Am J Transl Res 2019;11(4):1997-2012
www.ajtr.org /ISSN:1943-8141/AJTR0083746

Original Article

Novel methylation-driven genes identified as prognostic indicators for lung squamous cell carcinoma

Yin Li, Jie Gu, Fengkai Xu, Qiaoliang Zhu, Di Ge, Chunlai Lu

Department of Thoracic Surgery, Zhongshan Hospital, Fudan University, Shanghai, P. R. China

Received August 8, 2018; Accepted January 28, 2019; Epub April 15, 2019; Published April 30, 2019

Abstract: Lung cancer remains the leading cause of cancer death. DNA methylation plays an essential role in carcinogenesis through regulating gene expression and gene alternative splicing. However, the role of methylation in the tumorigenesis of lung squamous cell carcinoma (SCC) and its association with prognosis remains unclear. Here, we used an integrative approach to evaluate the prognostic value of epigenetic processes in lung SCC by examining the data provided by The Cancer Genome Atlas (TCGA). We found that the mean methylation level was significantly decreased in lung SCC. We also identified methylation-driven genes which were associated with cancer-related pathways. The multivariate Cox regression analysis showed four methylation-driven genes, GCSAM, GPR75, NHLRC1, and TRIM58, could be served as prognostic indicators for lung SCC. Validation on two external GEO datasets showed consistent methylation alterations of the four genes. These findings may have important implications in the understanding of the potential therapeutic method for lung SCC.

Keywords: Methylation, lung squamous cell carcinoma, prognostic indicators, TCGA

Introduction

Lung cancer remains the leading cause of cancer death in both males and females around the world [1]. The overall 5-year survival rate for lung cancer patients remains low at about 17% [2]. The two major histological classes of lung cancer are non-small-cell lung cancer (NSCLC) and small-cell lung cancer (SCLC). Non-small cell lung cancer (NSCLC) can be further classified as non-squamous carcinoma and lung squamous cell carcinoma (SCC), among which the proportion of lung SCC was as high as nearly 40% [3]. For patients with NSCLC, the best prognosis can be achieved by complete surgical resection of stage IA disease. However, this requires early detection and accurate disease prognosis prediction. Lung cancer can be detected early by a computed tomography scan in high-risk individuals, but this method has a high false-positive rate that can lead to unnecessary treatment [4]. Therefore, more precise prognostic indicators are urgently demanded.

As one of the critical epigenetic modifications, DNA methylation, the addition of a methyl group to DNA, plays a vital role in carcinogenesis through regulating gene expression and gene alternative splicing [5]. Aberrant change of methylation level of genes is a common problem in cancer development. Previous studies regarding methylation events of lung SCC significantly improved our understanding of this devastating disease, yet the exact role of methylation in the tumorigenesis of lung SCC and association with prognosis remains largely unknown. Advanced high-throughput technologies have been applied to identifying the epigenetic abnormalities in various diseases [6, 7]. However, many methylation events found to be statistically significant using high-throughput technologies are not correlated with gene expression changes. Therefore, integrating data across multiple platforms to determine the epigenetic events that are most likely to be involved in lung SCC is needed.

The Cancer Genome Atlas (TCGA), a project supported by the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), has generated comprehensive, multi-dimensional maps of the key genomic changes in various types of cancers. In this study, we used an integrative approach to evaluate the prognostic value of epigenetic processes and found four methylation-driven signatures that could effectively predict patient survival. We hope that these genes can be used as a novel prognostic tool for lung SCC and provide more insights into the molecular mechanisms of this prevalent and devastating disease.

Materials and methods

Lung SCC data acquisition from TCGA and preprocessing

Clinical information of lung SCC patients and lung SCC DNA methylation data based on the Illumina Infinium Human Methylation 450 platform, including 370 tumor samples and 42 normal samples, were downloaded from TCGA using the Data Transfer Tool provided by GDC Apps (https://portal.gdc.cancer.gov/). The downloaded methylation data had already been pre-processed via TCGA pipelines and contained information of each probe's beta-value, chromosomal location with corresponding predictive genes. Probe-level data was condensed to a summary beta value for each gene by calculating average methylation value for all CpG sites associated with a gene in each sample. Next, all data were normalized in the R computing environment using the limma package [8]. Differential methylation analysis was conducted using TCGAbiolinks [9]. Gene expression quantification data (502 tumor samples and 49 normal samples) was also obtained by the same approach and processed by edgeR [10].

Integrative analysis

Aberrant DNA methylation is an important mechanism that contributes to oncogenesis. Although many algorithms exist that exploit this vast dataset to identify hypo- and hypermethylated genes in cancer, few are capable of identifying differentially methylated genes that are also predictive of transcription. The R package MethylMix was designed to perform an analysis integrating methylation data and gene expression data [11]. There are three steps to the MethylMix analysis: first, genes are filtered to identify methylation events that result in gene expression changes; second, a univariate beta mixture modeling is used to define a methylation state across a large number of patients; and third, Wilcoxon rank sum test is conducted to compare DNA methylation level in tumor tissues vs. normal tissues. In this present study, correlation coefficient < -0.3 (adjusted *p*-value < 0.05) between DNA methylation and matched gene expression was considered significant (minus signifies negatively correlated). No methylation fold change cut-offs were implemented so that a gene whose methylation level change is dramatically related to expression would not be excluded even if this change is rather slight.

Enrichment analysis

Gene ontology (GO) and pathway analysis were performed with ConsensusPathDB to identify in which aspect these methylation-driven genes are involved in the development and progression of lung SCC. ConsensusPathDB is a database that integrates different types of functional interactions to assemble a more complete and a less biased picture of cellular biology. The database content is updated every three months. For our analysis, GO level 2, 3, 4, and 5 categories were included (*p*-value < 0.01). For pathway analysis, the following pathway databases were selected for our study: Reactome, Inoh, Pid, Biocarta, Netpath, Humancyc, Kegg, Wikipathways, Smpdb, Pharmgkb, Ehmn, and Signalink. We used the default settings: minimum overlap and the *p*-value cutoff of 0.01.

Multivariate Cox proportional hazard regression analysis

The prognostic value of the methylation-driven genes was first assessed by the univariate Cox proportional hazards regression analysis. These methylation-driven genes with a *p*-value < 0.05 were regarded as prognostic methylation-driven genes. These prognostic methylation-driven genes were then used for prognostic model construction. To evaluate the relative contribution of these prognostic methylation-driven genes to lung SCC survival prediction, they were fitted into a multivariate Cox regression analysis. A methylation-based prognostic risk score model was constructed by the linear combination of the methylation levels of the methylation-driven genes with the multivariate Cox regression coefficient (β) as the weight. The risk score formula was as follows: risk score = methylation of gene$_1$ × β_1 gene$_1$ + methylation of gene$_2$ × β_2 gene$_2$ +...methylation of

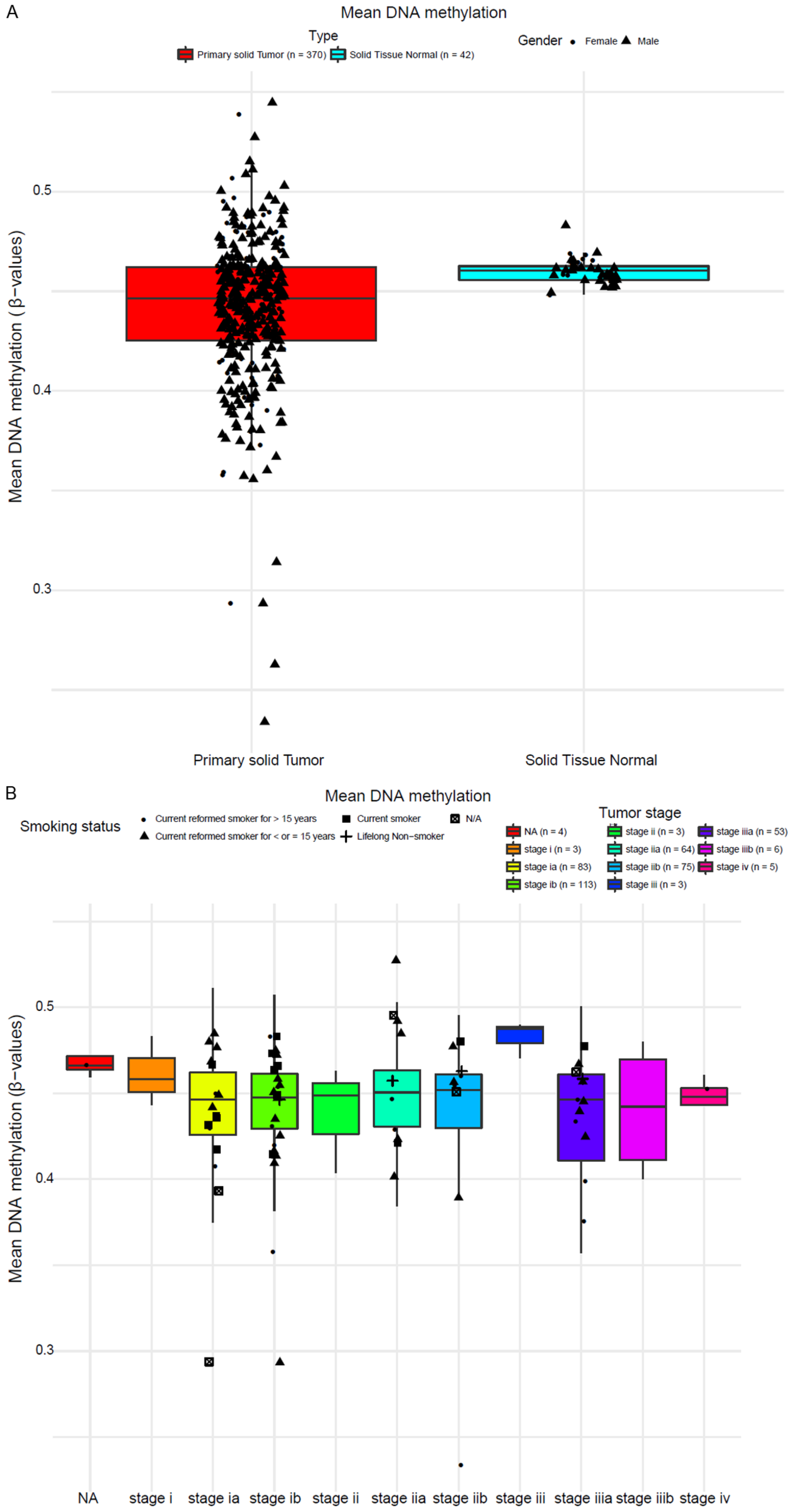

A
Mean DNA methylation
Type
Primary solid Tumor (n = 370)
Solid Tissue Normal (n = 42)
Gender
Female
Male
Mean DNA methylation (β−values)
0.5
0.4
0.3
Primary solid Tumor
Solid Tissue Normal
B
Mean DNA methylation
Smoking status
Current reformed smoker for > 15 years
Current reformed smoker for < or = 15 years
Current smoker
Lifelong Non−smoker
N/A
Tumor stage
NA (n = 4)
stage i (n = 3)
stage ia (n = 83)
stage ib (n = 113)
stage ii (n = 3)
stage iia (n = 64)
stage iib (n = 75)
stage iii (n = 3)
stage iiia (n = 53)
stage iiib (n = 6)
stage iv (n = 5)
Mean DNA methylation (β−values)
0.5
0.4
0.3
NA
stage i
stage ia
stage ib
stage ii
stage iia
stage iib
stage iii
stage iiia
stage iiib
stage iv

Figure 1. Clinical information of lung SCC DNA methylation samples queried from TCGA. A. Overall mean DNA methylation level alteration in tumor tissues (n = 370) vs. normal tissues (n = 42), the tumor tissues of TCGA cohort shows global lower methylation level compared to normal tissues (p-value < 0.05). B. DNA methylation level distributions concerning different stages of cancer and smoking history (p-value > 0.05).

$gene_n \times \beta_n\ gene_n$. This prognostic model could divide the lung SCC patients into high- and low-risk groups. The time-dependent receiver-operating characteristic (ROC) curve was performed using the survival ROC package on the R platform to evaluate the predictive accuracy. To comprehensively assess our prognostic model, univariate and multivariate Cox regression analyses were conducted to determine the independence of our prognostic model in survival prediction with other clinical variables in lung SCC.

Validating the methylation-related signature

To assess whether the prognostic model related methylation-driven genes exist consistent methylation alterations in other lung SCC patients, two external datasets (GSE39279 and GSE75008) from Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) were used as the validation datasets [12, 13]. GSE39279 contains 122 lung SCC samples, GSE75008 contains 40 normal lung samples and 16 lung SCC samples. All these samples were hybridized to the Illumina Infinium 450k Human Methylation Beadchip. We integrated these two datasets to significantly improve the number of samples and ended up having 40 normal samples vs. 138 tumor samples. The batch effect was removed using sva package [14]. Probe-level data was condensed to a summary beta value for each gene by calculating average methylation value for all CpG sites associated with a gene in each sample.

On the other hand, a least absolute shrinkage and selector operation (LASSO) algorithm was used to further substantiate the accuracy of the four-methylation-driven-gene model [15, 16]. The LASSO analysis was performed using the glmnet package.

CpG site methylation analysis of the four methylation-driven genes

To further investigate the methylation level alterations of the CpG sites related to the four methylation-driven genes, we extracted information of all CpG sites associated with these four genes. Differentially methylation analysis was conducted using limma package in R environment (|beta-value| > 0.10, p-value < 0.05). We then calculated the correlation coefficient between each site and its corresponding gene expression. To promote the reliability of our analysis, we validated our results in MEXPRESS (http://mexpress.be). MEXPRESS is a web tool for the integration and visualization of the expression, DNA methylation and clinical TCGA data on a single-gene level. The data of this tool was downloaded from the TCGA ftp site: level 3 per-gene RNA-seq v2 expression data (UNC IlluminaHiSeq_RNASeqV2), level 3 DNA methylation data (JHU_USC HumanMethylation450) and clinical data (both clinical patient and tumor sample data).

Results

Mean DNA methylation level alteration in tumor tissues

Analysis of TCGA data showed an altered DNA methylation level in lung SCC relative to corresponding normal tissues. The mean methylation level was significantly lower in primary tumors (n = 370) than normal tissues (n = 42) (**Figure 1A**, p-value < 0.05). However, there was no significant difference in mean methylation level among tumor stages (**Figure 1B**, p-value > 0.05). We performed the differential methylation analysis on the probe level to find that there were 19,553 dys-methylated sites (Δ|beta-value| > 0.25, p-value < 0.05 considered significant). The corresponding chromosomal locations of these methylation sites were visualized as a heat-map (**Figure 2**). Chromosome mapping revealed chromosome distribution and relative methylation level changes, with chromosomes 2 and 7 containing the highest number of dys-methylated sites in lung SCC, of 1825 and 1716 sites respectively and chromosome Y containing the least (Supplementary Table 1).

Methylation-driven genes in lung SCC

The analysis of DNA methylation-driven genes using MethylMix was based on a significant

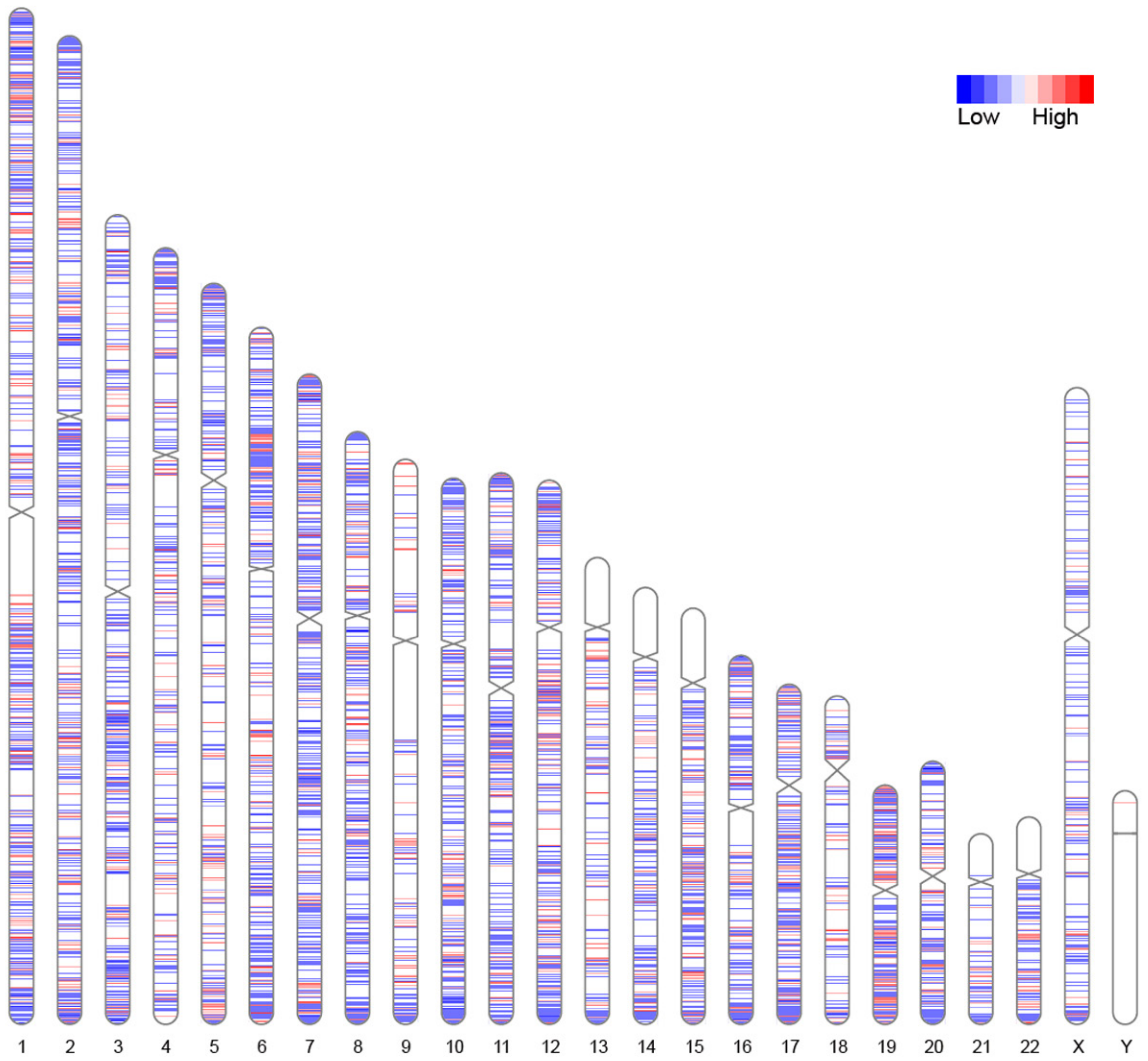


Figure 2. Visualization of chromosomal positions of dys-methylated loci in lung SCC. Red indicates hypermethylation and blue indicates hypomethylation.

predictive effect on gene expression in cancer [17]. As the heatmap shown (**Figure 3A**), a total of 182 differentially methylated-driven genes (136 hypermethylated and 46 hypomethylated) were found based on 370 tumor samples and 42 normal samples in lung SCC (adjusted *p*-value < 0.05). Here, we exhibited the top 10 hypermethylated (left), and hypomethylated genes (right) (**Figure 4**), which displays the MethylMix model showing the low methylation state matches the normal methylation and the high methylation state corresponds to hypermethylation for hypermethylated genes, and the high methylation state matches the normal methylation, and the low methylation state corresponds to hypomethylation for hypomethylated genes. **Figure 5** showed the inverse correlation between DNA methylation and matched gene expression of these 20 genes. Among these 182 methylation-driven genes, ALG1L is the most hypomethylated gene (correlation coefficient: approximately -0.44), whereas ZNF382 is the most hypermethylated gene (correlation coefficient: about -0.43). The most significant effect of DNA methylation on gene expression goes with MKRN3 (around -0.70) (Supplementary Table 2).

To further investigate the biological functions of methylated-driven genes in lung SCC, pathway and function analyses were performed with ConsensusPathDB. The methylation-driven genes were mainly involved in pathways of

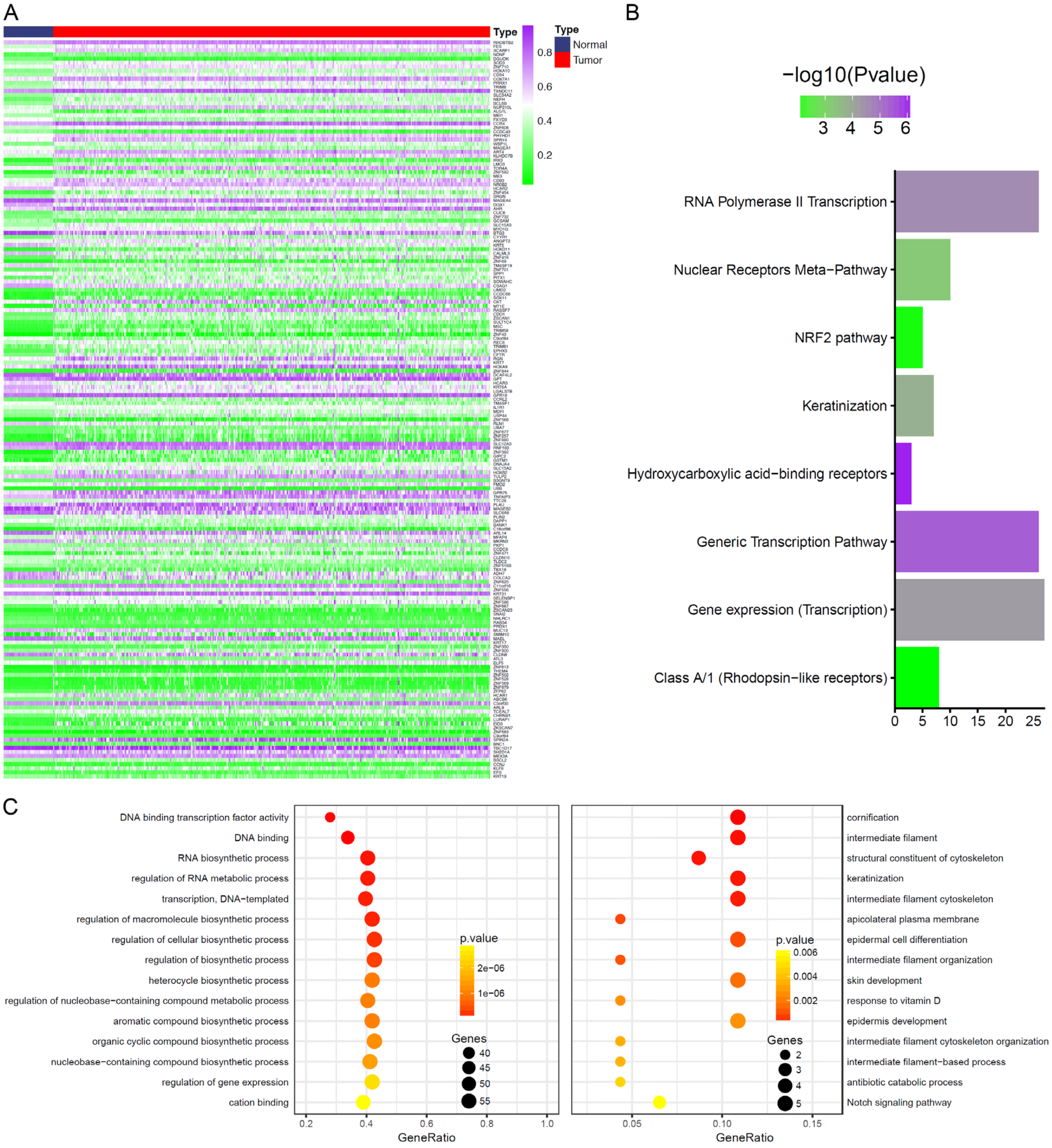


Figure 3. Identification of methylation-driven genes in lung SCC as well as function and pathway analyses. A. Heatmap showing the methylation changes of 182 methylation-driven genes in tumor vs. normal tissues, purple indicates hypermethylation and green indicates hypomethylation. B. Pathways enriched for 182 methylation-driven genes (*p*-value < 0.01). Horizontal axis demonstrates the number of genes enriched in their corresponding pathways; the vertical axis represents the significantly enriched pathways. C. Bubble plots showing GO annotation results (*p*-value < 0.05) for genes that were hypermethylated (left) or hypomethylated (right), respectively.

hydroxycarboxylic acid-binding receptors, generic transcription pathway, RNA Polymerase II Transcription, Nuclear Receptors Meta-Pathway, Class A/1 (Rhodopsin-like receptors), and NRF2 pathway (**Figure 3B**). The bubble plots show GO enrichment data for methylated genes (**Figure 3C**). For hypermethylated genes (left), they were predominantly associated with DNA binding transcription factor activity, RNA biosynthetic process, regulation of RNA metabolic process, regulation of biosynthetic process and regulation of gene expression. For hypomethylated genes (right), they were mainly involved in cornification, intermediate filament, Notch signaling pathway, antibiotic catabolic process, and intermediate filament-based process.

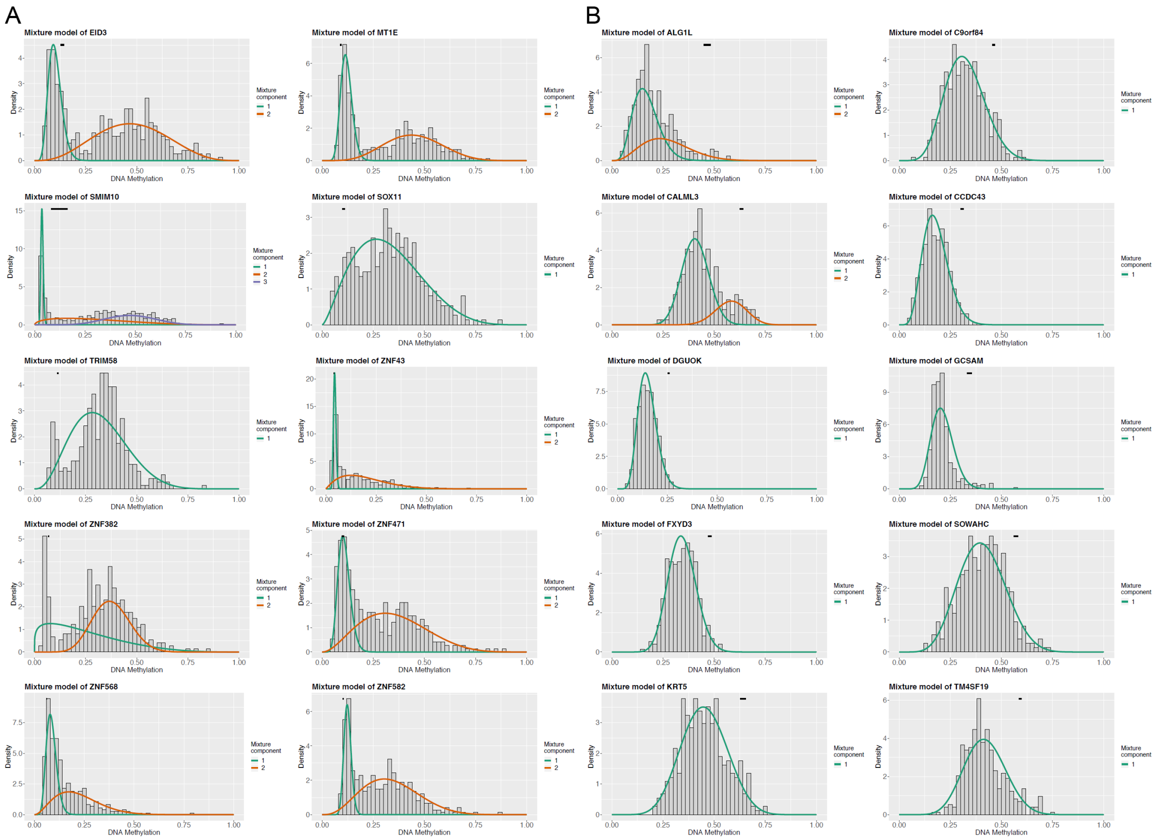


Figure 4. Summary of top 10 hypermethylated (A) and hypomethylated genes (B). The histogram indicates the distribution of methylation across all tumor samples. The black bar above represents the distribution of methylation values across normal samples.

Construction of prognostic signature based on methylation-driven genes

The univariate Cox regression analysis results suggested that there were 24 methylation-driven genes gained significant prognostic value for lung SCC with a *p*-value less than 0.05 (Supplementary Table 3). Next, the multivariate Cox regression analysis was conducted to evaluate the relative contribution of these prognostic methylation-driven genes in survival prediction. We ended up having four methylation-driven genes, GCSAM, GPR75, NHLRC1, and TRIM58, as independent prognostic indicators (*p*-value < 0.05). Individual survival analysis of these four genes was performed. GCSAM and NHLRC1 were hypomethylated in lung SCC whereas GPR75 and TRIM58 were hypermethylated. Patients with low methylation levels of GCSAM and GPR75 had longer survival times while patients with high methylation levels of NHLRC1 and TRIM58 had longer survival times (**Figure 6A**). A prognostic prediction model was constructed based on these four genes. Each patient was assigned with a risk score (Supplementary Table 4). Survival analysis showed that patients with high-risk scores had a shorter survival time than those with low-risk scores (**Figure 6B**). Time-dependent ROC analysis with the survival ROC package demonstrated that prognostic model performed relatively well in lung SCC overall survival prediction, and the area under the curve (AUC) of the time-dependent ROC curve was 0.717, 0.704, 0.708, and 0.707 for 2-, 3-, 4-, and 5-year survival (**Figure 6C**). To assess whether the survival prediction ability of our prognostic model is independent of other clinical or pathological factors of the patients with lung SCC, Cox regression analysis was performed. Selected variables included gender, age, race, tumor stage, T stage, N stage, cigarettes per day and our risk-score model. The results showed that our risk-score model (HR = 1.7773, 95% CI 1.361954 to 2.3194, *p*-value = 2.29e-05) was an independent prognostic factor for patients with lung SCC (**Table 1**).

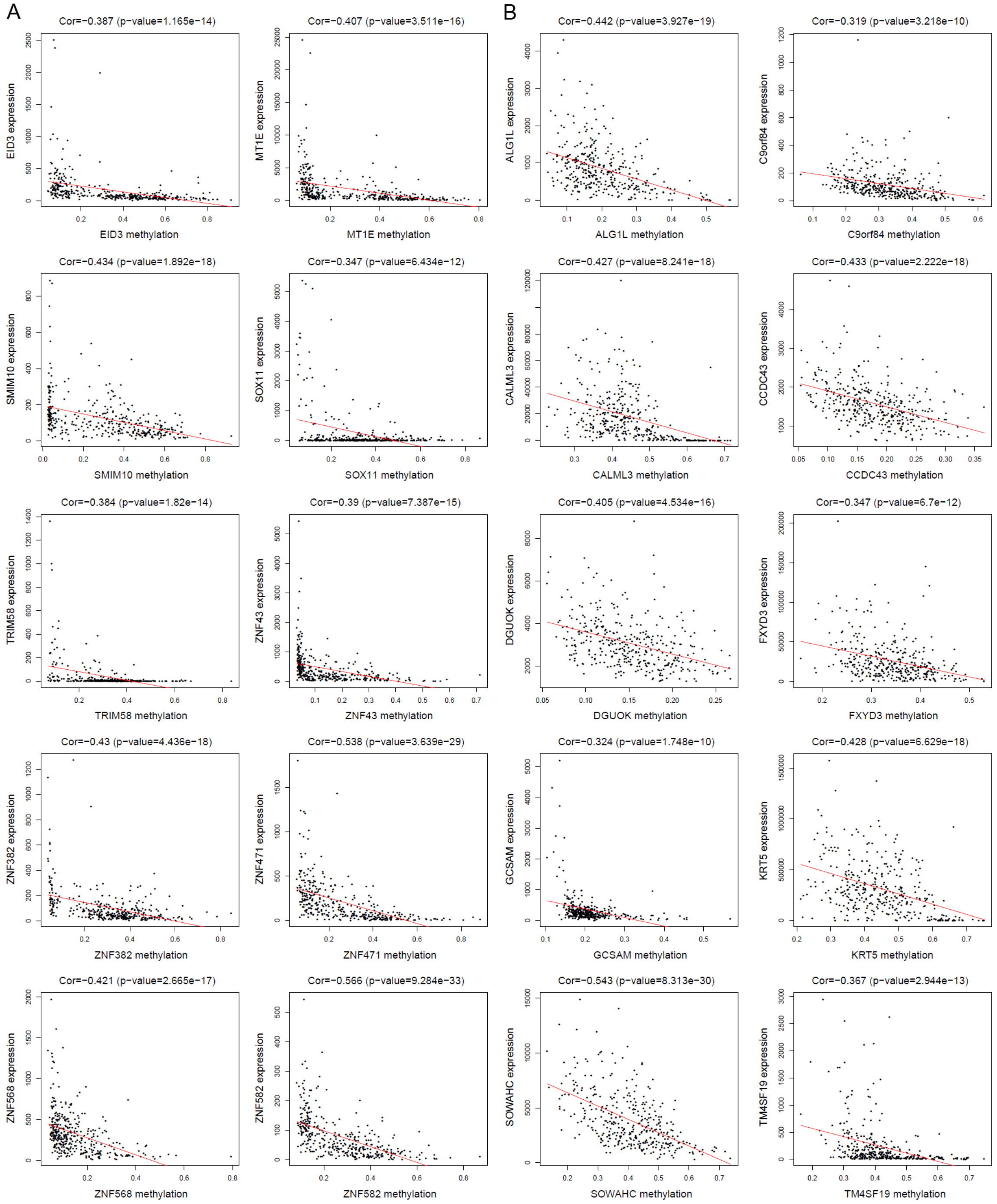


Figure 5. The correlation coefficients between gene expression values and methylation values of the top 10 hypermethylated (A) and hypomethylated genes (B). The higher the correlation coefficient, the stronger interaction between gene expression and methylation.

Verifying the signature

The methylation alterations of each of the four genes, GCSAM, GPR75, NHLRC1, and TRIM58, were evaluated on two GEO datasets (GSE39279 and GSE75008). Consistent with the findings in TCGA cohort, GCSAM and NHLRC1 were hypomethylated while GPR75 and TRIM58 were hypermethylated in tumor samples (**Figure 7**, *p*-value < 0.05).

Next, the four methylation-driven genes were fitted into LASSO Cox regression analysis with 10-fold cross-validation. When the Log lambda

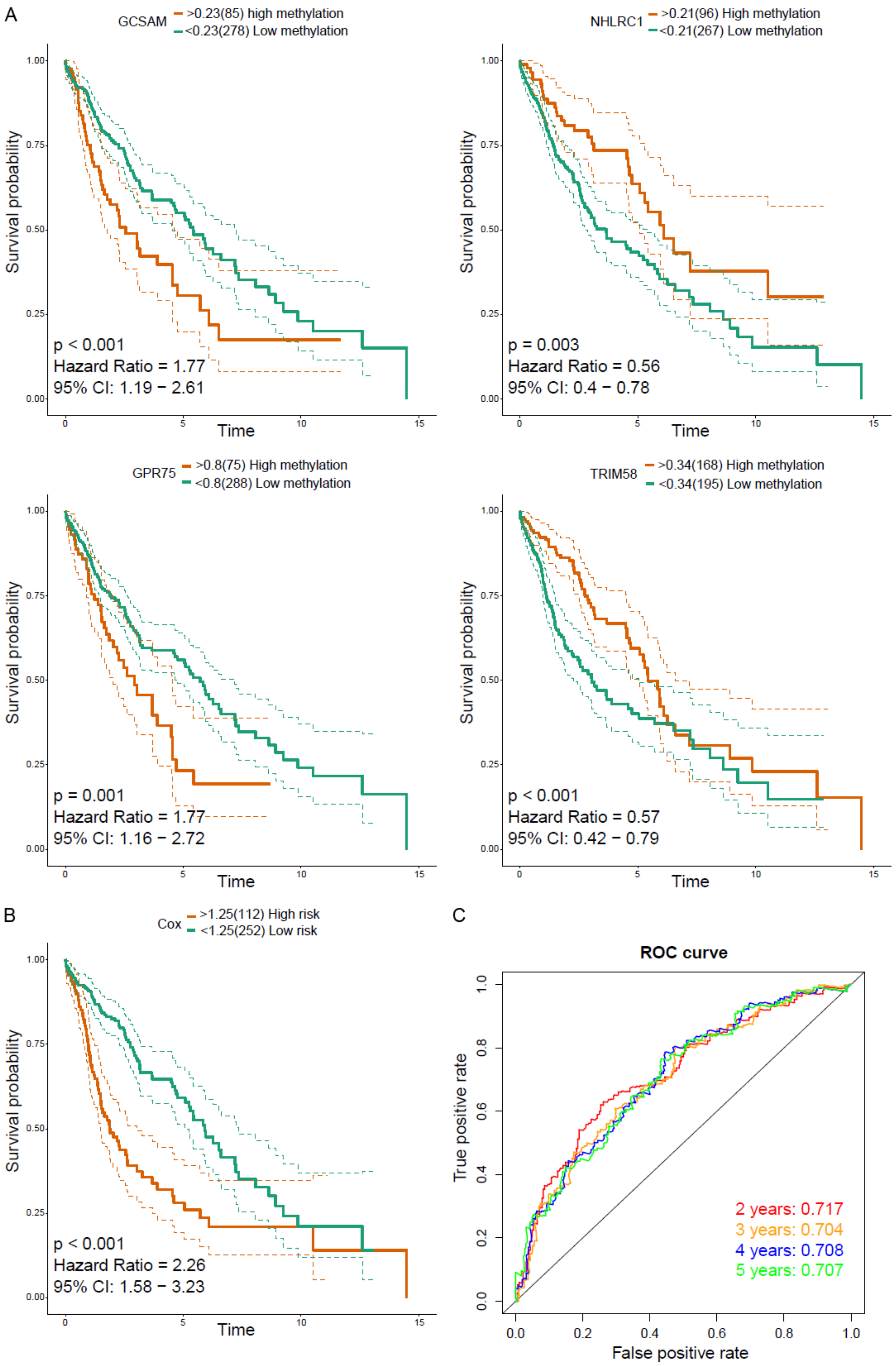

A
GCSAM
>0.23(85) high methylation
<0.23(278) Low methylation
Survival probability
1.00
0.75
0.50
0.25
0.00
p < 0.001
Hazard Ratio = 1.77
95% CI: 1.19 − 2.61
0
5
10
15
Time
NHLRC1
>0.21(96) High methylation
<0.21(267) Low methylation
Survival probability
p = 0.003
Hazard Ratio = 0.56
95% CI: 0.4 − 0.78
Time
GPR75
>0.8(75) High methylation
<0.8(288) Low methylation
Survival probability
p = 0.001
Hazard Ratio = 1.77
95% CI: 1.16 − 2.72
Time
TRIM58
>0.34(168) High methylation
<0.34(195) Low methylation
Survival probability
p < 0.001
Hazard Ratio = 0.57
95% CI: 0.42 − 0.79
Time
B
Cox
>1.25(112) High risk
<1.25(252) Low risk
Survival probability
p < 0.001
Hazard Ratio = 2.26
95% CI: 1.58 − 3.23
Time
C
ROC curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
2 years: 0.717
3 years: 0.704
4 years: 0.708
5 years: 0.707

Figure 6. Cox regression analysis showed four methylation-driven genes were significantly related patients' overall survival. A. Individual survival analysis of the four methylation-driven genes, GCSAM, NHLRC1, GPR75, and TRIM58, in lung SCC. The dashed line represents 95% confident interval (CI). The *p*-value, hazard ratio, and 95% CI of each gene is shown at the left lower corner of each plot. Green indicates low methylation and brown means high methylation. The optimal beta value separating two groups is displayed above. B. The Kaplan-Meier curve drawn using the four genes to divide patients into high- and low-risk groups, the high-risk group had shorted survival time compared to low-risk group (*p*-value < 0.001). C. ROC curve showing the suggested area under the curve for 2-, 3-, 4-, and 5 years.

Table 1. Univariate and multivariate analyses for the risk-score model

	Univariate analyses			Multivariable analyses		
	HR	95% CI	*p*-value	HR	95% CI	*p*-value
Gender	1.001	0.6923-1.448	0.995	1.3696	0.864064-2.1710	0.18079
Age	1.156	0.642-1.165	0.339	0.9818	0.667570-1.4440	0.9258
Race: Black or African American	1.3979	0.3173-6.159	0.658	0.5106	0.083615-3.1180	0.46654
Race: White	0.8536	0.2098-3.473	0.825	0.2533	0.047230-1.3586	0.10909
Stage IA	0.2867	0.08575-0.9586	0.0425*	0.1923	0.049213-0.7515	0.01775*
Stage IB	0.3991	0.12307-1.2940	0.1259	0.2461	0.065297-0.9274	0.03833*
Stage II	0.8914	0.14861-5.3469	0.8999	0.2801	0.020781-3.7761	0.33765
Stage IIA	0.369	0.10781-1.2631	0.1123	0.124	0.025754-0.5966	0.00921**
Stage IIB	0.4284	0.13019-1.4098	0.1631	0.1546	0.029807-0.8014	0.02618*
Stage III	3.6471	0.72258-18.4086	0.1172	0.5492	0.031587-9.5507	0.6809
Stage IIIA	0.6239	0.18852-2.0647	0.4397	0.1708	0.021925-1.3305	0.09153
Stage IIIB	0.5712	0.12643-2.5803	0.4666	0.0524	0.003278-0.8374	0.03703*
Stage IV	1.957	0.43636-8.7771	0.3805	1.1504	0.196732-6.7269	0.87644
T1a	2.071	0.8549-5.017	0.1068	2.5952	1.006607-6.6908	0.04843*
T1b	1.319	0.5663-3.074	0.52063	1.5796	0.641007-3.8927	0.32044
T2	1.664	0.9160-3.023	0.09454	1.1918	0.434448-3.2692	0.7333
T2a	1.784	0.9268-3.434	0.0832	1.3041	0.465897-3.6506	0.61312
T2b	2.377	1.0291-5.488	0.04265*	2.687	0.762900-9.4641	0.12388
T3	2.591	1.3495-4.974	0.00423**	1.6748	0.398057-7.0463	0.48179
T4	3.184	1.3723-7.389	0.00700**	5.3926	0.898577-32.3619	0.06533
N2	1.699	1.0100-2.857	0.0458*	1.7326	0.380119-7.8977	0.47759
Cigarettes	1.026	0.9272-1.134	0.624	1.0628	0.950816-1.1880	0.28358
Risk Score	1.793	1.471-2.186	7.57e-09***	1.7773	1.361954-2.3194	2.29e-05***

*: *p*-value < 0.05, **: *p*-value < 0.01, ***: *p*-value < 0.001.

value was set to -4.818, which corresponds to four genes, the model has the lowest deviance (Supplementary Figure 1A). Next, each of the four genes was assigned with a LASSO index with which the LASSO score of each patient can be calculated (LASSO score = 3.279008786 × methylation of GCSAM + 1.509542569 × methylation of GPR75 + -2.084452271 × methylation of NHLRC1 + -1.698699597 × methylation of TRIM58). With LASSO score, lung SCC patients were classified into high- and low-risk group using survminer package to find the optimal cut-off value. Consistent with our multivariate Cox model, there are 112 high-risk patients and 251 low-risk patients, and the former have significantly shorter survival time (Supplementary Figure 1C, *p*-value < 0.001). Decision curve analysis was performed to predict the net benefit that patients can receive with our prognostic model [18]. It is evident that patients can obtain the best net benefit from our LASSO model, which is also the combination of the four methylation-driven genes (Supplementary Figure 1D).

CpG sites of the four prognostic-related methylation-driven genes

To further investigate which parts of the four methylation-driven genes were methylated in lung SCC, we performed differentially methylation analysis of all CpG sites associated with

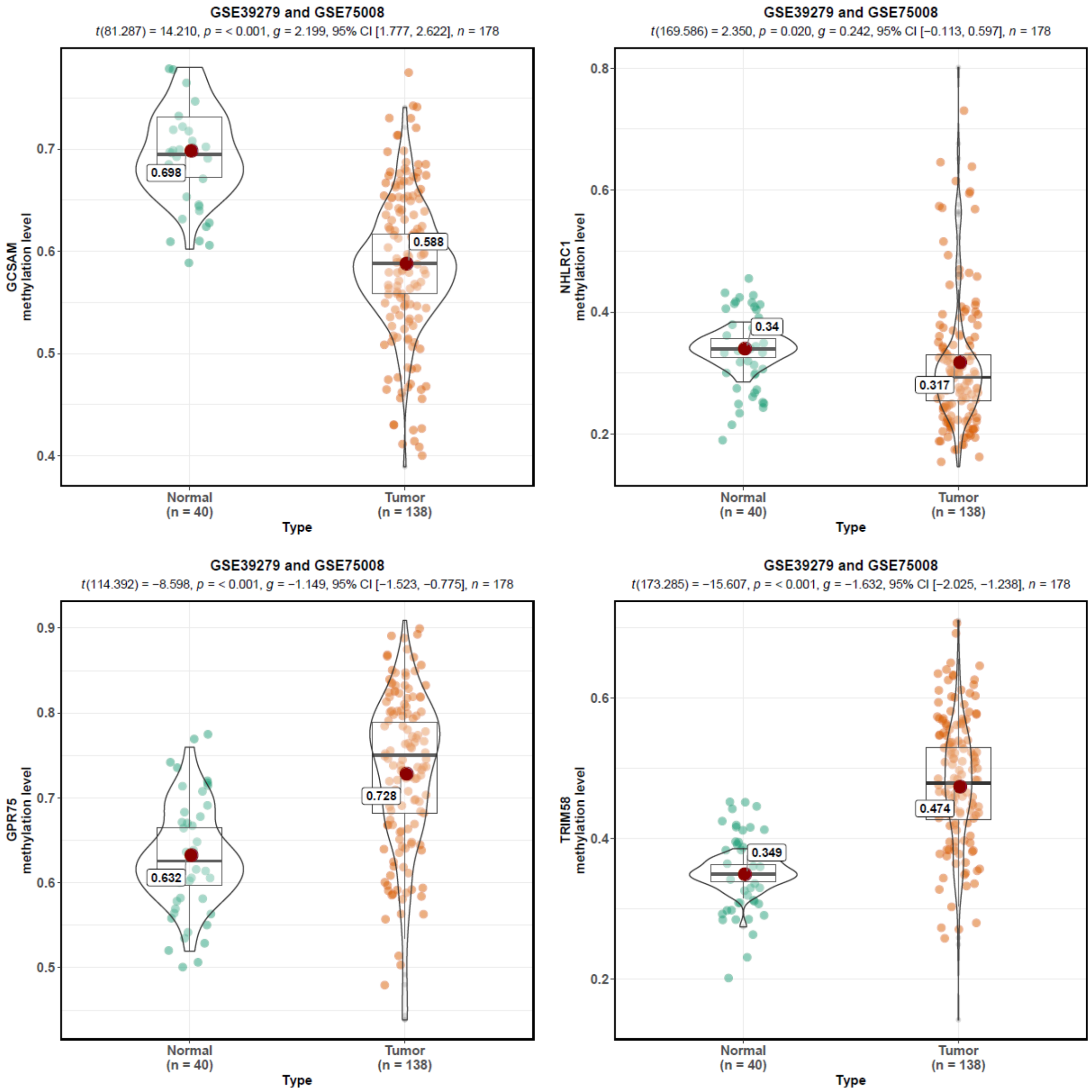


Figure 7. Methylation status of GCSAM, NHLRC1, GPR75 and TRIM58 in GEO datasets which contains 40 normal samples and 138 lung SCC tumor samples. GCSAM and NHLRC1 were hypomethylated, GPR75 and TRIM58 were hypermethylated in tumor samples of GEO data (*p*-value < 0.05).

these four genes. Validation of our results though MEXPRESS turned out to be consistent (**Figure 8** and Supplementary Table 5). In general, their methylation level alterations were inversely correlated with their gene expressions, but not associated with gender (*p*-value > 0.05). There are two methylation sites of GCSAM which were significantly related to the gene expression, cg06510261 is located in the promoter region, and its methylation level has a dramatic inverse relationship with the expression (r = -0.713). For GPR75, cg1959227 is the most predominant methylation site altered.

Discussion

All cancers arise as a result of changes that have occurred in somatic mutations, copy number alterations, DNA methylation, and gene expression in genomes. Using systematic studies of cancer genomes to guide the development and application of therapies in the clinic might be the most effective means for cancer treatment [19]. Mammalian DNA methylation plays critical roles in genomic imprinting, cell-fate determination, chromatin architecture organization, and regulation of gene expression.

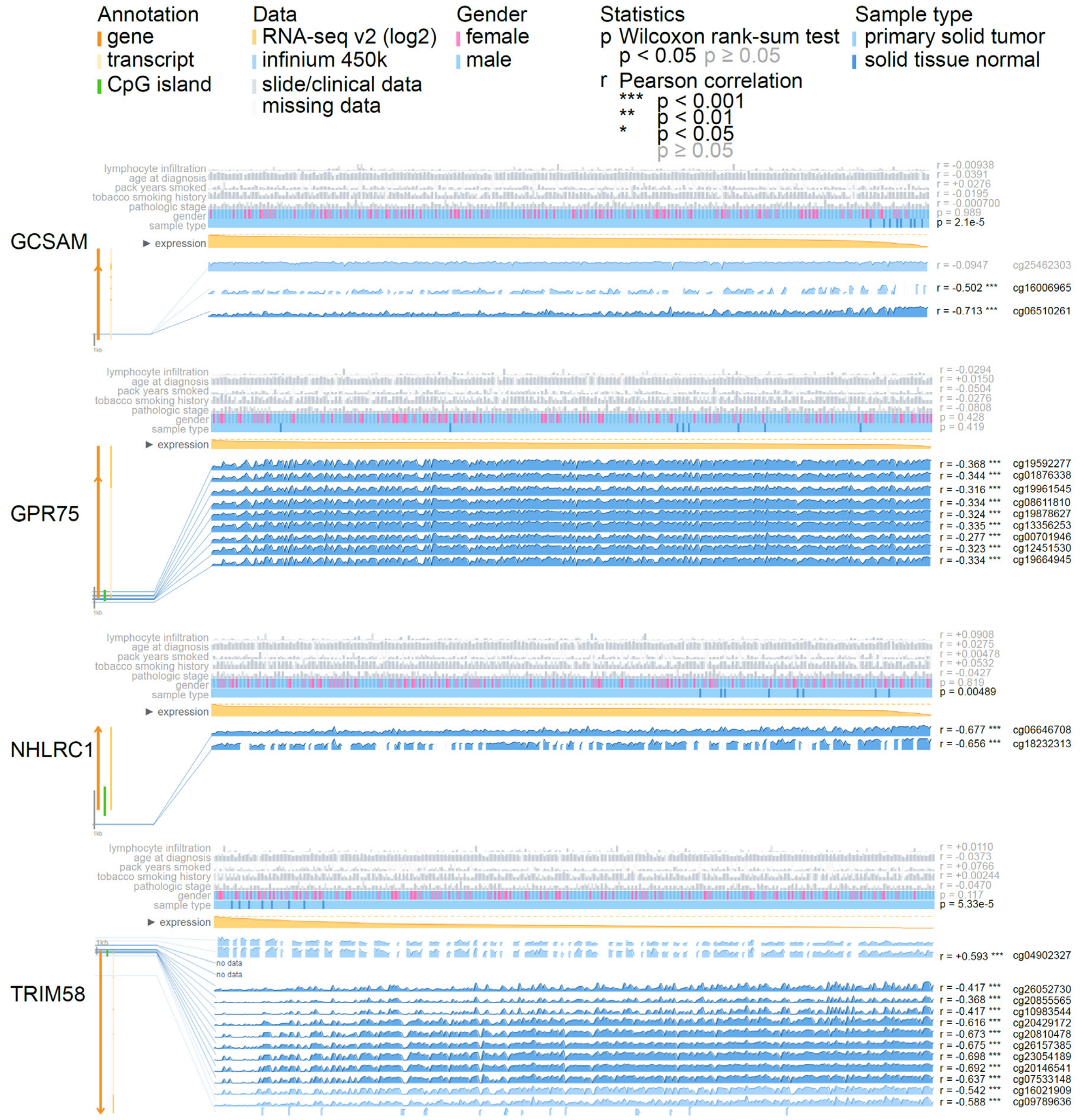


Figure 8. Detailed Methylation data of the four prognostic-related genes. From top to bottom, clinical data, the expression values, and the methylation data, the samples are ordered by their expression value. The arrow on the left indicates its direction. When the arrow points down, the gene is located on the + strand. If it points up, the gene lies on the - strand. Thin blue lines connect the probes to their respective genomic locations. The height of the blue lines indicates the beta value for a probe. The values on the far right represent the Pearson correlation coefficient between the methylation values for a probe and the expression values. Probes highlighted indicate they are located in a gene's promoter region.

Genetic studies have revealed abnormal DNA methylation in cancer disorders [20]. Chronic cigarette smoke-induced time-dependent epigenetic alterations are associated with abnormal DNA methylation, which may prime for changing oncogene senescence to addiction for a single key oncogene involved in lung cancer initiation [21]. Previous studies showed that cancer cells have a global decrease in DNA methylation, but at specific genes, DNA methylation is increased in association with silencing of genes that control cell growth, often including tumor suppressors [22]. Studies from TCGA have generated comprehensive catalogs of cancer genes involved in tumorigenesis across a broad range of cancer types [19].

Here, we found that lung SCC has a global decrease in DNA methylation compared to normal control. However, there was no significant difference among tumor stage I, tumor stage II, tumor stage III and IV.

Several previous studies have compared methylation patterns of chromosome arms in tumors. Epigenetic deregulation across chromosome 2q14.2 provides a regional panel of novel DNA methylation cancer biomarkers [23]. In our present studies, we characterized CpG island methylation and methylation patterns of entire chromosome arms with chromosome mapping. Dys-methylated sites at CpGs in lung SCC are mainly represented on chromosomes 2 and 7 containing the highest number of dys-methylated sites in lung SCC, of 1825 and 1716 sites respectively. The Y-chromosome is the sex-determining chromosome in many species and plays a critical role in tumor suppression. Studies by Yao L et al. indicated that aberrant DNA methylation on the Y-chromosome could serve as a potential diagnostic biomarker with high sensitivity and specificity in prostate cancer [24]. Furthermore, we found chromosome Y contains the least dys-methylated sites in lung SCC, which may be associated with Y chromosome inactivation.

Many previous studies on aberrant methylation in lung SCC merely focused on epigenetic alterations without excavating their effects on expression. Therefore, it is entirely necessary that we distinguish the epigenetic changes that act as effectors from “passenger” alterations with no biologic effect. Thus, we used MethylMiX to identify those genes with aberrant methylation and linked these data to transcriptome data reflecting gene expression. It has been shown that MethylMix is capable of generating consistent results with other tools but also to produce novel and unique findings [25].

In our study, we identified a total of 182 methylation-driven genes. Some of these genes have been previously reported to be candidates in a variety of diseases, but there are a certain number of genes to be considered associated with lung SCC for the first time. Pathway and function analyses allow us to define methylation-driven genes to cancer-associated pathways. A pathway-centric view highlighted the methylation-driven genes are involved in many relevant oncologic pathways, including hydroxyl carboxylic acid-binding receptors, generic transcription pathway, and RNA Polymerase II transcription, and gene expression, which are associated with canonical cancer-associated pathways. Hydroxyl carboxylic acid receptors respond to organic acids. There is evidence that these receptors can mediate anti-inflammatory effects [26]. Merlo A et al. found that 5’ CpG island methylation is associated with transcriptional silencing of tumor suppressor in human cancers [27]. Sanford T et al. reported that differential methylation exhibited better or worse prognosis after in patients with cystectomy [28]. Guo D et al. demonstrated that low expression resulting from hypermethylation of the tumor suppressor gene Kelch-like ECH-associating protein 1 (Keap1) promoter abrogates binding of the transcription factor Sp1 in lung cancer cells. Hypermethylation of DNA binding transcription factor may be a possible gene silencing mechanism [29]. Here, GO analysis showed hypermethylated genes were involved DNA binding transcription factor activity, RNA biosynthetic process, regulation of RNA metabolic process, regulation of biosynthetic process, and regulation of gene expression in lung SCC. RNA biosynthetic and RNA metabolic process involved in the stability of RNAs, which represents a crucial point for cell death and aging [30]. Yu Q et al. reported that miR-142 hypermethylation promotes TGF-β-mediated tumor growth and metastasis through loss-of-function of miR-142 in hepatocellular carcinoma [31]. Cornification, a basic types of cell death, is distinguished from necrosis, apoptosis, and autophagy [32]. Targeting programmed cell death provides a series of possible targets in cancer therapy [33]. The cytoskeleton is involved in cell adhesion and cell cycle progression. The cytoskeleton plays an essential role in tumor invasion and metastasis [34]. Notch signaling engaged in the development and cell fate determination, and it is deregulated in solid tumors [35]. In our present studies, GO analysis indicated hypomethylated genes were involved in cornification, intermediate filament, Notch signaling pathway, cytoskeleton, epidermis development, cell differentiation.

Even though lung SCC is a devasting cancer subtype, prognostic tools for lung SCC are limited. If we could precisely predict the tumor behavior in the initial stage, the prognosis of patients with lung SCC would be dramatically improved. In the current study, we have developed a relatively accurate risk score model for lung SCC prognosis prediction through a comprehensive survival analysis of the prognostic signature. Patients can be divided into high- and low-risk groups according to their risk scores, and the clinical outcome of lung SCC patients was significantly different between high- and low-risk groups. ROC analysis also suggested that this risk score model has relatively good accuracy. GCSAM (Germinal center-associated signaling and motility protein) is reported to be associated with regulation of lymphocyte motility in lymphoma, but it also functions in the regulation of kinase activation and B-cell signaling, indicating its potential role in lung SCC [36]. GPR75 is a G protein-coupled receptor and may play a role in the signaling pathway involving the PI3, Akt and MAP kinases, previous studies have identified it could be a novel methylated gene in colorectal cancer [37, 38]. NHLRC1 is a kind of ubiquitin-protein ligase and is involved in the pathway protein ubiquitination [39]. As for TRIM58, studies have shown that its methylation was significantly related to many prognostic genes in lung SCC [40]. As a result, the methylation level of these four genes may have potential prognostic and therapeutic significance in lung SCC.

Our current study has identified numerous novel and unique methylation-driven genes, which could have potential value in lung SCC diagnosis and prognosis. Validation on external datasets showed consistent methylation alterations of GCSAM, GPR75, NHLRC1, and TRIM58, which may suggest their values in potential clinical applications.

In summary, our results indicate lung SCC exhibited a global decrease in DNA methylation. Methylation-driven genes are associated with cancer-associated pathways. We further developed a relative accurate risk score model for lung SCC prognosis prediction through a comprehensive survival analysis of the prognostic signature. Moreover, we found four methylation-driven genes, GCSAM, GPR75, NHLRC1, and TRIM58, as independent prognostic indicators.

Acknowledgements

The study was supported by Outstanding study plan of Fudan University.

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Chunlai Lu, Department of Thoracic Surgery, Zhongshan Hospital, Fudan University, 180 Fenglin Road, Shanghai 200032, P. R. China. Tel: +86 021 64041990-2559; Fax: +86 021 64041990-2559; E-mail: lu.chunlai@zs-hospital.sh.cn

References

[1] Campbell JD, Lathan C, Sholl L, Ducar M, Vega M, Sunkavalli A, Lin L, Hanna M, Schubert L, Thorner A, Faris N, Williams DR, Osarogiagbon RU, van Hummelen P, Meyerson M and MacConaill L. Comparison of prevalence and types of mutations in lung cancers among black and white populations. JAMA Oncol 2017; 3: 801-809.

[2] Miller KD, Siegel RL, Lin CC, Mariotto AB, Kramer JL, Rowland JH, Stein KD, Alteri R and Jemal A. Cancer treatment and survivorship statistics, 2016. CA Cancer J Clin 2016; 66: 271-289.

[3] Piperdi B, Merla A and Perez-Soler R. Targeting angiogenesis in squamous non-small cell lung cancer. Drugs 2014; 74: 403-413.

[4] Aberle DR, Abtin F and Brown K. Computed tomography screening for lung cancer: has it finally arrived? Implications of the national lung screening trial. J Clin Oncol 2013; 31: 1002-1008.

[5] Pu W, Geng X, Chen S, Tan L, Tan Y, Wang A, Lu Z, Guo S, Chen X and Wang J. Aberrant methylation of CDH13 can be a diagnostic biomarker for lung adenocarcinoma. J Cancer 2016; 7: 2280-2289.

[6] Kim KY, Tanaka Y, Su J, Cakir B, Xiang Y, Patterson B, Ding J, Jung YW, Kim JH, Hysolli E, Lee H, Dajani R, Kim J, Zhong M, Lee JH, Skalnik D, Lim JM, Sullivan GJ, Wang J and Park IH. Uhrf1 regulates active transcriptional marks at bivalent domains in pluripotent stem cells through Setd1a. Nat Commun 2018; 9: 2583.

[7] Yang C, Zhang Y, Song Y, Lu X and Gao H. Genome-wide DNA methylation analysis of the regenerative and non-regenerative tissues in sika deer (Cervus nippon). Gene 2018; 676: 249-255.

[8] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W and Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015; 43: e47.

[9] Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM,

Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G and Noushmehr H. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res 2016; 44: e71.

[10] McCarthy DJ, Chen Y and Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res 2012; 40: 4288-4297.

[11] Cedoz PL, Prunello M, Brennan K and Gevaert O. MethylMix 2.0: an R package for identifying DNA methylation genes. Bioinformatics 2018; 34: 3044-3046.

[12] Marwitz S, Depner S, Dvornikov D, Merkle R, Szczygieł M, Müller-Decker K, Lucarelli P, Wäsch M, Mairbäurl H, Rabe KF, Kugler C, Vollmer E, Reck M, Scheufele S, Kröger M, Ammerpohl O, Siebert R, Goldmann T and Klingmüller U. Downregulation of the TGFβ pseudoreceptor BAMBI in non-small cell lung cancer enhances TGFβ signaling and invasion. Cancer Res 2016; 76: 3785-3801.

[13] Sandoval J, Mendez-Gonzalez J, Nadal E, Chen G, Carmona FJ, Sayols S, Moran S, Heyn H, Vizoso M, Gomez A, Sanchez-Cespedes M, Assenov Y, Muller F, Bock C, Taron M, Mora J, Muscarella LA, Liloglou T, Davies M, Pollan M, Pajares MJ, Torre W, Montuenga LM, Brambilla E, Field JK, Roz L, Lo Iacono M, Scagliotti GV, Rosell R, Beer DG and Esteller M. A prognostic DNA methylation signature for stage I non-small-cell lung cancer. J Clin Oncol 2013; 31: 4140-4147.

[14] Leek JT, Johnson WE, Parker HS, Jaffe AE and Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 2012; 28: 882-883.

[15] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 1996; 58: 267-288.

[16] Qiu J, Peng B, Tang Y, Qian Y, Guo P, Li M, Luo J, Chen B, Tang H, Lu C, Cai M, Ke Z, He W, Zheng Y, Xie D, Li B and Yuan Y. CpG Methylation signature predicts recurrence in early-stage hepatocellular carcinoma: results from a multicenter study. J Clin Oncol 2017; 35: 734-742.

[17] Thaper D, Vahid S, Nip KM, Moskalev I, Shan X, Frees S, Roberts ME, Ketola K, Harder KW, Gregory-Evans C, Bishop JL and Zoubeidi A. Targeting Lyn regulates Snail family shuttling and inhibits metastasis. Oncogene 2017; 36: 3964-3975.

[18] Vickers AJ and Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006; 26: 565-574.

[19] Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, Aben N, Goncalves E, Barthorpe S, Lightfoot H, Cokelaer T, Greninger P, van Dyk E, Chang H, de Silva H, Heyn H, Deng X, Egan RK, Liu Q, Mironenko T, Mitropoulos X, Richardson L, Wang J, Zhang T, Moran S, Sayols S, Soleimani M, Tamborero D, Lopez-Bigas N, Ross-Macdonald P, Esteller M, Gray NS, Haber DA, Stratton MR, Benes CH, Wessels LFA, Saez-Rodriguez J, McDermott U and Garnett MJ. A landscape of pharmacogenomic interactions in cancer. Cell 2016; 166: 740-754.

[20] Liu XS, Wu H, Ji X, Stelzer Y, Wu X, Czauderna S, Shu J, Dadon D, Young RA and Jaenisch R. Editing DNA methylation in the mammalian genome. Cell 2016; 167: 233-247 e217.

[21] Vaz M, Hwang SY, Kagiampakis I, Phallen J, Patil A, O'Hagan HM, Murphy L, Zahnow CA, Gabrielson E, Velculescu VE, Easwaran HP and Baylin SB. Chronic cigarette smoke-induced epigenomic changes precede sensitization of bronchial epithelial cells to single-step transformation by KRAS mutations. Cancer Cell 2017; 32: 360-376, e6.

[22] Licht JD. DNA methylation inhibitors in cancer therapy: the immunity dimension. Cell 2015; 162: 938-939.

[23] Devaney J, Stirzaker C, Qu W, Song JZ, Statham AL, Patterson KI, Horvath LG, Tabor B, Coolen MW, Hulf T, Kench JG, Henshall SM, Pe Benito R, Haynes AM, Mayor R, Peinado MA, Sutherland RL and Clark SJ. Epigenetic deregulation across chromosome 2q14.2 differentiates normal from prostate cancer and provides a regional panel of novel DNA methylation cancer biomarkers. Cancer Epidemiol Biomarkers Prev 2011; 20: 148-159.

[24] Yao L, Ren S, Zhang M, Du F, Zhu Y, Yu H, Zhang C, Li X, Yang C, Liu H, Wang D, Meng H, Chang S, Han X, Sun Y and Sun Y. Identification of specific DNA methylation sites on the Y-chromosome as biomarker in prostate cancer. Oncotarget 2015; 6: 40611-40621.

[25] Gevaert O, Tibshirani R and Plevritis SK. Pancancer analysis of DNA methylation-driven genes using MethylMix. Genome Biol 2015; 16: 17.

[26] Graff EC, Fang H, Wanders D and Judd RL. Anti-inflammatory effects of the hydroxycarboxylic acid receptor 2. Metabolism 2016; 65: 102-113.

[27] Merlo A, Herman JG, Mao L, Lee DJ, Gabrielson E, Burger PC, Baylin SB and Sidransky D. 5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. Nat Med 1995; 1: 686-692.

[28] Sanford T, Meng MV, Railkar R, Agarwal PK and Porten SP. Integrative analysis of the epigenetic basis of muscle-invasive urothelial carcinoma. Clin Epigenetics 2018; 10: 19.

[29] Guo D, Wu B, Yan J, Li X, Sun H and Zhou D. A possible gene silencing mechanism: hypermethylation of the Keap1 promoter abrogates binding of the transcription factor Sp1 in lung cancer cells. Biochem Biophys Res Commun 2012; 428: 80-85.

[30] Falcone C and Mazzoni C. RNA stability and metabolism in regulated cell death, aging and diseases. FEMS Yeast Res 2018; 18.

[31] Yu Q, Xiang L, Yin L, Liu X, Yang D and Zhou J. Loss-of-function of miR-142 by hypermethylation promotes TGF-beta-mediated tumour growth and metastasis in hepatocellular carcinoma. Cell Prolif 2017; 50.

[32] Toton E, Lisiak N, Sawicka P and Rybczynska M. Beclin-1 and its role as a target for anticancer therapy. J Physiol Pharmacol 2014; 65: 459-467.

[33] Ke B, Tian M, Li J, Liu B and He G. Targeting programmed cell death using small-molecule compounds to improve potential cancer therapy. Med Res Rev 2016; 36: 983-1035.

[34] Ntantie E, Fletcher J, Amissah F, Salako OO, Nkembo AT, Poku RA, Ikpatt FO and Lamango NS. Polyisoprenylated cysteinyl amide inhibitors disrupt actin cytoskeleton organization, induce cell rounding and block migration of non-small cell lung cancer. Oncotarget 2017; 8: 31726-31744.

[35] Takebe N, Nguyen D and Yang SX. Targeting notch signaling pathway in cancer: clinical development advances and challenges. Pharmacol Ther 2014; 141: 140-149.

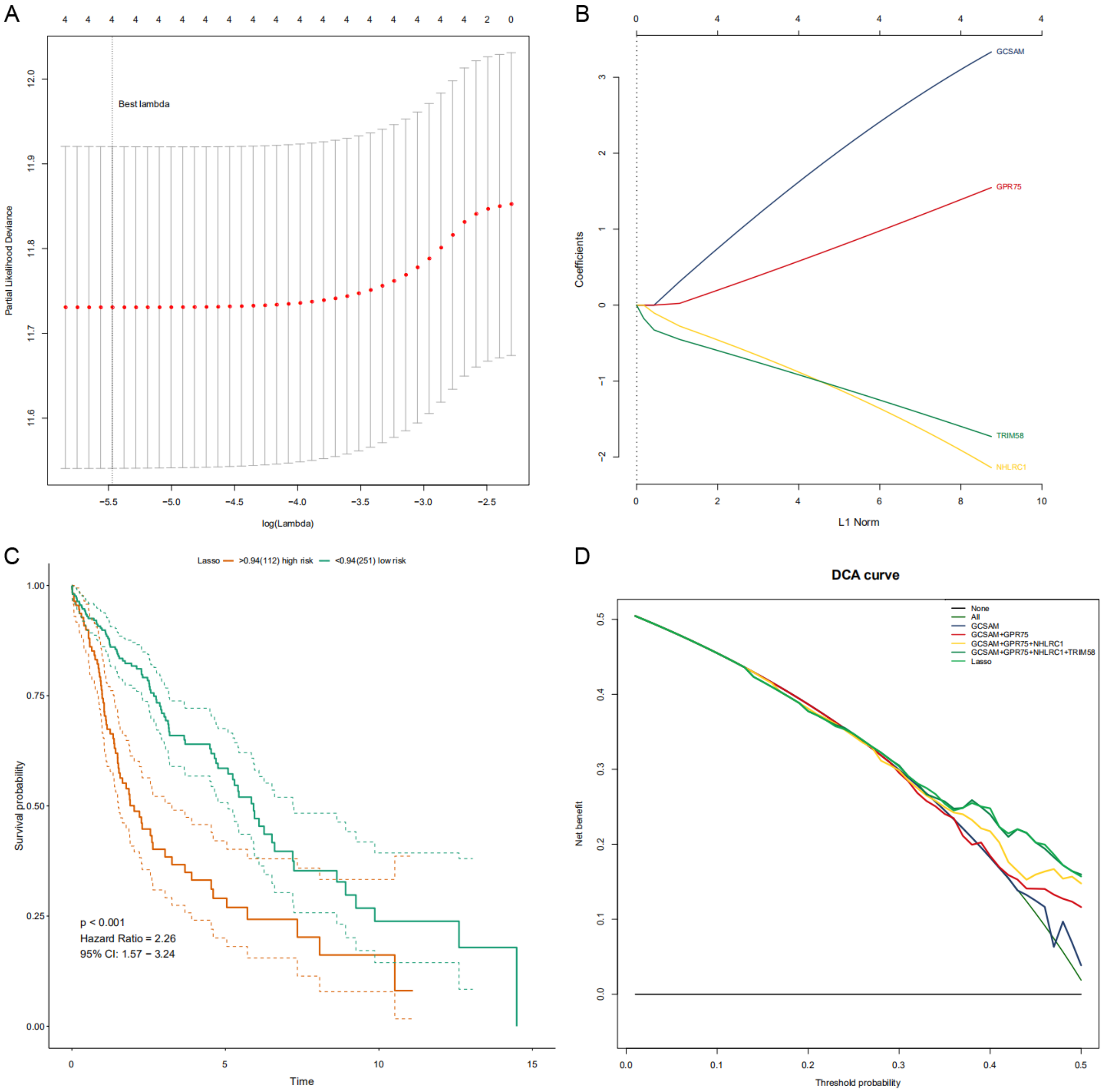
[36] Lu X, Chen J, Malumbres R, Cubedo Gil E, Helfman DM and Lossos IS. HGAL, a lymphoma prognostic biomarker, interacts with the cytoskeleton and mediates the effects of IL-6 on cell migration. Blood 2007; 110: 4268-4277.

[37] Liu B, Hassan Z, Amisten S, King AJ, Bowe JE, Huang GC, Jones PM and Persaud SJ. The novel chemokine receptor, G-protein-coupled receptor 75, is expressed by islets and is coupled to stimulation of insulin secretion and improved glucose homeostasis. Diabetologia 2013; 56: 2467-2476.

[38] Ashktorab H, Daremipouran M, Goel A, Varma S, Leavitt R, Sun X and Brim H. DNA methylome profiling identifies novel methylated genes in African American patients with colorectal neoplasia. Epigenetics 2014; 9: 503-512.

[39] Liu W, Duan X, Fang X, Shang W and Tong C. Mitochondrial protein import regulates cytosolic protein homeostasis and neuronal integrity. Autophagy 2018; 14: 1293-1309.

[40] Zhang W, Cui Q, Qu W, Ding X, Jiang D and Liu H. TRIM58/cg26157385 methylation is associated with eight prognostic genes in lung squamous cell carcinoma. Oncol Rep 2018; 40: 206-216.



Supplementary Figure 1. Validation on the prognostic model using LASSO Cox regression analysis. A. Partial likelihood deviance for the LASSO coefficient profiles. A dashed vertical. line stands for the minimum partial likelihood deviance (logγ = -4.818). B. LASSO coefficient profiles of the four methylation-driven genes. A dashed vertical line is drawn atthe value (γ = 0.006708057) chosen by 10-fold cross-validation. C. The survival curve using the four genes to divide patients into high- and low-risk groups, the high-risk group had shorted survival time compared to low-risk group (*p*-value < 0.001). D. DCA curve showing the combination of the four genes can provide more benefits for patients' survival prediction.