www.aging-us.com AGING 2020, Vol. 12, No. 14

Research Paper

Identifying CpG methylation signature as a promising biomarker for recurrence and immunotherapy in non–small-cell lung carcinoma

Ruihan Luo[1], Jing Song[1], Xiao Xiao[2], Zhengbo Xie[3], Zhiyuan Zhao[3], Wanfeng Zhang[1], Shiqi Miao[1], Yongyao Tang[4], Longke Ran[1]

[1]Department of Bioinformatics, The Basic Medical School of Chongqing Medical University, Chongqing, China
[2]Department of Endocrine and Breast Surgery, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China
[3]Information Center Department, Chongqing Medical University, Chongqing, China
[4]Molecular and Tumor Research Center, Chongqing Medical University, Chongqing, China

Correspondence to: Longke Ran; **email:** ranlongke@cqmu.edu.cn

Keywords: CpG methylation signature, recurrence, non–small-cell lung carcinoma, tumor mutation burden, immunotherapy
Received: March 19, 2020 **Accepted:** June 4, 2020 **Published:** July 28, 2020

ABSTRACT

Epigenetic alterations are crucial to oncogenesis and regulation of gene expression in non–small-cell lung carcinoma (NSCLC). DNA methylation (DNAm) biomarkers may provide molecular-level prediction of relapse risk in cancer. Identification of optimal treatment is warranted for improving clinical management of NSCLC patients. Using machine learning algorithm we identified 4 recurrence predictive CpG methylation markers (cg00253681/ART4, cg00111503/KCNK9, cg02715629/FAM83A, cg03282991/C6orf10) and constructed a risk score model that potently predicted recurrence-free survival and prognosis for patients with NSCLC (P = 0.0002). Integrating genomic, transcriptomic, proteomic and clinical data, the DNAm-based risk score was observed to significantly associate with clinical stage, cell proliferation markers, somatic alterations, tumor mutation burden (TMB) as well as DNA damage response (DDR) genes, and potentially predict the efficacy of immunotherapy. In general, our identified DNAm signature shows a significant correlation to TMB and DDR pathways, and serves as an effective biomarker for predicting NSCLC recurrence and response to immunotherapy. These findings demonstrate the utility of 4-DNAm-marker panel in the prognosis, treatment decision-making and evaluation of therapeutic responses for NSCLC.

INTRODUCTION

Lung cancer is a prominent global health issue and economic burden, with an estimated 40.9 million disability-adjusted life-years in 2017 [1]. Non-small cell lung carcinoma (NSCLC) is the most frequent tumor type of lung cancer, accounting for approximately 87% of lung cancer cases, most of which were diagnosed with advanced stage [2]. In recent years, NSCLC still poses a huge health threat to all mankind in the case of high morbidity and mortality as well as poor prognosis due to late disease diagnosis and not being eligible for curative surgery. Even though traditional adjuvant platinum-based chemotherapy or target therapies have been beneficial for advanced resected tumors, most still have a high relapse risk [3–5]. One potential explanation for these clinical phenomena could be that traditional prognosis indicators are not helpful in making optimal therapy strategies for NSCLC patients with different states of recurrence risk. It would be of great significance to investigate a better prognostic molecular signature to predict recurrence and determine the patients with NSCLC who might benefit most from adjuvant therapies.

It has been well-known that DNA methylation (DNAm), as an epigenetic modulator, regulates gene expression in cancer [6]. The pattern of DNAm alterations which are locus dependent, includes hypo- and hyper-methylation of oncogene and tumor suppressor genes respectively, and has been proved to correlate with oncogenesis, progression and treatment [7, 8]. Methylome profiling carries several benefits: lots of altered CpG sites within DNA methylation target region, relatively stable methylation aberrance, and higher clinical sensitivity in cancer detection [9]. Besides, epigenetic therapy (low doses of DNMTi) exerts durable anti-tumor effect, avoiding acute cytotoxicity [10]. Prior studies showed well-renowned SEPT9 in colorectal cancer (CRC) [11] and MGMT in CRC with metastasis [12] were sensitive and effective methylation markers for diagnosis and prognosis. DNAm locus aberrance was also proved in lung tumor tissues and the epigenetic alterations might associate with prognosis of patients with stage I lung cancer [13]. Nevertheless, little is known about molecular function of specific DNAm markers or a methylation panel, and few of which are with clinical utility and widely accepted for NSCLC patients. Therefore, investigation into clinically effective and reliable DNAm signature is warranted for evaluating relapse risk of NSCLC.

Immune checkpoint blockades (ICBs) therapies in advanced NSCLC patients demonstrated prominent durable response, and higher tumor mutation burden (TMB) correlated to improved relapse-free survival, durable objective response as well as elevated clinical efficacy [14]. Both mutation load and methylation loss accumulate during mitotic cell division [15], and chromosome instability may arise from mutations in a DNA methyltransferase gene [16]. Hence, it's imperative that the DNAm signature and its contribution to immunotherapy responding patient stratification in NSCLC be explored to discover routine and potent biomarkers for identification of potential responders to ICBs treatment.

In this study, we initially identified 4 CpG biomarkers associated with recurrence of NSCLC. Base on TCGA NSCLC cohort comprised of lung adenocarcinomas (LUAD) and lung squamous cell carcinomas (LUSC), a promising DNAm-based risk score model predictive of relapse was constructed and then validated in the other 3 datasets. We further explored molecular mechanism and clinical utility of the DNAm signature. At last we investigated the relevance of the combined DNAm panel with TMB and clinical response to ICBs.

RESULTS

Patient and clinical characteristics in NSCLC cohorts

NSCLC patients included were mainly derived from TCGA LUNG Cancer cohort, GSE39279, GSE66836 and GSE119144 cohorts with clinical characteristics presented in Supplementary Table 1. DNAm data for a total of 827 TCGA NSCLC and 60 GSE119144 tumor samples were available at initial analysis, whereas RFS information of only 662 TCGA NSCLC patients (393 non-recurrence and 271 recurrence tumor tissue samples) and 59 GSE119144 NSCLC patients (10 non-recurrence and 49 recurrence tumors) was complete and available for analysis in training and validation phase, respectively. Beyond all that, we also utilized DNAm data of GSE66836 (164 LUAD samples) from GEO repository. Work flowchart of DNAm prognostic marker selection was depicted in Figure 1.

DNA methylation and gene expression profiles in NSCLC

Analyses of DNAm differences between NCSLC tumors and normal lungs were conducted on TCGA and GSE66836 datasets, revealing that DNA methylation in 11641 overlapping CpGs representing 5359 unique genes were of significant aberration. To keep consistent with loci in GSE39279 and GSE119144 datasets, DMPs were further filtered and 9367 consistent CpGs among training and validation sets were retained. For transcriptomic profiling, differential expression analysis on TCGA RNA-seq data that matched with DNAm profiles showed 1717 significant DEGs, including 1282 upregulated and 435 downregulated genes, and from which 270 potentially hyper- and hypo-methylated genes nearby DMPs mentioned above were identified and used for further shrinking CpGs. After Spearman's correlation test ($r < 0$, Bonferroni corrected $P < 0.05$) on association between DNAm and mRNA expression levels, 102 CpGs representing 87 unique DEGs finally yielded. These DMPs were deemed as biologically meaningful where DNAm changes probably epigenetically regulated and negatively correlated to reference gene expression nearby. On the basis of results above, unsupervised hierarchical clustering of 87 DEGs separated 849 TCGA NSCLC samples into tumor and normal subgroups, revealing 53 genes were upregulated and 34 genes were downregulated in tumor tissues (Figure 2A). Unsupervised clustering analysis of 102 significant DMPs also presented a clear distinction between tumor and normal samples of TCGA DNAm data, in which 57 DMPs were hypomethylated while 45 were hypermethylated in NSCLC tissues (Figure 2B).

Identification of recurrence predictive CpGs for NSCLC

To screen out the DNAm markers predictive of relapse risk for NSCLC patients, methylation values of DMPs in 664 TCGA tumor samples were included into following analyses with machine learning algorithms.

We firstly implemented two methods: LASSO-Logistic regression and Random Forest on modeling 102 aforementioned DMPs for narrowing down markers, identifying 14 and 21 CpGs respectively (Figure 3A, 3B, Supplementary Figure 1, Supplementary Table 2). A total of 11 CpGs were overlapped in results of two algorithms, and 24 CpGs unioned together were then incorporated into LASSO-Cox model, yielding 9 robust prognostic CpG markers (Figure 3C, 3D, Supplementary Table 2). Plus, univariable Cox regression analysis was performed with relapse-free survival data of training set, and 8 CpGs were screened out, with 4 most significant CpG markers identified simultaneously by LASSO-Cox and univariate Cox methods (Figure 3E, Supplementary Table 2, Supplementary Figure 1). By combining CpGs selected from LASSO-Cox and univariate Cox models, 13 predictive biomarkers were obtained. Subsequently, multivariable analysis was conducted on clinicopathological factors in combination with the 13 CpGs (Supplementary Table 3). Using DNAm profile multiplied by coefficients of the multivariate Cox regression model, based on 4 ultimate CpGs (Table 1), a risk score model generated for prediction of recurrence in NSCLC. The median of risk score was set as the cutoff value and NSCLC patients were divided into high-risk group (risk score ≥ -0.0416) and low-risk group (risk score < -0.0416) (Supplementary Figure 2).

We further investigated the potential of risk score model in NSCLC prognostic prediction. Survival analysis on the combined risk score showed that RFS probabilities of NSCLC patients in high- and low-risk groups were of salient difference, and the 4-DNAm-marker panel also presented favorable potential in predicting overall survival (OS) in TCGA NSCLC cohort (Figure 3F).

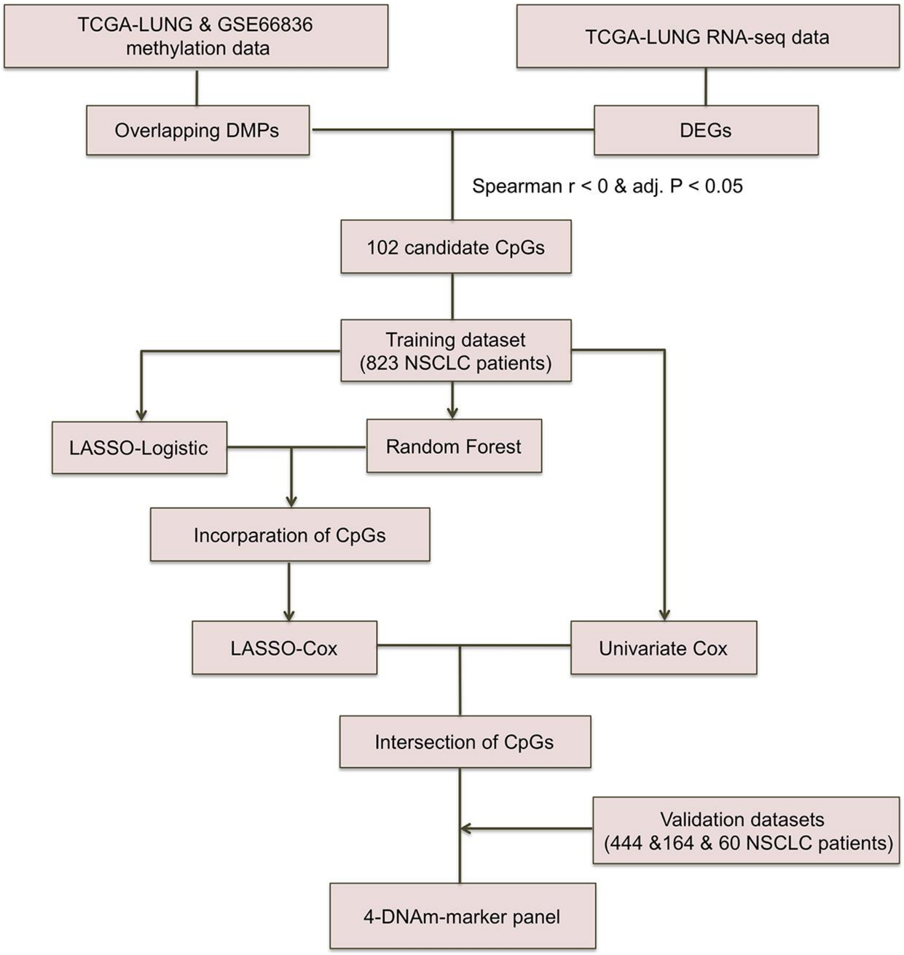


Figure 1. Workflow chart of CpG marker selection. Two DNA methylation (DNAm) datasets and TCGA RNA-seq dataset were used for identifying 102 candidate CpG markers. Based on recurrence-free survival data of the training cohort (823 TCGA NSCLC patients), LASSO-Logistic and Random Forest methods were applied to identify recurrence associated CpG markers. With the incorporation of CpGs identified by two methods above, LASSO-Cox were implemented to select robust DNAm signatures. Using the CpGs overlapped in results of univariate Cox and LASSO-Cox models, the 4-DNAm-marker panel was finally identified and verified in validation cohorts. adj.P: Bonferroni corrected P.

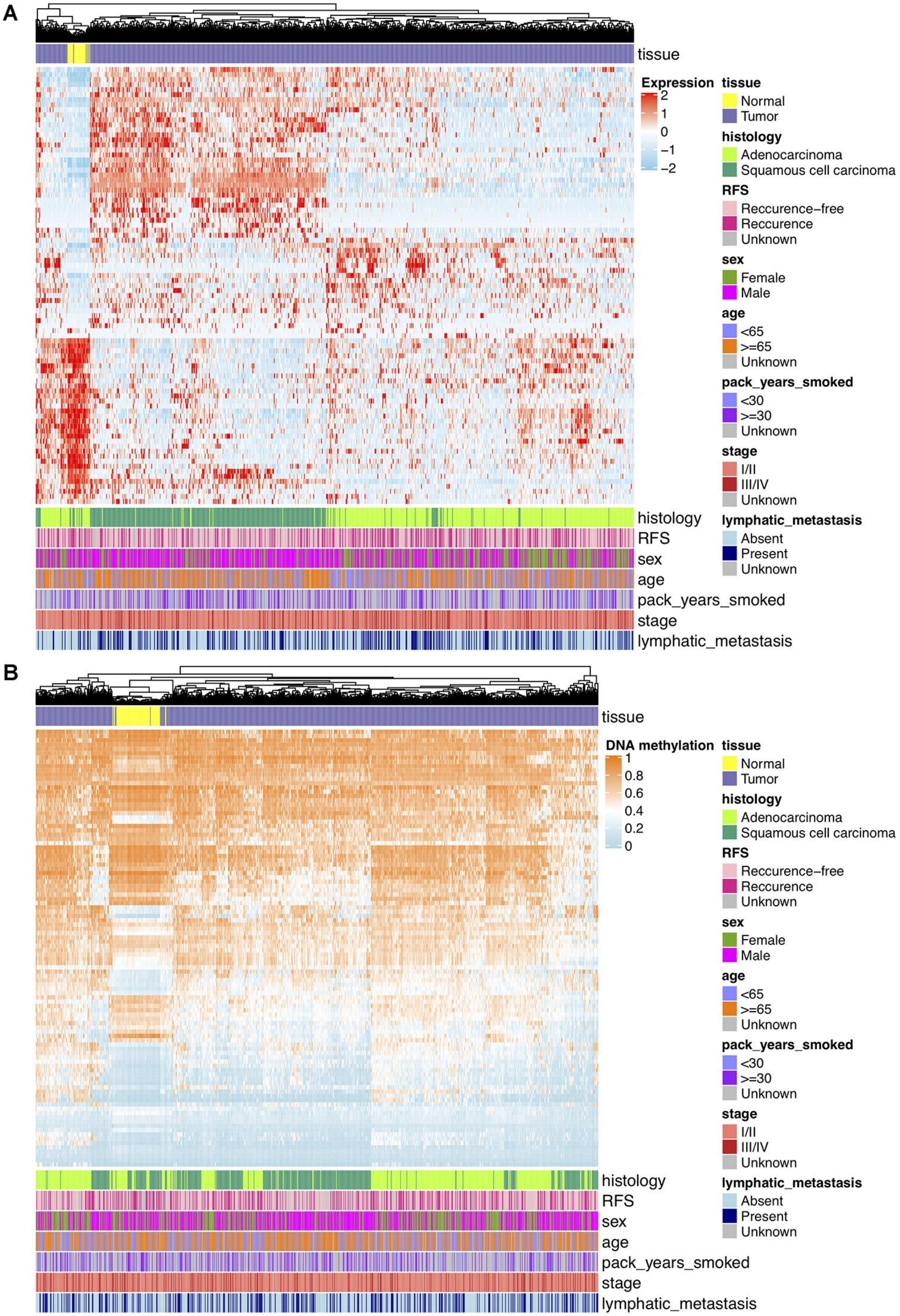


Figure 2. (**A**) Hierarchical clustering of 87 unique DEGs that potentially regulated by changes in DNAm levels at 102 selected loci based on TCGA NSCLC gene expression data (821 tumor and 28 normal samples). Clinical and demographic features, including age, sex, pack-years smoked, histology, stage, lymphatic metastasis and clinical outcome (RFS, recurrence-free survival status). High expression, red; low expression, skyblue. (**B**) Hierarchical clustering of 102 significantly DMPs between NSCLC (n=827) and normal (n=74) samples. Hopomethylated CpGs, skyblue; hypermethylated CpGs, orange.

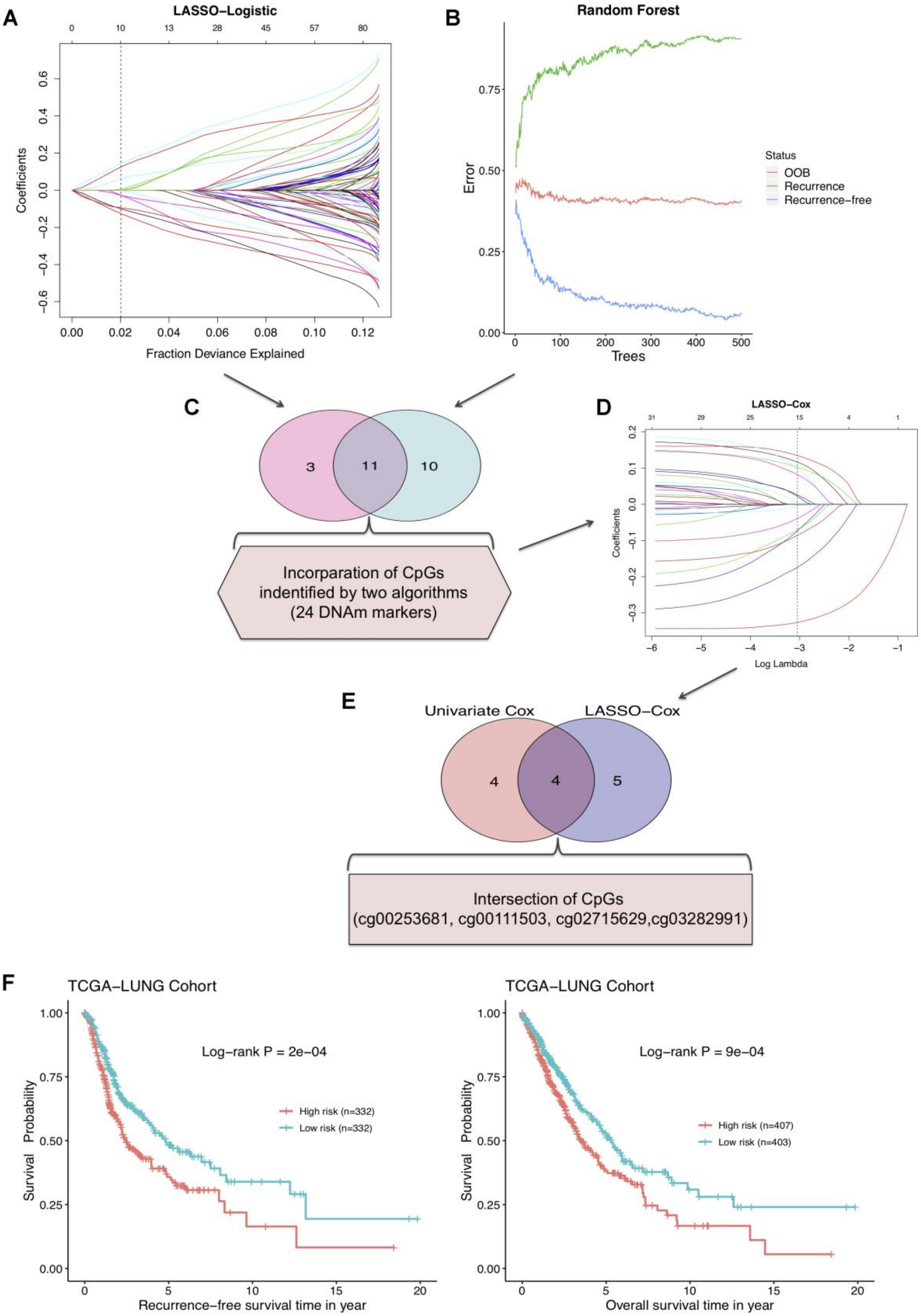


Figure 3. (**A**) LASSO-Logistic and (**B**) Random Forest methods applied to identify recurrence associated DNAm markers in training cohort. (**C**) A total of 11 overlapping CpGs and 24 combined CpGs in two algorithms. (**D**) LASSO-Cox analysis performed to select robust relapse predictive CpGs. (**E**) Four final identified CpGs in the intersection of univariate Cox and LASSO-Cox results. (**F**) The RFS curve (left) and overall survival curve (right) of training cohort based on 4-DNAm-maker panel.

Table 1. Characteristics of four prognostic CpG markers and their coefficients on the basis of multivariate Cox proportional hazard model.

Marker	Ref Gene	Coefficient	Hazard Ratio
cg00253681	ART4	0.339	1.4
cg00111503	KCNK9	-0.337	0.71
cg02715629	FAM83A	-0.138	0.87
cg03282991	C6orf10	0.037	1.04

Moreover, according to results of multivariable survival analysis, the risk score was an independent prognostic indicator for NSCLC patients and significantly associated with both RFS and OS (Supplementary Figure 3, Supplementary Table 4). AJCC stage system was extensively used for clinical prognostic evaluation, whereas this DNAm-based risk score was found to show better discriminative power of relapse status than clinical stage and other four separate CpGs in training and validation sets (Supplementary Figure 4).

Clinical and molecular features, and mutations associated with the prognostic DNA methylation signature

We next assessed the correlation of the DNAm signature with clinical characteristics and molecular features. The association between the risk score and tumor progression was found to be positive in TCGA NSCLC patients. Risk scores of tumors with recurrence were significantly higher than those without, and TCGA NSCLC samples at different stages had significant different relapse risk (Figure 4A). Besides, the similar results were also obtained on GSE39279 and GSE66836 datasets (Figure 4A). Then, we sought to investigate the molecular implications hidden behind current correlations. GSEA on two TCGA NSCLC sample subgroups divided by predictive recurrence risk revealed that high-risk group endowed significant enrichment of gene signatures mainly related to E2F targets, G2M checkpoint and MYC targets V1 (Figure 4B). According to RPPA analysis of LUAD tumors based upon TCGA repository, higher risk score was significantly correlated with higher expression of FOXM1 and CYCLINB1 proteins. FOXM1, a transcription factor upstream of CYCLINB1 (a G2/M transition marker), has been reported to associate with cell proliferation, whose overexpression is related to metastasis and poor prognosis in ovarian [17] and clear cell renal carcinoma [18], which suggested high cellular cycling and cell proliferation in NSCLC patients with high recurrence risk score. Also, epigenetic treatment of combining DNA-demethylating agents with histone deacetylase inhibitors decreased MYC signaling, exerting robust anti-tumor effect in NSCLC [19]. Notably, prominent correlations with risk score were also observed for pathways including allograft rejection, inflammatory response and IL2-STAT5 signaling (Figure 4B). Tumor microenvironment (TME) is deemed as intricate and dynamic in exacerbating and inhibiting tumor cell proliferation, migration, invasion and metastasis. On account that salient results of pathway enrichment analysis on the DNAm signature primarily consisted in cell cycle, proliferation and immune-related pathways, we were then committed to evaluating DNAm-based risk score in the context of TME.

Within TME, component system activation plays an important role in the connection of inflammation and anti-tumor immune response as well as oncogenesis [20]. Prior studies have proposed that epigenetic changes can regulate inflammation and immune signaling [19]. We thus tried to assess cellular composing in NSCLC based on DNAm profiles of TCGA and GSE66836 cohorts. After implementing HEpiDISH function, we found escalating risk score was associated with an increased fibroblasts and a reduced immune-cell fraction. In Figure 4C presented significant correlations between the DNAm signature and estimated compositions of B-cells, CD4+ T-cells, CD8+ T-cells, eosinophils, monocytes, and fibroblasts in TCGA NSCLC samples. There were no significant correlations for recurrence risk score with assessed enrichments of neutrophils and NK cells (data not shown). We also noted that the fraction of fibroblasts was associated with lymphatic metastasis (Figure 4C), suggesting fibroblasts in TME might induce tumor inflammation and boost NSCLC progression as well as metastasis [21]. Likewise, strong associations of B-cells, CD4+ T-cells, monocytes, eosinophils and fibroblasts abundance with DNAm-based risk score were observed in GSE66836 cohort (Supplementary Figure 5). These findings indicated that the higher fibroblasts fraction and lower infiltrating levels of immune cells might emerge in NSCLC patients with high recurrence risk, which were also consistent with previous reports demonstrating significant association between low density of CD8 +T-cells and poor RFS in papillary thyroid cancer [22].

Subsequently, we linked the DNAm signature to mutations in genes. Based on mutation profiles of TCGA NSCLC tumors, several SMGs identified by

MutsigCV v.1.41 and also correlated with DNAm-based risk score were presented in Figure 4D. A notable correlation of our risk score with somatic mutations in gene KRAS, KEAP1, STK11, and co-occurring KRAS/KEP1A mutations was found in NSCLC (Figure 4E). Furthermore, we found a significant association of this DNAm signature with mRNA expression of KEP1A and STK11 (Figure 4F). Higher NRF2, ACC1 and lower KEAP1, PTEN, p-AMPK, NF2 and RB protein abundance were observed in high-risk group (Figure 4F). Loss-of-function type mutation in KEAP1 results in activation of NRF2, which accelerates lung cancer cell growth [23]. Combined loss of PTEN and KEAP1 promotes LUAD formation in mice model [24]. In current study, high DNAm-based risk score was connected to low expression of STK11, low AMPK activation as well as STK11 mutation, indicating the mTOR activation of TCGA NSCLC samples in high-risk group [25]. Elevated ACC1 levels in patients with hepatocellular carcinoma were correlated to vascular invasion and disease recurrence [26]. The inactivation of NF2 in malignant Pleural Mesothelioma with mTOR activity aberrantly upregulated, fails to inhibit cell proliferation, leading to a poor prognosis [27]. Absence of RB in a mouse model of LUAD was demonstrated to drive disease progression and metastasis [28].

Methylation signature correlates to TMB and DDR genes

In recent time, tumor mutation burden (TMB) has been well-reported as an emerging and promising indicator for clinical benefit of immunotherapy [29]. DNA hypomethylation was proposed to induce chromosomal aberrations, prompting chromosome instability [30]. Based upon nonsynonymous coding mutations in somatic mutation data matched with methylome data of TCGA NSCLC tumors, we then quested for the connection between DNAm signature and TMB. It turned out to be noteworthy higher TMB in high-risk group (Figure 5A), suggesting DNAm-based risk score might predict immune evasion of NSCLC to some extent. Methylation loss was linked to an increase in mutation density and cell cycle gene expression by mitotic cell division [31]. Methylation status of four selected DMPs and expression levels of nearby reference genes at four CpG sites in our risk score model were also correlated with TMB (Supplementary Figure 6). We further inspected the underlying biological structure with respect to our DNAm signature to inquire into interpretation for distinct TMB between high- and low-risk groups. GSVA were performed on recurrence risk status, which revealed that gene sets of cell cycle, DNA replication, homologous recombination, nonhomologous end-joining, mismatch repair, base excision repair and nucleotide excision repair were significantly upregulated in high-risk group (Figure 5A). All these observations implied relevance of cell proliferative processes activation for high risk status, and also implicated that this DNAm signature might have connection to alterations in cell cycle, DNA replication and DDR pathway genes. Increased TMB and improved efficacy of ICBs were proposed to independently associate with alterations in DDR genes [32]. In order to determine whether DNAm signature might underlie and account for differential TMB by influencing molecular mechanisms analyzed above, we subsequently aimed at figuring out the correlation of DDR-related genes and DNAm-based risk score.

The related gene signatures involved in six of aforementioned DDR pathways were listed in Table 2 and defined as BER, NER, HRR, MMR, NHEJ and Checkpoint genes [33], respectively. Also, six kinds of co-mutations in DDR genes above were analyzed in this study. It has been well-established that TP53 mutation accelerates cell cycle and DNA replication, and TP53/KRAS co-mutation exhibits a remarkable increased mutational loads and owns a potential of predicting response to immunotherapy in LUAD [34]. In this study, higher frequency of TP53 mutation in high-risk group was observed in patients from TCGA NSCLC as well as GSE66836 cohort (Figure 5B), indicating NSCLC tumors with higher risk score were more susceptible to yielding DNA replication errors. Additionally, the presence of TP53 alterations without EGFR or STK11 mutations is revealed to identify LUAD patients who respond to anti-PD1 therapies [35]. Significant associations of risk scores with TP53/KRAS co-mutation, combination of TP53-Mut/EGFR-Wt mutation as well as somatic mutations in STK11 and in other co-analyzed DDR genes were also identified in this study (Figure 5B). The mTORC1-S6K pathway is aberrantly activated due to STK11 loss, leading to DNA damage response defects and prompting genome instability [36]. The mutated ATM is reported to correlate with genomic instability and ATM predominantly responds to DNA double-strand breaks (DSB) [37, 38]. What’s more, we found significant associations of the DNAm signature with waning protein abundance of ATM, PARP1, and raised expression of PCNA and CYCLINE1 protein (Figure 5C). Overexpressed CCNE1, an oncogene encoding cell cycle protein CYCLINE, induced DNA replication stress by premature S phase entry, leading to genomic instability [39]. All analyses above suggested that changes of DNAm patterns in NSCLC might predispose epigenetically impacting on TMB by mediating alteration in cell-cycle regulating and DDR genes, resulting in more neoantigens formation and changes of the tumor antigenicity.

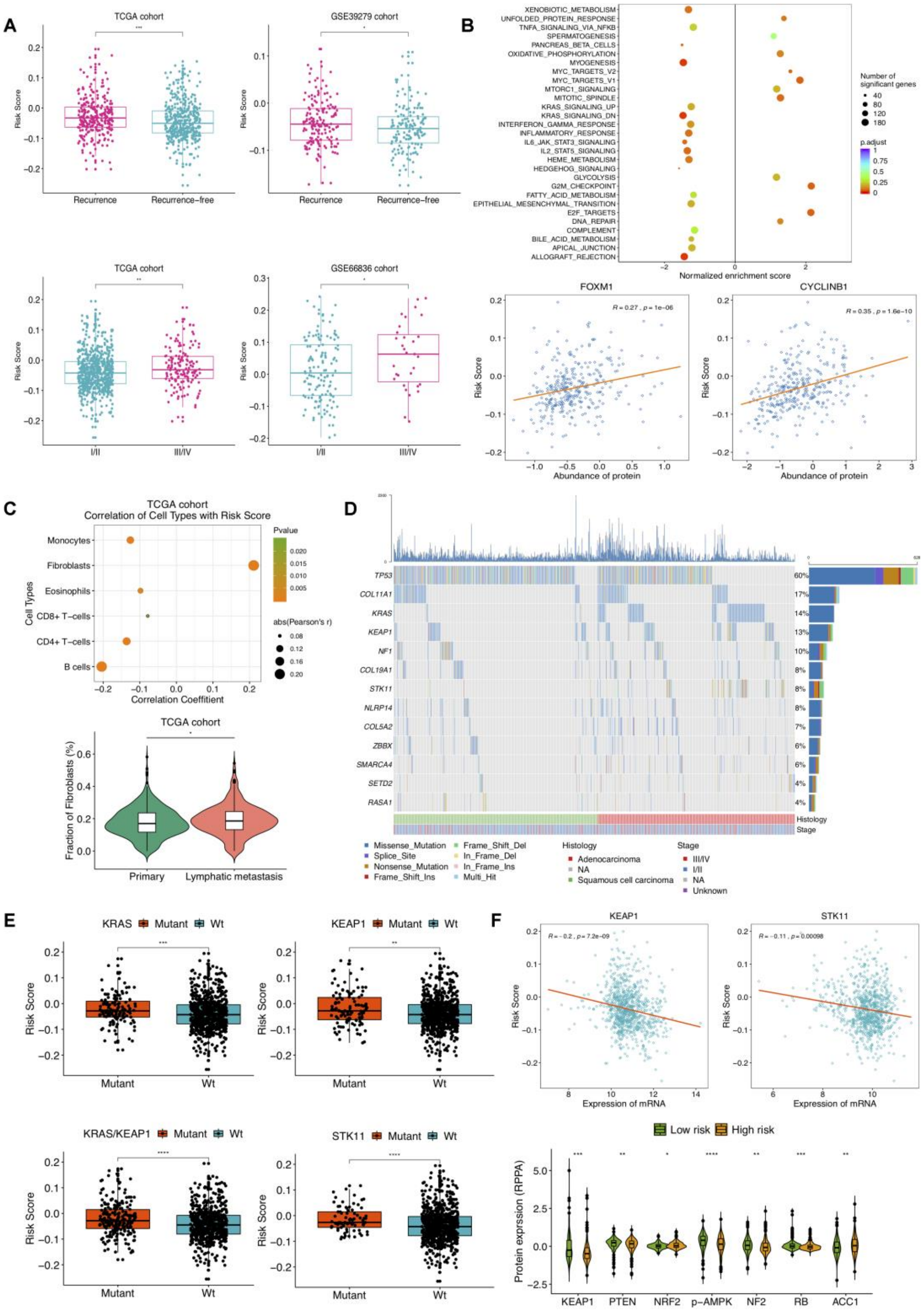


Figure 4. Selected clinical, molecular features and mutations associated with DNAm-based risk score. (**A**) Relationship of clinical characteristics and the DNAm signature. DNAm-based risk score stratified by different stages and recurrence status from TCGA NSCLC patients (left) and from GSE39279 (top right) and GSE66836 cohorts (bottom right). (**B**) GSEA on a set of hallmark gene signatures revealing the impact of the identified DNAm signature on cell cycle, proliferation and immune-related pathways (top); DNAm-based risk score strongly correlated to expression of FOXM1 and CYCLINB1 protein (bottom). (**C**) Relevance of estimated cell-type fractions with risk score (top); The abundance of fibroblasts in NSCLC patients relates to lymphatic metastasis status (bottom). (**D**) Mutation profile of TCGA NSCLC samples showing 13 SMGs of which mutational proportion correlated with DNAm signature. (**E**) Association of the DNAm signature with mutation in genes. DNAm-based risk score stratified by mutations in KRAS, KEAP1, STK11 and KRAS/KEAP1 co-mutations. (**F**) Correlation of DNAm signature with representative gene (top) and protein (bottom) expression.

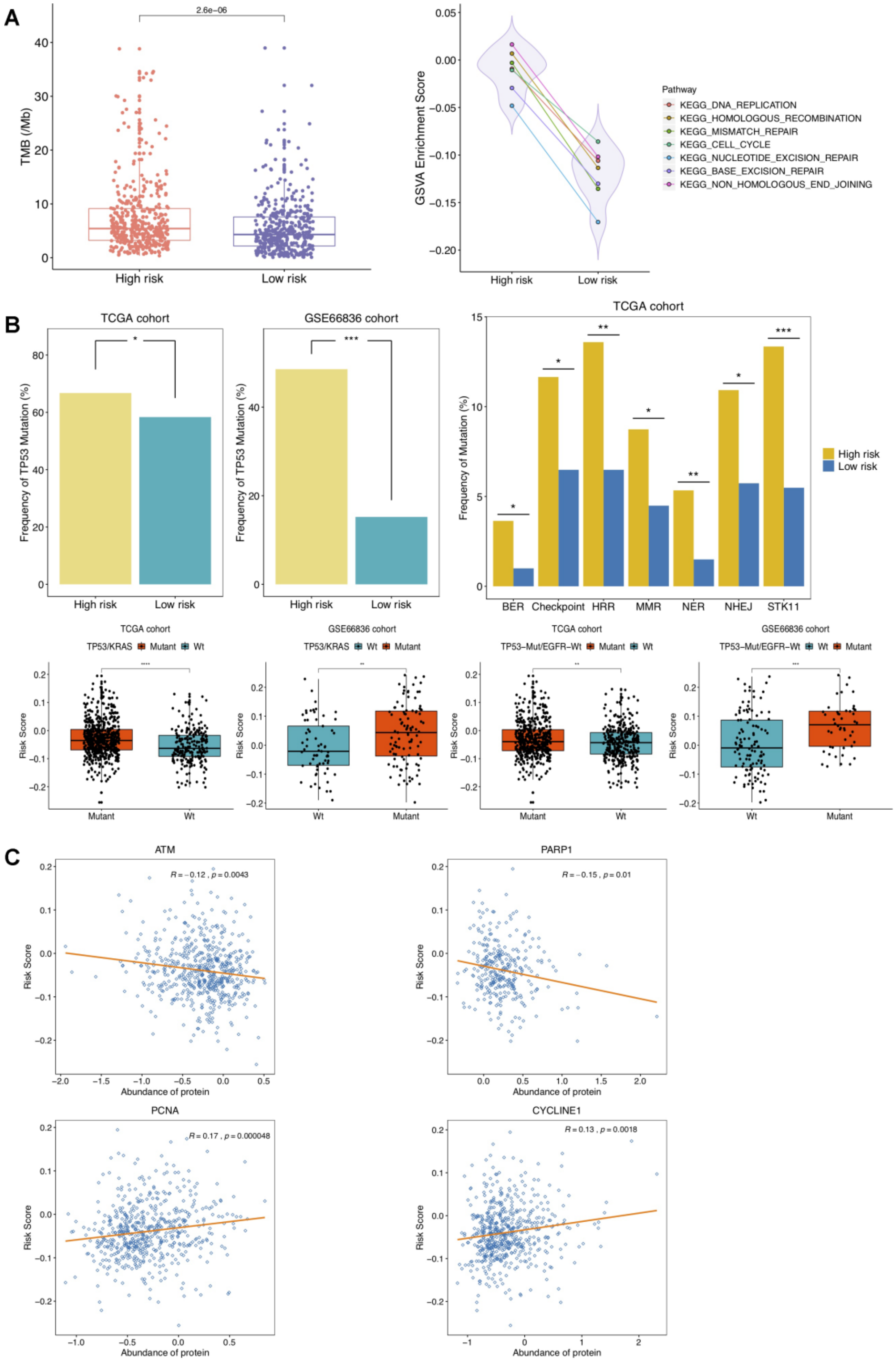


Figure 5. Correlation of DNAm-based risk score with TMB, cell cycle, DNA damage response (DDR) genes. (**A**) Left, TMB estimation of TCGA NSCLC patients in high- and low-risk group; right, GSVA presenting DDR pathways significantly enriched in high risk group (adjusted P < 0.01). (**B**) Estimated frequencies of mutations in TP53 (top left), STK11 and DDR genes (top right), TP53/KRAS co-mutations and TP53-Mut/EGFR-Wt mutations (bottom) under different recurrence risk status. (**C**) Protein expression of ATM, PARP1, PCNA and CYCLINE1 associated with DNAm-based risk score in TCGA cohort.

Table 2. Gene list for six related DNA damage repair response pathways.

Pathway	Genes
BER	NEIL3, PARP1, PCNA
Checkpoint	ATM, TIMELESS, TP53
HRR	BRCA1, BRCA2, XRCC2
MMR	MLH1, MSH3, MSH4
NER	ERCC1, ERCC6, TCEB3
NHEJ	DCLRE1C, PRKDC

*BER, base excision repair; HRR, homologous recombination repair; MMR, mismatch repair; NER, nucleotide excision repair; NHEJ, non-homologous end-joining.

NSCLC patients with high risk score present favorable clinical benefit to immunotherapy

To go further, the relevance of DNAm-based risk score with response to ICBs was investigated on GSE119144 cohort. We observed RFS of immunotherapeutic patients in high-risk group was significantly superior to that of those in low-risk group (Figure 6A). Besides, a greater percentage of NSCLC patients with high risk score owned a durable clinical benefit, while most of low-risk group patients had no durable clinical benefit (Figure 6B). Notably, the combination of DNAm signature and TMB saliently improved the ability to predict clinical responses to immunotherapy (AUC = 0.965, Figure 6C). As was revealed in Kaplan–Meier curves, NSCLC patients separated by this combination of two variables harbored significantly different clinical outcomes (P = 0.01, Figure 6D). These results

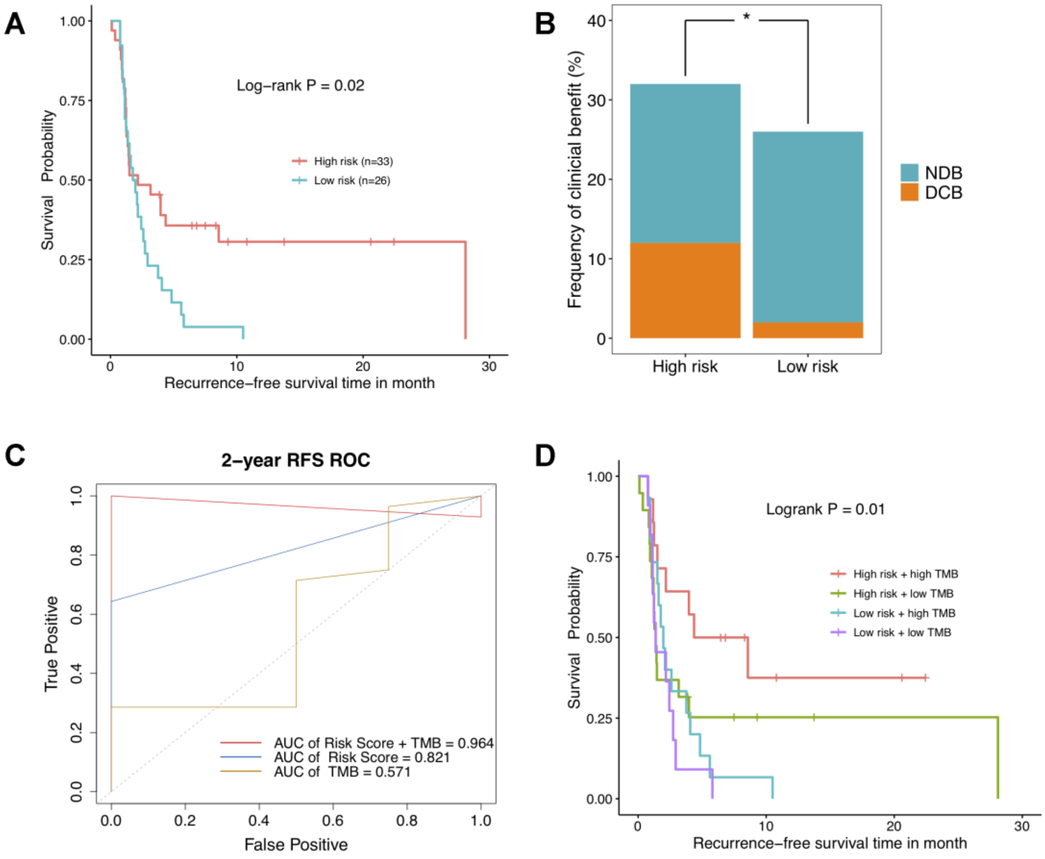


Figure 6. The relationship between DNA methylation signature and clinical response to immunotherapy investigated in GSE119144 cohort. (**A**) Relapse-free survival curves comparing high-risk with low-risk groups in NSCLC patients received anti-PD-1/PD-L1 therapies, according to DNAm-based risk score from validation set. (**B**) Proportion of clinical benefit to immunotherapy in the indicated groups stratified by our DNAm signature (DCB: durable clinical benefit and NDB: no durable benefit). (**C**) Time-dependent ROC curves for DNAm-based risk score, TMB, and risk score combined with TMB. (**D**) RFS curves of NSCLC patients with combinations of risk score and TMB.

implicated high risk score and high mutational burden might represent robust biomarkers to determine best responders to ICBs treatments. And anyway, the clinical application effect deserves further research and being corroborated in larger cohorts with longer follow-up data.

DISCUSSION

In present study, based on methylation and transcriptome data as well as corresponding clinical information of training and validation cohort, we initially selected 4 DMPs predictive of NSCLC relapse by applying machine learning methods, and built a risk score model comprised of 4 CpG markers. To demonstrate clinical utility of the 4-DNAm-marker panel, training and validation NSCLC cohorts could be effectively divided into high-risk and low-risk groups of tumor relapse. In this manner, it's convenient and beneficial for clinicians to conduct individualized medical treatment and heath management.

Given that higher DNAm-based risk score linked to adverse clinical outcomes, we assumed our DNAm signature might underlie and facilitate development and progression in NSCLC. To better understand implications of DNAm signatures in clinical events of NSCLC, we then sought to explore the biological function, molecular mechanism as well as hidden compounding somatic variants, and also evaluated cell-type composition from an epigenetic perspective. In this work, we observed the recurrence predictive risk score of TCGA NSCLC patients to be correlated with clinical characteristics, fractions of immune cell infiltrates, molecular features in the layers of both mRNA and protein expression, as well as somatic mutations in genes involved in specific signaling pathways. Interestingly, a remarkable association between the DNAm signature and TMB was identified in TCGA NSCLC cohort. Additionally, we noted DNAm-based risk score was connected to several DDR genes, which could be a favorable interpretation of high TMB in high-risk group. Ultimately, this DNAm signature was demonstrated to be a potential biomarker that predicted clinical response to immunotherapy and survival of NSCLC.

This is, to the best of our knowledge, a comprehensive investigation on NSCLC by integrating TCGA NSCLC data for DNAm, mutations, clinical features, expression of mRNAs and proteins, which provides insight into molecular mechanism, prognostic, and therapeutic implications. We propose that the 4-DNAm-marker panel acts as an effective biomarker for predicting metastasis and recurrence in NSCLC. The hematogenic metastasis formation can be characterized by the extravasation of leukocytes and tumor cells [40]. In the context of TME, it's of necessity to investigate individual cell-type enrichments for reflecting tumor-immune interactions. Prior studies have established the notion that CD4+ /CD8+ T-cells are capable of recognizing cancer antigens and positively associated with favorable RFS in ovarian cancer [41]. We noted that higher risk score was correlated to increased fibroblasts and reduced leukocyte fractions, indicating 4-DNAm-marker panel might have the potential to serve as an indicator of characterizing immune infiltration landscape in NSCLC. The combination epigenetic therapy (Aza + ITF-2357) induced the increased levels of CCL5, a secreted chemokine attracting functional lymphocytes, restraining tumor growth with increased number of CD8+ T cells in mice model of NSCLC [19]. Our observations also suggested that several immune cell subtypes in NSCLC were indispensable for tumor progression.

Previous studies put forward that genetic characterization revealed by WES or targeted sequencing might have influence over therapeutic options, assessment of treatment response and patients' prognosis in some solid cancers [42]. Investigations on targeted therapies directed towards several somatically altered pathways are thus essential for medical decision and clinical implementation. We presented that mutational patterns of TCGA NSCLC tumors were associated with the DNAm signature, in which a strong correlation of higher DNAm-based risk score with recurrently mutated driver genes KRAS, KEAP1 and STK11 was observed in this study. We also noted that high risk score was correlated to co-mutations in KRAS/KEAP1, and associated with expression of KEAP1, STK11 mRNA as well as several other key protein abundance. KEAP1 deletion contributes to tumor aggressiveness, metastasis, and increases radio-resistance in LUSC [43]. Somatic oncogenic point mutations in KRAS were proposed to be crucial to progression and drug resistance in 90% of patients with pancreatic ductal adenocarcinoma, a highly metastatic disease with a high mortality rate [44]. A higher incidence of metastasis emerged in KRAS-mutated CRC patients, whose relapse pattern depends on the KRAS mutational status with down-regulation of p-MAPK signaling prompting and forming distant lung metastasis [45]. Somatic KEAP1 mutation leads to activation of the NRF2 pathway, and NSCLC patients with KEAP1 mutation in addition to an activating KRAS mutation were demonstrated to have a shorter duration of platinum-based chemotherapy and a worse prognosis than other patients with KRAS-mutant [46]. We speculated that epigenetic patterns might describe genetic alterations in NSCLC tumors and further reflect tumor aggressiveness and resistance to therapy. Our identified DNAm signature of TCGA NSCLC tumors were also observed to connect with recurrence-associated and especially cell-cycle related gene signatures that induce tumor metastasis for many tumor types, suggesting that epigenetic changes may impact on activation of cell cycle and disease progression.

The unmethylated status of FOXP1 indicates durable response to ICB therapies and improved survival of a subset of NSCLC patients, and could correlate to validated and up-to-date biomarkers such as mutational load [29]. Earlier studies proposed that DNA replication stress leads to genomic instability in cancer, which can be characterized by the rates of high mutation as well as epigenetic perturbation, especially for DNA methylation loss [15, 47]. TMB of TCGA NSCLC patients was found prominently associated with the DNAm-based risk score in our study. We also revealed underlying mechanisms hidden behind this correlation. It turned out that DNAm signature also interplayed with genomic alterations in DDR pathways, implicating NSCLC patients in high-risk group potentially endowed endogenous replication stress and genomic instability. In the light of results analyzed above, we speculate that DNAm-based risk score may contribute to the identification of those individuals who will be more susceptible to immunotherapy and predicting clinical efficacy of ICBs. More recently, several studies have proposed DNAm alterations might work as biomarkers of immune evasion with higher predictive power than mutation burden [48], but the more specific DNAm signature remained to be elucidated. The sensitive measure of TMB estimation demands WES or a minimum gene panel size of 150 [49]. Clinically, our 4-DNAm-maker panel, by contrast, can avoid the high cost of WES or deep sequencing and be cost-efficient in practice. Combining epigenetic treatment, depletion of Myc reverses immune invasion of lung cancer, enhancing effectiveness of immune checkpoint treatment [19]. It's conceivable that a promising epigenetic therapy, possibly coupled with immunotherapies will generate remarkable clinical benefit.

Several limitations to our study should be noted: Firstly, another independent NSCLC cohort which is performed with DNAm assay and also received ICBs therapy with detailed follow-up data will be required to validate our observation. Second, this study was based on DNAm data of NSCLC tissue biopsy and it's preferable that our results could be also validated by ctDNA methylation analysis which is relatively noninvasive and clinically feasible. Thirdly, more detailed biologic mechanisms of final selected markers remains to be investigated on laboratory experiments.

To sum up, the combined DNAm signature (cg00253681, cg00111503, cg02715629, and cg03282991) is a reliable biomarker for predicting clinical benefit to ICBs treatments and recurrence in NSCLC. Our study shed light on the implications of epigenetic modulation in disease recurrence prediction, treatment strategy selection and evaluation of responses to immunotherapy.

MATERIALS AND METHODS

NSCLC patient data

The Cancer Genome Atlas (TCGA) LUNG Cancer, GSE66836, GSE39279 and GSE119144 cohorts included in this study were derived from online public data repository, with NSCLC patients who received neo-adjuvant chemotherapy excluded. For DNAm data, a total of 901 TCGA NSCLC samples were available using the Illumina Infinium HumanMethylation450 platform, including 827 tumor tissues and 74 non-tumor tissues. DNAm level 3 data were obtained at the website: https://tcga.xenahubs.net. In addition to TCGA data, we also analyzed GSE66836 dataset (164 LUADs, 19 normal lungs), GSE39279 and GSE119144 cohorts that recruited 444 and 60 NSCLC patients respectively. Three methylation microarray datasets and corresponding clinical data were downloaded from Gene Expression Omnibus (GEO) database. DNAm levels of 4 datasets above were all measured by beta values for each CpG probe, which ranged from 0 (completely unmethylated) to 1 (completely methylated). TCGA LUNG Cancer gene expression data consisted of 1116 samples, including 1007 NSCLC tissues and 109 normal tissues. RNA sequencing (RNA-seq) level 3 expression data (normalized read counts) and related clinical data were available at https://tcga.xenahubs.net. TCGA NSCLC somatic mutation data comprised of somatic variant calls in TCGA-LUAD (n=562) and TCGA-LUSC (n=486) cohorts were retrieved from https://gdc.xenahubs.net.

Analysis of epigenetic profiles

Right at the beginning, we aimed at identifying significant differentially methylated positions (DMPs) in TCGA NSCLC tumor versus normal lung tissues. TCGA DNAm level 3 data for 485578 CpGs were parsed into the limma package [50] with limma function implemented to assess the differential DNAm. A robust DMP was defined as containing CpG that yielded a Benjamini-Hochberg (BH) adjusted $P < 0.05$, without "NA" for the average beta in each group. To obtain more reliable DMPs, the same analysis steps were conducted on Methylation 450K Beadchip data (normalized beta values) of GSE66836, then overlapping DMPs of 2 datasets were retained for subsequent analyses. The DMPs with $\log_2$ fold change < 0 were regarded as hypomethylated and > 0 were regarded as hypermethylated. To pursuit meaningful epigenetic profiling, 450k data in our study were annotated by R package lluminaHumanMethylation450kanno.ilmn 12.hg19.

Gene expression data and reverse phase protein array profiling

Subsequently, analysis of aberrant gene expression in 821 TCGA NSCLC tissues compared with 28 matched normal lung tissues (consistent with samples in DNAm data) was performed by edgeR package. A threshold value of false discovery rate (FDR) < 0.05 and $|\log_2$ fold change$|$ > 1 was used for screening significant differentially expressed genes (DEGs). The upregulated genes (FDR < 0.05, $\log_2$ fold change >1) and downregulated genes (FDR < 0.05, $\log_2$ fold change < -1) were further utilized for screening hypomethylated CpGs showing higher expression and hypermethylated CpGs showing lower expression, respectively. To integrate mRNA expression and epigenetic profile, the associations between DEGs and differentially methylated loci located within 2 kb of transcript start site were then assessed by Spearman correlation. The Bonferroni corrected P < 0.05 and r < 0 were used as cut-off criteria for further filtration.

As is described in previous researches, it's preferable that we performed analysis of mRNA expression incorporated with protein expression profiles for reflecting biological complexity more systematically and comprehensively [51]. To assess protein levels of TCGA cohort, we downloaded the reverse phase protein array (RPPA) profiles of 687 NSCLC samples (including 362 LUAD and 325 LUSC samples) from MD Anderson (http://app1.bioinformatics.mdanderson.org/tcpa/_design/basic/index.html).

Clinical data and predictive modeling

To identify potential recurrence predictive methylation markers for NSCLC patients, four aforementioned cohorts were included for prognostic prediction using candidate CpGs in combination with corresponding clinical characteristics (adjuvant radiation and chemotherapy, histology, sex, age, stage, pack-years smoked, lymphatic and distant metastasis).

In training phase, TCGA LUNG Cancer cohort was used as the training dataset, in which 664 NSCLC specimens have detailed recurrence-free survival (RFS) information. First, Random Forest and Least Absolute Shrinkage and Selection Operator (LASSO) Logistic models were applied to select recurrence-related DNAm markers. Meanwhile, univariate cox regression was performed to identify CpG markers associated with RFS of NSCLC patients. According to theoretical basis of LASSO method, Cox proportional hazards regression model with LASSO penalty generates regression coefficients that strictly equal to 0, removing some of variables with lower weights on the purpose of data dimension reduction, and preventing overfitting resulting from collinearity of the covariates [52]. To further identify more significant markers, LASSO-Cox method was subsequently implemented on the incorporation of CpGs selected by LASSO-Logistic and Random Forest algorithms. Next, the consistent CpGs identified by univariate Cox and LASSO-Cox methods were retained and then incorporated into a multivariable Cox regression model. Eventually, GSE39279, GSE66836 and GSE119144 were taken as validation datasets, in which GSE119144 cohort was consisted of 60 NSCLC patients who underwent anti-PD-1/PD-L1 ICBs treatments, with complete follow-up information available about 59 samples.

Gene set enrichment analysis and gene set variation analysis

To investigate the association between NSCLC recurrence risk status predicted by the DNAm-based risk score and gene signatures, using R package "clusterprofiler" [53] we implemented gene set enrichment analysis (GSEA) on TCGA mRNA expression data. Two NSCLC subgroups (High risk vs. Low risk) were divided according to median risk score, on which fold change values were calculated and used for GSEA on a set of 50 hallmark signatures [54]. Additionally, the "GSVA" package [55] was utilized for identifying pathways most related to the DNAm signature. The gene set variation analysis (GSVA) was performed with a set of 186 KEGG pathway signatures. Gene signatures with adjusted P < 0.05 were considered significant differentially enriched. Two reference gene sets above were downloaded from the Molecular Signature Database (MSigDB): http://software.broadinstitute.org/gsea/msigdb/index.jsp.

Somatic mutation profiling and cell-type fraction estimating

Highly confident somatic variants, including single nucleotide variation and short insertion/deletion polymorphism, from a total of 1048 TCGA NSCLC samples were integrated in this study. Using MutsigCV v.1.41 [56] somatic variant calls from TCGA NSCLC tumors were analyzed for significantly mutated genes (SMGs); among them, SMGs with FDR value below 0.05 were retained. The visualized part and summarization for MAF files of TCGA NSCLC whole-exome sequencing (WES) data were implemented by R package maftools [57]. To evaluate the tumor mutation burden (TMB), we computed the number of non-synonymous somatic mutations in coding region for each tumor sample. Based on a primary reference containing fibroblasts and a secondary reference that contains 7 immune cells subtypes (B-cells, NK cells, CD4+ and CD8+ T-cells, monocytes, neutrophils and eosinophils) [58], we applied the HEpiDISH method on DNAm profiles of TCGA and GSE66836 datasets by R package "EpiDISH" to infer individual cell-type fractions for NSCLC patients included. The correlations between the DNAm signature and estimated enrichments of cell types were subsequently investigated.

Statistical analyses

All statistical analyses were implemented using R version 3.6.2. Unsupervised hierarchical clustering was conducted by package ComplexHeatmap in R using the DNA methylation levels of selected DMPs as well as the mRNA expression levels of DEGs nearby those CpGs. The R package "randomForest" and "glmnet" were used for Random Forest and LASSO model. The 10-fold cross validation was performed on two algorithms to optimize Random Forest model with minimum misclassification rate and obtain the optimal lambda values (the minimum lambda value) in LASSO models. Kaplan-Meier curves analyses and log-rank tests were performed by the survminer package. Furthermore, the survival package was used for survival analysis with DNAm signature and clinicopathological parameters combined in a multivariable Cox proportional hazards regression model. To estimate the performance of CpG markers in training and validation sets, we conducted receiver operating characteristic (ROC) curve analyses using pROC package. In addition, time depended ROC analysis was performed by the survival ROC package. We performed the Wilcoxon test followed by multiple testing using the BH correction approach to figure out difference of DNAm-based risk score between mutant subgroups, between related clinical-factor groups, and difference of TMB estimation in high-risk vs. low-risk group. Spearman correlation analysis was used to assess the relationship of DNAm signature with estimated cell-type fractions, DNA damage response (DDR) genes and proteins. The mutation frequencies in DDR genes between high- and low-risk groups were compared using Chi-square ($\chi2$) test. For all statistical tests, two-tailed $P < 0.05$ denoted statistical significance, which is indicated by *, $P < 0.05$, **, $P < 0.01$, ***, $P < 0.001$, ****, $P < 0.0001$.

Data accessibility

The methylation chip data for NSCLC samples included in our study are accessible through GEO accession number GSE66836, GSE119144 and GSE39279. RPPA data are available on MD Anderson TCGA database. All TCGA NSCLC data can be accessed at https://tcga.xenahubs.net.

Abbreviations

BH: Benjamini-Hochberg; DDR: DNA damage response; DEGs: differentially expressed genes; DNAm: DNA methylation; DMPs: differentially methylated positions; FDR: false discovery rate; GEO: Gene Expression Omnibus; GSEA: gene set enrichment analysis; GSVA: gene set variation analysis; ICBs: immune checkpoint blockades; LASSO: Least Absolute Shrinkage and Selection Operator; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; NSCLC: non–small-cell lung carcinoma; OS: overall survival; RFS: recurrence-free survival; RNA-seq: RNA sequencing; RPPA: reverse phase protein array; ROC: receiver operating characteristic; SMGs: significantly mutated genes; TCGA: The Cancer Genome Atlas; TMB: tumor mutation burden; TME: tumor micro-environment; WES: whole-exome sequencing.

AUTHOR CONTRIBUTIONS

Ruihan Luo and Longke Ran conceived of and designed the project. Ruihan Luo conducted the data analysis, interpretation and manuscript writing. Longke Ran and Jing Song performed paper revision. All of the authors participated in the disussion and approved the manuscript.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to declare.

FUNDING

This work was supported by a grant from The Intelligence Medicine Project of Chongqing Medical University, China (No.ZHYX2019015 to Longke Ran).

REFERENCES

1. Fitzmaurice C, Abate D, Abbasi N, Abbastabar H, Abd-Allah F, Abdel-Rahman O, Abdelalim A, Abdoli A, Abdollahpour I, Abdulle AS, Abebe ND, Abraha HN, Abu-Raddad LJ, et al, and Global Burden of Disease Cancer Collaboration. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2017: a systematic analysis for the global burden of disease study. JAMA Oncol. 2019; 5:1749–68. https://doi.org/10.1001/jamaoncol.2019.2996 PMID:31560378

2. Morgensztern D, Ng SH, Gao F, Govindan R. Trends in stage distribution for patients with non-small cell lung cancer: a national cancer database survey. J Thorac Oncol. 2010; 5:29–33.
https://doi.org/10.1097/JTO.0b013e3181c5920c
PMID:19952801

3. Rotolo F, Dunant A, Le Chevalier T, Pignon JP, Arriagada R, and IALT Collaborative Group. Adjuvant cisplatin-based chemotherapy in nonsmall-cell lung cancer: new insights into the effect on failure type via a multistate approach. Ann Oncol. 2014; 25:2162–66.
https://doi.org/10.1093/annonc/mdu442
PMID:25193990

4. Pogrebniak KL, Curtis C. Harnessing tumor evolution to circumvent resistance. Trends Genet. 2018; 34:639–51.
https://doi.org/10.1016/j.tig.2018.05.007
PMID:29903534

5. Rotow J, Bivona TG. Understanding and targeting resistance mechanisms in NSCLC. Nat Rev Cancer. 2017; 17:637–58.
https://doi.org/10.1038/nrc.2017.84
PMID:29068003

6. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M, Ji H, Potash J, Sabunciyan S, Feinberg AP. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. 2009; 41:178–86.
https://doi.org/10.1038/ng.298
PMID:19151715

7. Esteller M. Epigenetics in cancer. N Engl J Med. 2008; 358:1148–59.
https://doi.org/10.1056/NEJMra072067
PMID:18337604

8. Ibrahim AE, Arends MJ, Silva AL, Wyllie AH, Greger L, Ito Y, Vowler SL, Huang TH, Tavaré S, Murrell A, Brenton JD. Sequential DNA methylation changes are associated with DNMT3B overexpression in colorectal neoplastic progression. Gut. 2011; 60:499–508.
https://doi.org/10.1136/gut.2010.223602
PMID:21068132

9. Xu RH, Wei W, Krawczyk M, Wang W, Luo H, Flagg K, Yi S, Shi W, Quan Q, Li K, Zheng L, Zhang H, Caughey BA, et al. Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. Nat Mater. 2017; 16:1155–61.
https://doi.org/10.1038/nmat4997
PMID:29035356

10. Tsai HC, Li H, Van Neste L, Cai Y, Robert C, Rassool FV, Shin JJ, Harbom KM, Beaty R, Pappou E, Harris J, Yen RW, Ahuja N, et al. Transient low doses of DNA-demethylating agents exert durable antitumor effects on hematological and epithelial tumor cells. Cancer Cell. 2012; 21:430–46.
https://doi.org/10.1016/j.ccr.2011.12.029
PMID:22439938

11. Church TR, Wandell M, Lofton-Day C, Mongin SJ, Burger M, Payne SR, Castaños-Vélez E, Blumenstein BA, Rösch T, Osborn N, Snover D, Day RW, Ransohoff DF, and PRESEPT Clinical Study Steering Committee, Investigators and Study Team. Prospective evaluation of methylated SEPT9 in plasma for detection of asymptomatic colorectal cancer. Gut. 2014; 63:317–25.
https://doi.org/10.1136/gutjnl-2012-304149
PMID:23408352

12. Amatu A, Barault L, Moutinho C, Cassingena A, Bencardino K, Ghezzi S, Palmeri L, Bonazzina E, Tosi F, Ricotta R, Cipani T, Crivori P, Gatto R, et al. Tumor MGMT promoter hypermethylation changes over time limit temozolomide efficacy in a phase II trial for metastatic colorectal cancer. Ann Oncol. 2016; 27:1062–67.
https://doi.org/10.1093/annonc/mdw071
PMID:26916096

13. Brock MV, Hooker CM, Ota-Machida E, Han Y, Guo M, Ames S, Glöckner S, Piantadosi S, Gabrielson E, Pridham G, Pelosky K, Belinsky SA, Yang SC, et al. DNA methylation markers and early recurrence in stage I lung cancer. N Engl J Med. 2008; 358:1118–28.
https://doi.org/10.1056/NEJMoa0706550
PMID:18337602

14. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, Lee W, Yuan J, Wong P, Ho TS, Miller ML, Rekhtman N, Moreira AL, et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science. 2015; 348:124–28.
https://doi.org/10.1126/science.aaa1348
PMID:25765070

15. Shipony Z, Mukamel Z, Cohen NM, Landan G, Chomsky E, Zeliger SR, Fried YC, Ainbinder E, Friedman N, Tanay A. Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. Nature. 2014; 513:115–19.
https://doi.org/10.1038/nature13458
PMID:25043040

16. Xu GL, Bestor TH, Bourc'his D, Hsieh CL, Tommerup N, Bugge M, Hulten M, Qu X, Russo JJ, Viegas-Péquignot E. Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. Nature. 1999; 402:187–91.
https://doi.org/10.1038/46052
PMID:10647011

17. Wen N, Wang Y, Wen L, Zhao SH, Ai ZH, Wang Y, Wu B, Lu HX, Yang H, Liu WC, Li Y. Overexpression of FOXM1

predicts poor prognosis and promotes cancer cell proliferation, migration and invasion in epithelial ovarian cancer. J Transl Med. 2014; 12:134. https://doi.org/10.1186/1479-5876-12-134 PMID:24885308

18. Song J, Song F, Liu K, Zhang W, Luo R, Tang Y, Ran L. Multi-omics analysis reveals epithelial-mesenchymal transition-related gene FOXM1 as a novel prognostic biomarker in clear cell renal carcinoma. Aging (Albany NY). 2019; 11:10316–37. https://doi.org/10.18632/aging.102459 PMID:31743108

19. Topper MJ, Vaz M, Chiappinelli KB, DeStefano Shields CE, Niknafs N, Yen RC, Wenzel A, Hicks J, Ballew M, Stone M, Tran PT, Zahnow CA, Hellmann MD, et al. Epigenetic therapy ties MYC depletion to reversing immune evasion and treating lung cancer. Cell. 2017; 171:1284–300.e21. https://doi.org/10.1016/j.cell.2017.10.022 PMID:29195073

20. Reis ES, Mastellos DC, Ricklin D, Mantovani A, Lambris JD. Complement in cancer: untangling an intricate relationship. Nat Rev Immunol. 2018; 18:5–18. https://doi.org/10.1038/nri.2017.97 PMID:28920587

21. Ershaid N, Sharon Y, Doron H, Raz Y, Shani O, Cohen N, Monteran L, Leider-Trejo L, Ben-Shmuel A, Yassin M, Gerlic M, Ben-Baruch A, Pasmanik-Chor M, et al. NLRP3 inflammasome in fibroblasts links tissue damage with inflammation in breast cancer progression and metastasis. Nat Commun. 2019; 10:4375. https://doi.org/10.1038/s41467-019-12370-8 PMID:31558756

22. Cunha LL, Marcello MA, Nonogaki S, Morari EC, Soares FA, Vassallo J, Ward LS. CD8+ tumour-infiltrating lymphocytes and COX2 expression may predict relapse in differentiated thyroid cancer. Clin Endocrinol (Oxf). 2015; 83:246–53. https://doi.org/10.1111/cen.12586 PMID:25130519

23. Ohta T, Iijima K, Miyamoto M, Nakahara I, Tanaka H, Ohtsuji M, Suzuki T, Kobayashi A, Yokota J, Sakiyama T, Shibata T, Yamamoto M, Hirohashi S. Loss of Keap1 function activates Nrf2 and provides advantages for lung cancer cell growth. Cancer Res. 2008; 68:1303–09. https://doi.org/10.1158/0008-5472.CAN-07-5003 PMID:18316592

24. Best SA, De Souza DP, Kersbergen A, Policheni AN, Dayalan S, Tull D, Rathi V, Gray DH, Ritchie ME, McConville MJ, Sutherland KD. Synergy between the KEAP1/NRF2 and PI3K pathways drives non-small-cell lung cancer with an altered immune microenvironment. Cell Metab. 2018; 27:935–43.e4. https://doi.org/10.1016/j.cmet.2018.02.006 PMID:29526543

25. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014; 511:543–50. https://doi.org/10.1038/nature13385 PMID:25079552

26. Wang MD, Wu H, Fu GB, Zhang HL, Zhou X, Tang L, Dong LW, Qin CJ, Huang S, Zhao LH, Zeng M, Wu MC, Yan HX, Wang HY. Acetyl-coenzyme a carboxylase alpha promotion of glucose-mediated fatty acid synthesis enhances survival of hepatocellular carcinoma in mice and patients. Hepatology. 2016; 63:1272–86. https://doi.org/10.1002/hep.28415 PMID:26698170

27. Hylebos M, Van Camp G, van Meerbeeck JP, Op de Beeck K. The genetic landscape of Malignant pleural mesothelioma: results from massively parallel sequencing. J Thorac Oncol. 2016; 11:1615–26. https://doi.org/10.1016/j.jtho.2016.05.020 PMID:27282309

28. Rubin SM, Sage J. Manipulating the tumour-suppressor protein rb in lung cancer reveals possible drug targets. Nature. 2019; 569:343–44. https://doi.org/10.1038/d41586-019-01319-y PMID:31076732

29. Duruisseaux M, Martínez-Cardús A, Calleja-Cervantes ME, Moran S, Castro de Moura M, Davalos V, Piñeyro D, Sanchez-Cespedes M, Girard N, Brevet M, Giroux-Leprieur E, Dumenil C, Pradotto M, et al. Epigenetic prediction of response to anti-PD-1 treatment in non-small-cell lung cancer: a multicentre, retrospective analysis. Lancet Respir Med. 2018; 6:771–81. https://doi.org/10.1016/S2213-2600(18)30284-4 PMID:30100403

30. Eden A, Gaudet F, Waghmare A, Jaenisch R. Chromosomal instability and tumors promoted by DNA hypomethylation. Science. 2003; 300:455. https://doi.org/10.1126/science.1083557 PMID:12702868

31. Zhou W, Dinh HQ, Ramjan Z, Weisenberger DJ, Nicolet CM, Shen H, Laird PW, Berman BP. DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat Genet. 2018; 50:591–602. https://doi.org/10.1038/s41588-018-0073-4 PMID:29610480

32. Wang Z, Zhao J, Wang G, Zhang F, Zhang Z, Zhang F, Zhang Y, Dong H, Zhao X, Duan J, Bai H, Tian Y, Wan R, et al. Comutations in DNA damage response pathways serve as potential biomarkers for immune checkpoint blockade. Cancer Res. 2018; 78:6486–96.

https://doi.org/10.1158/0008-5472.CAN-18-1814 PMID:30171052

33. Scarbrough PM, Weber RP, Iversen ES, Brhane Y, Amos CI, Kraft P, Hung RJ, Sellers TA, Witte JS, Pharoah P, Henderson BE, Gruber SB, Hunter DJ, et al. A cross-cancer genetic association analysis of the DNA repair and DNA damage signaling pathways for lung, ovary, prostate, breast, and colorectal cancer. Cancer Epidemiol Biomarkers Prev. 2016; 25:193–200. https://doi.org/10.1158/1055-9965.EPI-15-0649 PMID:26637267

34. Dong ZY, Zhong WZ, Zhang XC, Su J, Xie Z, Liu SY, Tu HY, Chen HJ, Sun YL, Zhou Q, Yang JJ, Yang XN, Lin JX, et al. Potential predictive value of TP53 and KRAS mutation status for response to PD-1 blockade immunotherapy in lung adenocarcinoma. Clin Cancer Res. 2017; 23:3012–24. https://doi.org/10.1158/1078-0432.CCR-16-2554 PMID:28039262

35. Biton J, Mansuet-Lupo A, Pécuchet N, Alifano M, Ouakrim H, Arrondeau J, Boudou-Rouquette P, Goldwasser F, Leroy K, Goc J, Wislez M, Germain C, Laurent-Puig P, et al. TP53, STK11, and EGFR mutations predict tumor immune profile and the response to anti-PD-1 in lung adenocarcinoma. Clin Cancer Res. 2018; 24:5710–23. https://doi.org/10.1158/1078-0432.CCR-18-0163 PMID:29764856

36. Xie X, Hu H, Tong X, Li L, Liu X, Chen M, Yuan H, Xie X, Li Q, Zhang Y, Ouyang H, Wei M, Huang J, et al. The mTOR-S6K pathway links growth signalling to DNA damage response by targeting RNF168. Nat Cell Biol. 2018; 20:320–31. https://doi.org/10.1038/s41556-017-0033-8 PMID:29403037

37. Rao Q, Liu M, Tian Y, Wu Z, Hao Y, Song L, Qin Z, Ding C, Wang HW, Wang J, Xu Y. cryo-EM structure of human ATR-ATRIP complex. Cell Res. 2018; 28:143–56. https://doi.org/10.1038/cr.2017.158 PMID:29271416

38. Caron P, Choudjaye J, Clouaire T, Bugler B, Daburon V, Aguirrebengoa M, Mangeat T, Iacovoni JS, Álvarez-Quilón A, Cortés-Ledesma F, Legube G. Non-redundant functions of ATM and DNA-PKcs in response to DNA double-strand breaks. Cell Rep. 2015; 13:1598–609. https://doi.org/10.1016/j.celrep.2015.10.024 PMID:26586426

39. Macheret M, Halazonetis TD. Intragenic origins due to short G1 phases underlie oncogene-induced DNA replication stress. Nature. 2018; 555:112–16. https://doi.org/10.1038/nature25507 PMID:29466339

40. Strell C, Entschladen F. Extravasation of leukocytes in comparison to tumor cells. Cell Commun Signal. 2008; 6:10. https://doi.org/10.1186/1478-811X-6-10 PMID:19055814

41. Pinto MP, Balmaceda C, Bravo ML, Kato S, Villarroel A, Owen GI, Roa JC, Cuello MA, Ibañez C. Patient inflammatory status and CD4+/CD8+ intraepithelial tumor lymphocyte infiltration are predictors of outcomes in high-grade serous ovarian cancer. Gynecol Oncol. 2018; 151:10–17. https://doi.org/10.1016/j.ygyno.2018.07.025 PMID:30078505

42. Fernandes MG, Jacob M, Martins N, Moura CS, Guimarães S, Reis JP, Justino A, Pina MJ, Cirnes L, Sousa C, Pinto J, Marques JA, Machado JC, et al. Targeted gene next-generation sequencing panel in patients with advanced lung adenocarcinoma: paving the way for clinical implementation. Cancers (Basel). 2019; 11:1229. https://doi.org/10.3390/cancers11091229 PMID:31443496

43. Jeong Y, Hoang NT, Lovejoy A, Stehr H, Newman AM, Gentles AJ, Kong W, Truong D, Martin S, Chaudhuri A, Heiser D, Zhou L, Say C, et al. Role of KEAP1/NRF2 and TP53 mutations in lung squamous cell carcinoma development and radiation resistance. Cancer Discov. 2017; 7:86–101. https://doi.org/10.1158/2159-8290.CD-16-0127 PMID:27663899

44. Mann KM, Ying H, Juan J, Jenkins NA, Copeland NG. KRAS-related proteins in pancreatic cancer. Pharmacol Ther. 2016; 168:29–42. https://doi.org/10.1016/j.pharmthera.2016.09.003 PMID:27595930

45. Urosevic J, Garcia-Albéniz X, Planet E, Real S, Céspedes MV, Guiu M, Fernandez E, Bellmunt A, Gawrzak S, Pavlovic M, Mangues R, Dolado I, Barriga FM, et al. Colon cancer cells colonize the lung from established liver metastases through p38 MAPK signalling and PTHLH. Nat Cell Biol. 2014; 16:685–94. https://doi.org/10.1038/ncb2977 PMID:24880666

46. Arbour KC, Jordan E, Kim HR, Dienstag J, Yu HA, Sanchez-Vega F, Lito P, Berger M, Solit DB, Hellmann M, Kris MG, Rudin CM, Ni A, et al. Effects of co-occurring genomic alterations on outcomes in patients with KRAS-mutant non-small cell lung cancer. Clin Cancer Res. 2018; 24:334–40. https://doi.org/10.1158/1078-0432.CCR-17-1841 PMID:29089357

47. Kotsantis P, Silva LM, Irmscher S, Jones RM, Folkes L, Gromak N, Petermann E. Increased global transcription activity as a mechanism of replication stress in cancer.

Nat Commun. 2016; 7:13087.
https://doi.org/10.1038/ncomms13087
PMID:27725641

48. Jung H, Kim HS, Kim JY, Sun JM, Ahn JS, Ahn MJ, Park K, Esteller M, Lee SH, Choi JK. DNA methylation loss promotes immune evasion of tumours with high mutation and copy number load. Nat Commun. 2019; 10:4278.
https://doi.org/10.1038/s41467-019-12159-9
PMID:31537801

49. Wang Z, Duan J, Cai S, Han M, Dong H, Zhao J, Zhu B, Wang S, Zhuo M, Sun J, Wang Q, Bai H, Han J, et al. Assessment of blood tumor mutational burden as a potential biomarker for immunotherapy in patients with non-small cell lung cancer with use of a next-generation sequencing cancer gene panel. JAMA Oncol. 2019; 5:696–702.
https://doi.org/10.1001/jamaoncol.2018.7098
PMID:30816954

50. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43:e47.
https://doi.org/10.1093/nar/gkv007
PMID:25605792

51. Bernhardt S, Bayerlová M, Vetter M, Wachter A, Mitra D, Hanf V, Lantzsch T, Uleer C, Peschel S, John J, Buchmann J, Weigert E, Bürrig KF, et al. Proteomic profiling of breast cancer metabolism identifies SHMT2 and ASCT2 as prognostic factors. Breast Cancer Res. 2017; 19:112.
https://doi.org/10.1186/s13058-017-0905-7
PMID:29020998

52. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010; 33:1–22.
PMID:20808728

53. 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

54. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database (MSigDB) hallmark gene set collection. Cell Syst. 2015; 1:417–25.
https://doi.org/10.1016/j.cels.2015.12.004
PMID:26771021

55. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013; 14:7.
https://doi.org/10.1186/1471-2105-14-7
PMID:23323831

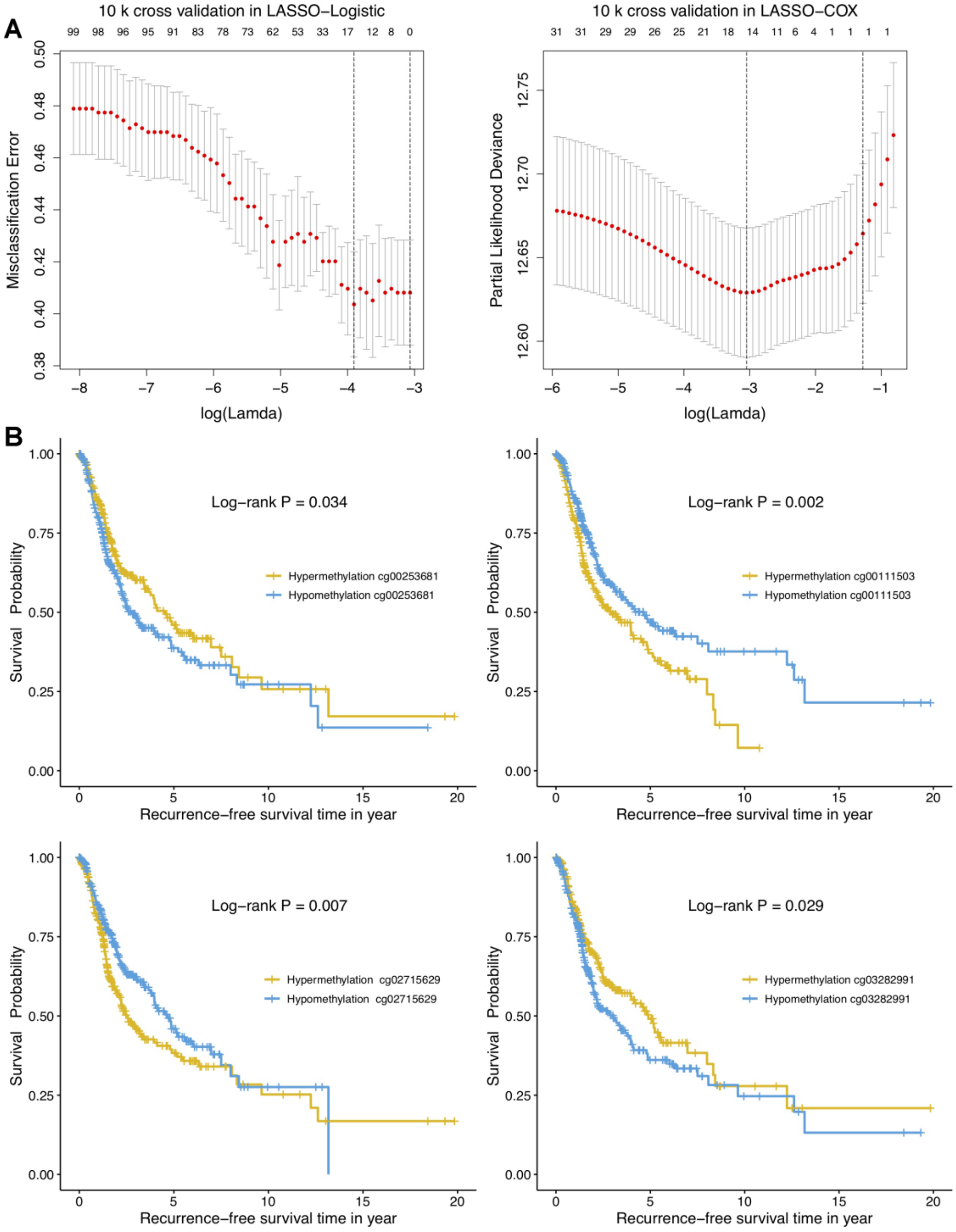
56. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499:214–18.
https://doi.org/10.1038/nature12213 PMID:23770567

57. 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

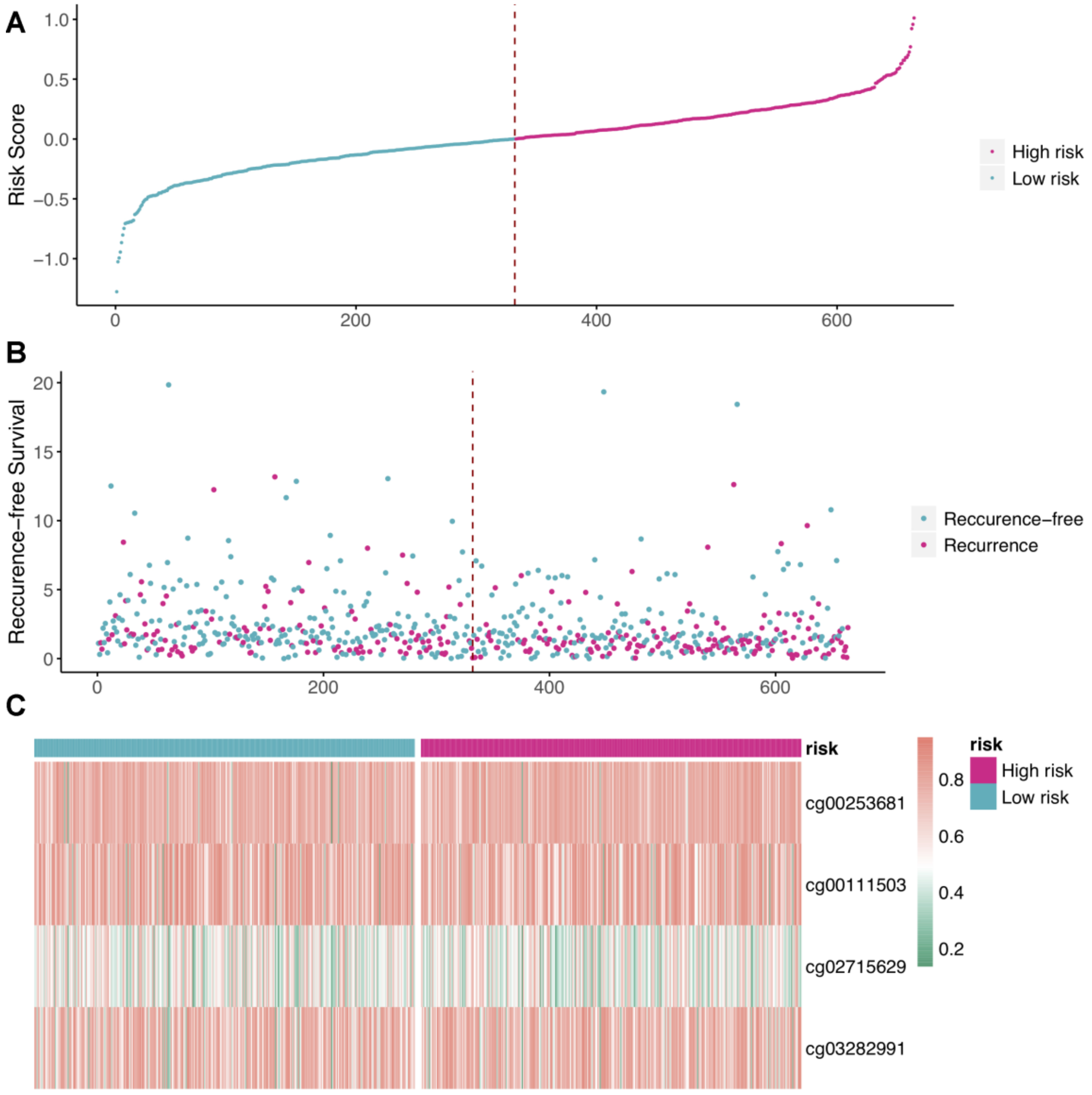
58. Zheng SC, Breeze CE, Beck S, Teschendorff AE. Identification of differentially methylated cell types in epigenome-wide association studies. Nat Methods. 2018; 15:1059–66.
https://doi.org/10.1038/s41592-018-0213-x
PMID:30504870

SUPPLEMENTARY MATERIALS

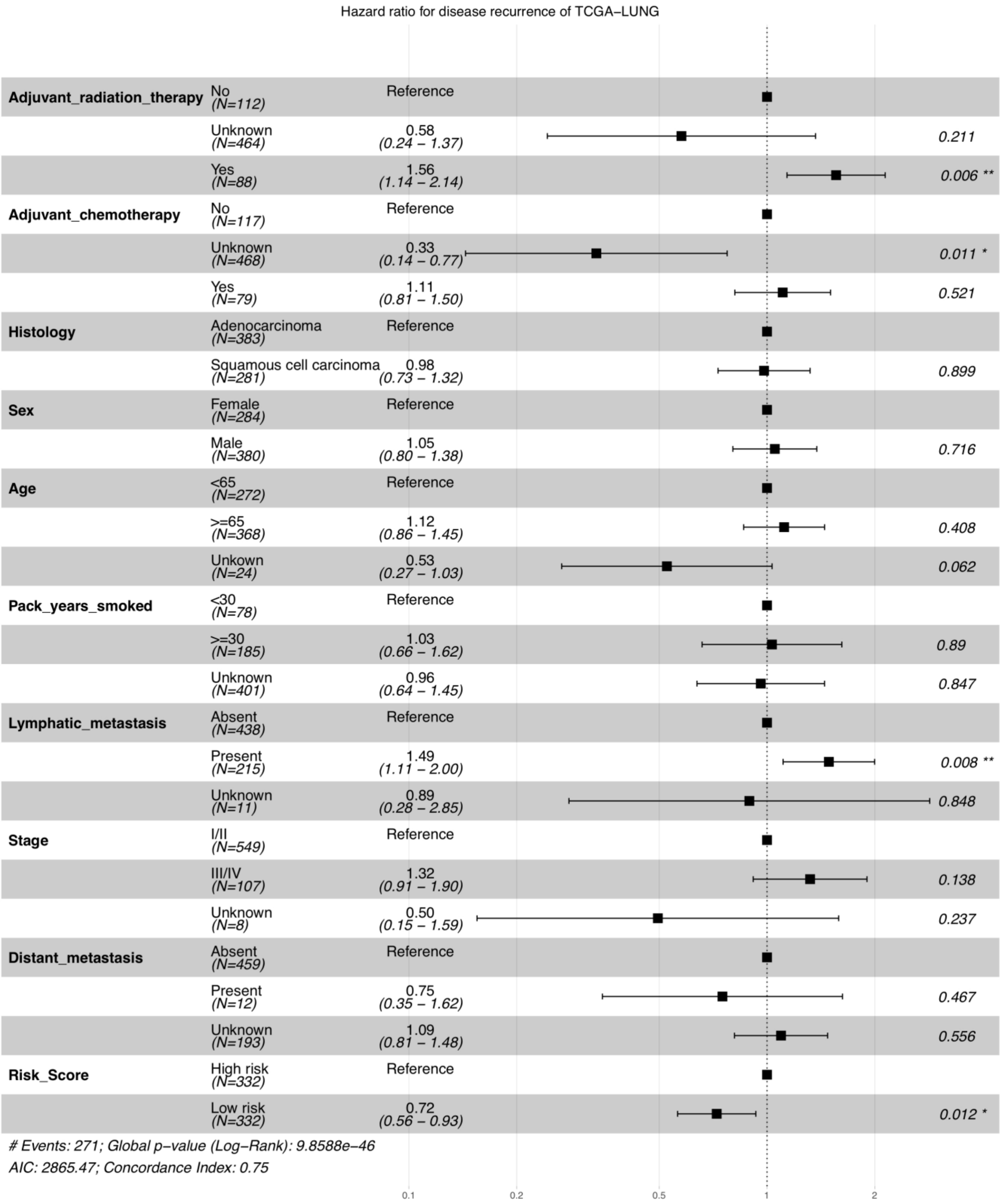
Supplementary Figures



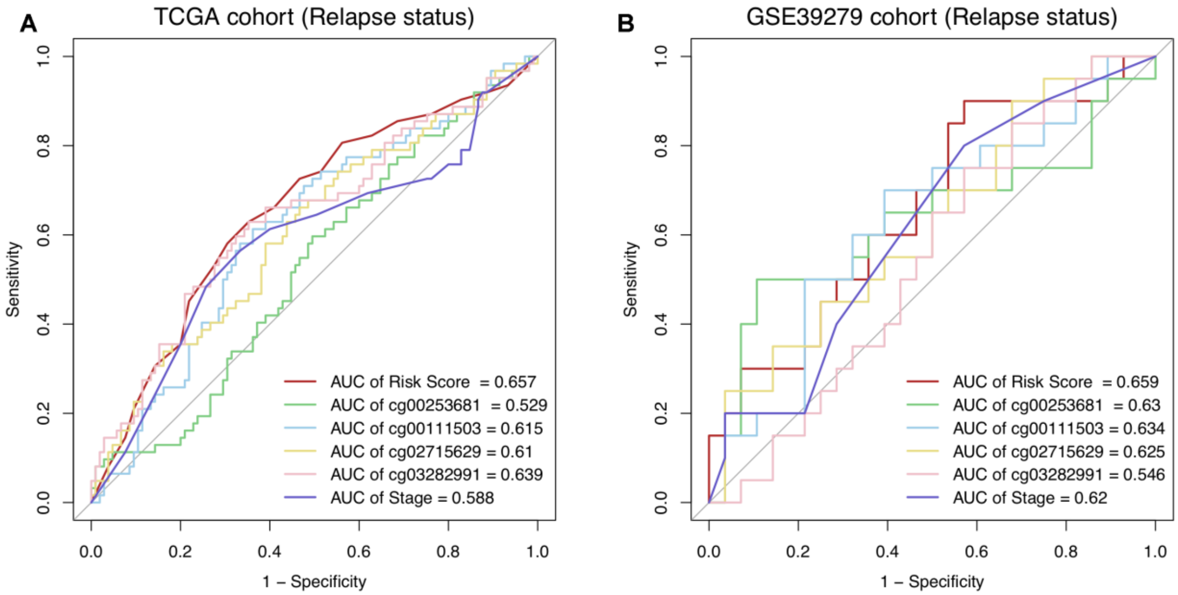
Supplementary Figure 1. (**A**) Ten-fold cross validations performed for obtaining optimal parameter lambda (λ) in LASSO-Logistic (left) and LASSO-Cox analysis (right). The dotted vertical lines were plotted at values of log (λ) by minimum criteria and 1-Standard Error criteria in two LASSO models, respectively. The optimal values of λ were determined by minimum criteria where two dotted vertical lines were drawn in Figure 3A and Figure3D, and thus 14 and 9 CpG markers with nonzero coefficients were screened out. (**B**) Kaplan-Meier curves of four final selected CpGs present their correlation with recurrence and prognostic prediction of NSCLC patients in training cohort.



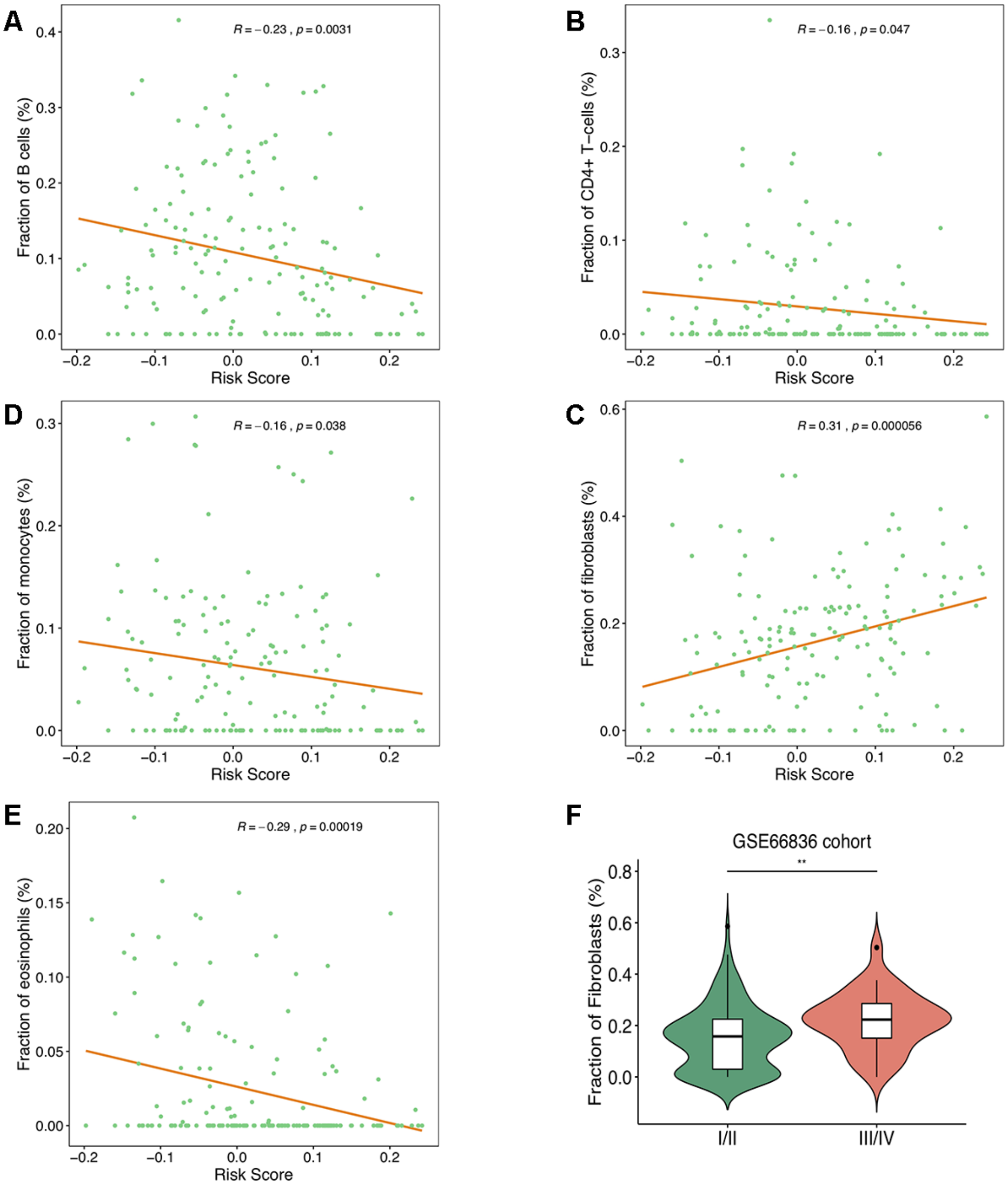
Supplementary Figure 2. Association chart of relapse risk factors. (**A**, **B**) Distribution of risk score and RFS of TCGA NSCLC patients in high- and low-risk groups. (**C**) The heatmap of DNAm profile of 4 final selected CpGs.



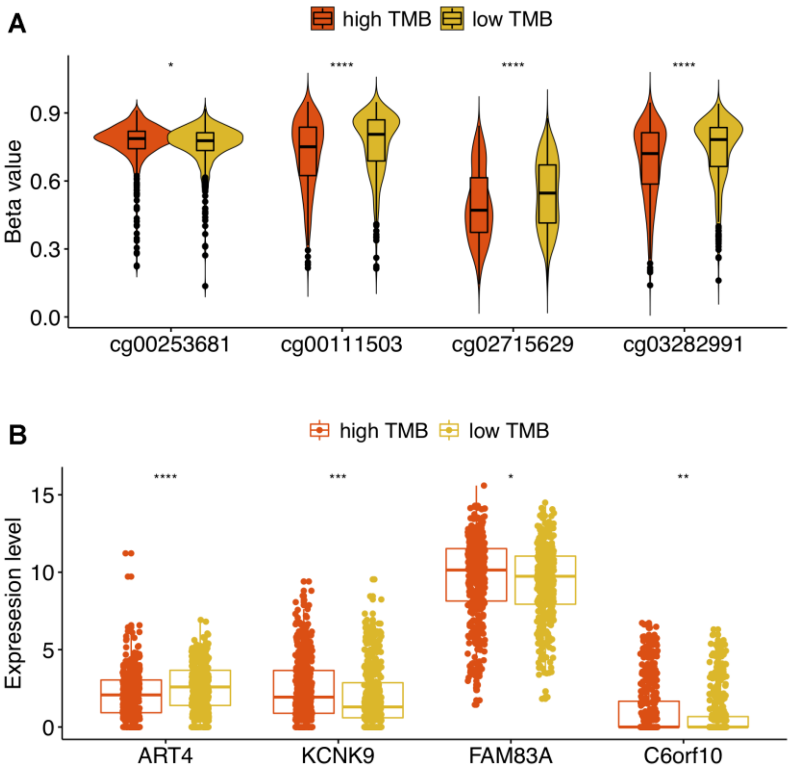
Supplementary Figure 3. Multivariate Cox regression analysis for RFS of TCGA NSCLC patients with combinations of DNAm-based risk score and clinical factors.



Supplementary Figure 4. Receiver operating characteristic (ROC) curves of the combined risk score, clinical stage and 4 separate CpGs demonstrate their performance of discrimination for relapse status in training (**A**) and validation (**B**) cohorts.



Supplementary Figure 5. Estimated compositions of B-cells (**A**), CD4+ T-cells (**B**), fibroblasts (**C**), monocytes (**D**) and eosinophils (**E**) were significantly correlated with DNAm-based risk score in GSE66836 cohort. (**F**) Fraction of fibroblasts in NSCLC samples at late stage was significantly higher than those at early stage.



Supplementary Figure 6. The correlation of TMB with four CpGs methylation status and expression of four nearby genes. (**A**) Methylation levels of 4 identified DMPs in high TMB compared with low TMB group. (**B**) Differential expression of 4 reference genes in high and low TMB.

Supplementary Tables

Supplementary Table 1. Clinical characteristics of patients for included study cohorts.

Characteristics		Training cohort (TCGA)	Validation cohort (GSE39279)	Validation cohort (GSE66836)	Validation cohort (GSE119144)
Total		n=823	n=444	n=164	n=60
Sex	Female	340	190	91	
	Male	483	254	73	
Age	<65	318	200		
	>=65	466	243		
	Unknown	39	1		
Pack-years smoked	<30	94	124		
	>=30	228	237		
	Unknown	501	83		
Histology	Adenocarcinoma	455	322	164	
	Squamous cell carcinoma	368	122	0	
Stage	I	420	237	93	
	II	244	94	40	
	III	127	102	29	
	IV	24	11	2	
	Unknown	8			
Lymphatic metastasis	Absent	534	239		
	Present	273	116		
	Unknown	16	89		
Distant metastasis	Absent	577	344		
	Present	23	11		
	Unknown	223	89		
Adjuvant radiation therapy	Yes	129	235		
	No	90	39		
	Unknown	604	170		
Adjuvant chemotherapy	Yes	136	250		
	No	79	24		
	Unknown	608	170		
Outcome(RFS)	Recurrence-free	399	150		10
	Recurrence	269	161		49
	Unknown	155	133		1
Outcome(OS)	Alive	502			
	Dead	321			
Follow up time(year/month)	Available(PFS)	662			59
	Unknown(PFS)	161			1
	Available(OS)	808			
	Unknown(OS)	15			

*RFS : recurrence-free survival status; OS : overall survival status.

Supplementary Table 2. Recurrence associated CpG markers identified by Univariate Cox, Random Forest and LASSO methods in training cohort.

Marker ID	Chr	Pos	Ref Gene	Location	logFC	adj.P	Method
cg00017489	chr7	153583318	DPP6	TSS1500	0.291	5.61E-39	LASSO-Logistic/Random Forest
cg00253681	chr12	14996583	ART4	TSS200	0.119	2.50E-21	LASSO-Logistic/Random Forest/Univariate Cox/LASSO-Cox
cg00682263	chr15	66188803	MEGF11	3'UTR	0.382	2.88E-87	LASSO-Logistic
cg02382109	chr8	22785456	PEBP4	TSS200	0.064	2.43E-06	LASSO-Logistic/Random Forest/LASSO-Cox
cg03389538	chr3	128779498	GP9	TSS200	0.043	7.37E-07	LASSO-Logistic/Random Forest
cg03502002	chr18	74962133	GALR1	1stExon;5'UTR	0.463	7.43E-52	LASSO-Logistic/Random Forest/LASSO-Cox
cg00111503	chr8	140631116	KCNK9	Body	-0.169	7.35E-20	LASSO-Logistic/Random Forest/Univariate Cox/LASSO-Cox
cg00814751	chr5	176072170	EIF4E1B	Body	-0.095	1.19E-10	LASSO-Logistic
cg01522296	chr22	50452415	IL17REL	TSS1500	-0.107	2.26E-13	LASSO-Logistic/Random Forest
cg02310286	chr8	88886432	DCAF4L2	TSS200	-0.141	9.58E-11	LASSO-Logistic/Random Forest/Univariate Cox
cg02407493	chr16	2068942	NPW	TSS1500	-0.148	5.53E-19	LASSO-Logistic/Random Forest/Univariate Cox
cg02715629	chr8	124193817	FAM83A	TSS1500;TSS1500	-0.247	1.75E-35	LASSO-Logistic/Random Forest/Univariate Cox/LASSO-Cox
cg02901006	chr19	8117024	CCL25	TSS1500	-0.117	1.57E-10	LASSO-Logistic/Random Forest
cg03282991	chr6	32294260	C6orf10	Body	-0.166	1.03E-17	LASSO-Logistic/Univariate Cox/LASSO-Cox
cg00446413	chr7	153749206	DPP6	TSS1500;Body	0.264	1.41E-50	Random Forest
cg02263813	chr16	56672640	MT1A	1stExon;5'UTR	0.233	2.38E-28	Random Forest
cg02099194	chr13	43149689	TNFSF11	Body;5'UTR	-0.153	6.42E-19	Random Forest
cg00914726	chr1	60539400	C1orf87	1stExon;5'UTR	0.233	3.82E-17	Random Forest
cg00472801	chr6	62995876	KHDRBS2	1stExon;5'UTR	0.132	8.77E-13	Random Forest
cg02096663	chr4	178650141	LOC285501	Body	-0.147	2.55E-10	Random Forest
cg00174500	chr14	23846479	CMTM5	1stExon;1stExon	0.045	0.000975951	Random Forest/LASSO-Cox
cg02062418	chr16	1494677	CCDC154	TSS200	-0.036	7.32E-06	Random Forest/LASSO-Cox
cg03322234	chr7	1022643	CYP2W1	TSS200	-0.033	0.046378826	Random Forest/LASSO-Cox
cg02992224	chr11	93822294	HEPHL1	Body	-0.123	1.25E-11	Random Forest
cg01466017	chr10	17496720	ST8SIA6	TSS1500	0.095	1.55E-07	Univariate Cox
cg03377767	chr2	17997138	MSGN1	TSS1500	-0.249	5.44E-47	Univariate Cox

*logFC: log2 fold change; adj.P: Benjamini-Hochberg adjusted P value.

Supplementary Table 3. Multivariable regression analysis for PFS of TCGA NSCLC patients conducted on clinical factors in combination with 13 biomarkers identified by LASSO-Cox and univariate Cox models.

Characteristics		Coefficient	Hazard Ratio
	Yes vs. No	0.461	1.59
	Unknown vs. No	-0.514	0.6
	Yes vs. No	0.206	1.23
	Unknown vs. No	-1.13	0.32
Histology	Squamous cell carcinoma vs. Adenocarcinoma	0.098	1.1
Sex	Male vs. Female	0.062	1.06
Age	>=65 vs. <65	0.17	1.19
	Unknown vs. <65	-0.552	0.58
Pack-years smoked	>=30 vs. <30	0.091	1.1
	Unknown vs. <30	-0.037	0.96
Lymphatic metastasis	Present vs. Absent	0.373	1.45
	Unknown vs. Absent	0.009	1.01
Stage	III/IV vs. I/II	0.347	1.41
	Unknown vs. I/II	-0.872	0.42
Distant metastasis	Present vs. Absent	-0.422	0.66
	Unknown vs. Absent	0.049	1.05
cg00253681	**Hypermethylation vs. Hypomethylation**	**0.339**	**1.4**
cg02382109	Hypermethylation vs. Hypomethylation	0.057	1.06
cg03502002	Hypermethylation vs. Hypomethylation	-0.27	0.76
cg00111503	**Hypermethylation vs. Hypomethylation**	**-0.337**	**0.71**
cg02715629	**Hypermethylation vs. Hypomethylation**	**-0.138**	**0.87**
cg03282991	**Hypermethylation vs. Hypomethylation**	**0.037**	**1.04**
cg00174500	Hypermethylation vs. Hypomethylation	-0.389	0.68
cg02062418	Hypermethylation vs. Hypomethylation	0.261	1.3
cg03322234	Hypermethylation vs. Hypomethylation	0.302	1.35
cg01466017	Hypermethylation vs. Hypomethylation	0.249	1.28
cg02310286	Hypermethylation vs. Hypomethylation	-0.18	0.84
cg02407493	Hypermethylation vs. Hypomethylation	0.064	1.07
cg03377767	Hypermethylation vs. Hypomethylation	-0.044	0.96

Supplementary Table 4. Multivariate Cox regression analysis for OS of TCGA NSCLC patients with combinations of DNAm-based risk score and clinical factors.

Characteristics		Hazard Ratio	CI	P Value
Adjuvant chemotherapy	Yes vs. No	0.71	0.48-1.06	0.094
	Unknown vs. No	0.43	0.2-0.93	0.032
Adjuvant radiation therapy	Yes vs. No	1.28	0.86-1.89	0.218
	Unknown vs. No	1.26	0.58-2.77	0.557
Age	>=65 vs. <65	1.23	0.96-1.57	0.095
	Unknown vs. <65	0.55	0.26-1.14	0.11
Distant metastasis	Present vs. Absent	1.72	0.96-3.09	0.069
	Unknown vs. Absent	1.17	0.88-1.56	0.278
Histology	Squamous cell carcinoma vs. Adenocarcinoma	1.17	0.9-1.52	0.232
Lymphatic metastasis	Present vs. Absent	1.53	1.17-2.02	0.002
	Unknown vs. Absent	1.58	0.69-3.61	0.279
Pack-years smoked	>=30 vs. <30	1.1	0.72-1.69	0.652
	Unknown vs. <30	1.08	0.73-1.6	0.707
Risk Model	High risk vs. Low risk	1.4	1.11-1.77	0.004
Sex	Male vs. Female	0.99	0.78-1.27	0.963
Stage	III/IV vs. I/II	1.41	1.02-1.96	0.04
	Unknown vs. I/II	0.95	0.3-3.03	0.935

*CI: 95% confidence interval.