the prognostic impact of 23 variables. **(D)** ROC curves of the risk score for predicting 1-year, 3-year and 5-year survival. **(E)** Kaplan-Meier curves show the independent relevance between overall survival time and risk scores.

Supplemental Tables

Table S1 The results of pairwise comparisons of estimate score and immune score on tumor size and stage.
(see Excel file)

Table S2 The list of the differentially mutated genes between high-immunity and low-immunity cohorts.
(see Excel file)

Table S3 The list of the differentially methylated probes.
(see Excel file)

Table S4 The list of the hypermethylation-associated DEGs in high-immunity cohort

Gene_symbol	Expression_status	Methylation_probes
PYHIN1	up-regulated	cg02490589,cg21576886,cg25464787,
PIK3R5	up-regulated	cg02505918,
SIRPA	up-regulated	cg25458147,
IL16	up-regulated	cg05042034,
NLRP3	up-regulated	cg25602756,cg18126557,cg07164722,
BIN2	up-regulated	cg02023138,
CD1E	up-regulated	cg09478478,cg10976626,
KLHL6	up-regulated	cg02939781,
DNAJC5B	up-regulated	cg21649629,
LSP1	up-regulated	cg23670188,
IDO1	up-regulated	cg24188163,
RGS1	up-regulated	cg22360028,
XIRP1	up-regulated	cg07165610,
AGAP2	up-regulated	cg23377551,cg11511175,cg00769161,cg15539944,cg14845962,cg12116027,
CD244	up-regulated	cg13240206,
SIGLEC14	up-regulated	cg03240663,
HLA-DQA2	up-regulated	cg17471365,
CLEC4G	up-regulated	cg05626117,
RASGRP2	up-regulated	cg22749107,
ACSM5	up-regulated	cg02147009,
IRF4	up-regulated	cg17336615,
CD1B	up-regulated	cg04574507,cg15952487,
KRT1	up-regulated	cg10538533,cg10926148,
S100B	up-regulated	cg11064537,
IFNG	up-regulated	cg01281450,
LCP2	up-regulated	cg21748244,
IKZF1	up-regulated	cg26746878,
ARHGAP15	up-regulated	cg17600630,cg27492839,
LCP1	up-regulated	cg11274337,
FCER1A	up-regulated	cg14696870,
CIITA	up-regulated	cg07371747,
MAP4K1	up-regulated	cg05258935,cg08230957,
APOE	up-regulated	cg14123992,cg01032398,
CD1C	up-regulated	cg09149672,
HLA-DOA	up-regulated	cg18030265,cg09132634,cg26175846,cg18043773,cg17777731,cg07699648,
MNDA	up-regulated	cg14216734,
BLK	up-regulated	cg24861686,
DOCK2	up-regulated	cg26794227,cg27591502,cg09175734,cg02732152,cg04730794,cg01057132,cg25076039,
APBB1IP	up-regulated	cg17954226,
TLR4	up-regulated	cg14629571,
ARHGAP9	up-regulated	cg03886558,
CMKLR1	up-regulated	cg13248299,cg07919167,
ZNF831	up-regulated	cg04068564,cg18826352,cg19787906,cg21296613,
TNFRSF8	up-regulated	cg13689066,
WIPF1	up-regulated	cg11778734,cg24401656,
TNFSF8	up-regulated	cg14187242,
KCNE1	up-regulated	cg03123557,

PRKCB	up-regulated	cg00795205,cg04086239,cg07959068,cg06401532,
ACAP1	up-regulated	cg08324090,cg25900902,cg19175364,
FPR1	up-regulated	cg17724366,
FPR2	up-regulated	cg02309029,
NLRC3	up-regulated	cg02819734,
ITGB2	up-regulated	cg27103348,
CADM3	up-regulated	cg11388664,cg24580655,cg01803258,
LCK	up-regulated	cg03243362,
RASAL3	up-regulated	cg01062942,cg22808633,
CD33	up-regulated	cg24809544,
ASB2	up-regulated	cg18485877,
SIRPG	up-regulated	cg23295754,
SIRPB1	up-regulated	cg26616780,cg04103198,cg27365103,
GIMAP4	up-regulated	cg00832923,
LILRB5	up-regulated	cg13535489,
MYO1F	up-regulated	cg22568423,cg08283130,
S100P	down-regulated	cg22266967,cg14140379,cg14323984,cg27027375,cg14900031,
EYS	down-regulated	cg23028841,
CABYR	down-regulated	cg05060976,cg21978251,
SLC38A8	down-regulated	cg02671204,
CGA	down-regulated	cg12078738,
DDC	down-regulated	cg19107055,
GPT2	down-regulated	cg04839706,
FSTL4	down-regulated	cg09024962,
VWA5B2	down-regulated	cg24363374,
MSMB	down-regulated	cg17030820,
F7	down-regulated	cg01089602,cg24269657,cg20329620,cg08266168,
CRABP1	down-regulated	cg07199257,
ATP4B	down-regulated	cg15154339,cg05201344,cg01020475,
LRFN2	down-regulated	cg18454510,
TMED6	down-regulated	cg18368125,cg27009392,cg16148454,
DEPDC7	down-regulated	cg00035630,
RNF186	down-regulated	cg23214395,cg04994456,
KCNQ2	down-regulated	cg04605816,cg01378999,cg13782274,
PROZ	down-regulated	cg16019436,
ATOH7	down-regulated	cg00001874,
HGD	down-regulated	cg04546097,
GREB1	down-regulated	cg25744579,cg14653284,
GNRH2	down-regulated	cg01583485,cg23848712,
OPRD1	down-regulated	cg09970593,
PROC	down-regulated	cg10021288,cg22856114,
PPARGC1A	down-regulated	cg27514608,
PLA2G10	down-regulated	cg27058221,
GAS2	down-regulated	cg25982561,cg17679980,cg23840044,
TFF3	down-regulated	cg21970261,
AKR7A3	down-regulated	cg07447773,
ADPRHL1	down-regulated	cg10624914,cg00767058,
GPX2	down-regulated	cg09643186,cg13844922,cg10880599,

Table S5 The list of the hypomethylation-associated DEGs in high-immunity cohort

Gene_symbol	Expression_status	Methylation_probes
HLA-B	up-regulated	cg25954539,
CASP1	up-regulated	cg05671385,
FAM78A	up-regulated	cg17936488,
SLAMF8	up-regulated	cg07625783,
PSMB9	up-regulated	cg07156249,
CARD16	up-regulated	cg13775050,
SLA2	up-regulated	cg07598052,cg03841065,

Table S6 The list of the genes with expression-methylation correlations in high-immunity cohort (see Excel file)

Table S7 The list of the genes with expression-methylation correlations in low-immunity cohort (see Excel file)

Table S8 The list of 52 reserved variables in the lasso regression model.

No.	Variables
1	PROZ_exp
2	FOXN4_exp
3	LCN15_exp
4	INSL4_exp
5	GAS2_exp
6	ANKRD18B_exp
7	ENO3_exp
8	UNC5D_exp
9	CLEC12A_exp
10	CLEC17A_exp
11	CD70_exp
12	CD200R1_exp
13	PTX3_exp
14	ACSM5_exp
15	STAP1_exp
16	BIRC3_exp
17	LILRA4_exp
18	MCOLN2_exp
19	CCR2_exp
20	cg00810292_methy
21	cg22706883_methy
22	cg26605164_methy
23	cg04969878_methy
24	cg10122865_methy
25	cg03237845_methy
26	cg08780166_methy
27	cg04240491_methy
28	cg01090026_methy
29	cg04275847_methy
30	cg10547387_methy
31	cg25407540_methy
32	cg22912095_methy
33	cg26904049_methy
34	cg23028841_methy
35	cg03462901_methy
36	cg00874480_methy
37	cg24484802_methy
38	cg07379581_methy
39	cg17665652_methy
40	cg13208438_methy
41	cg02711212_methy
42	cg02992777_methy
43	cg11288144_methy
44	cg15149938_methy
45	cg13261536_methy
46	cg13007207_methy
47	cg25826515_methy

48	cg00225623_methy
49	COL22A1_mutation
50	PTPRT_mutation
51	DNAH8_mutation
52	DST_mutation

Table S9 The table shows the contribution of 27 variables and risk score for each sample (see Excel file)