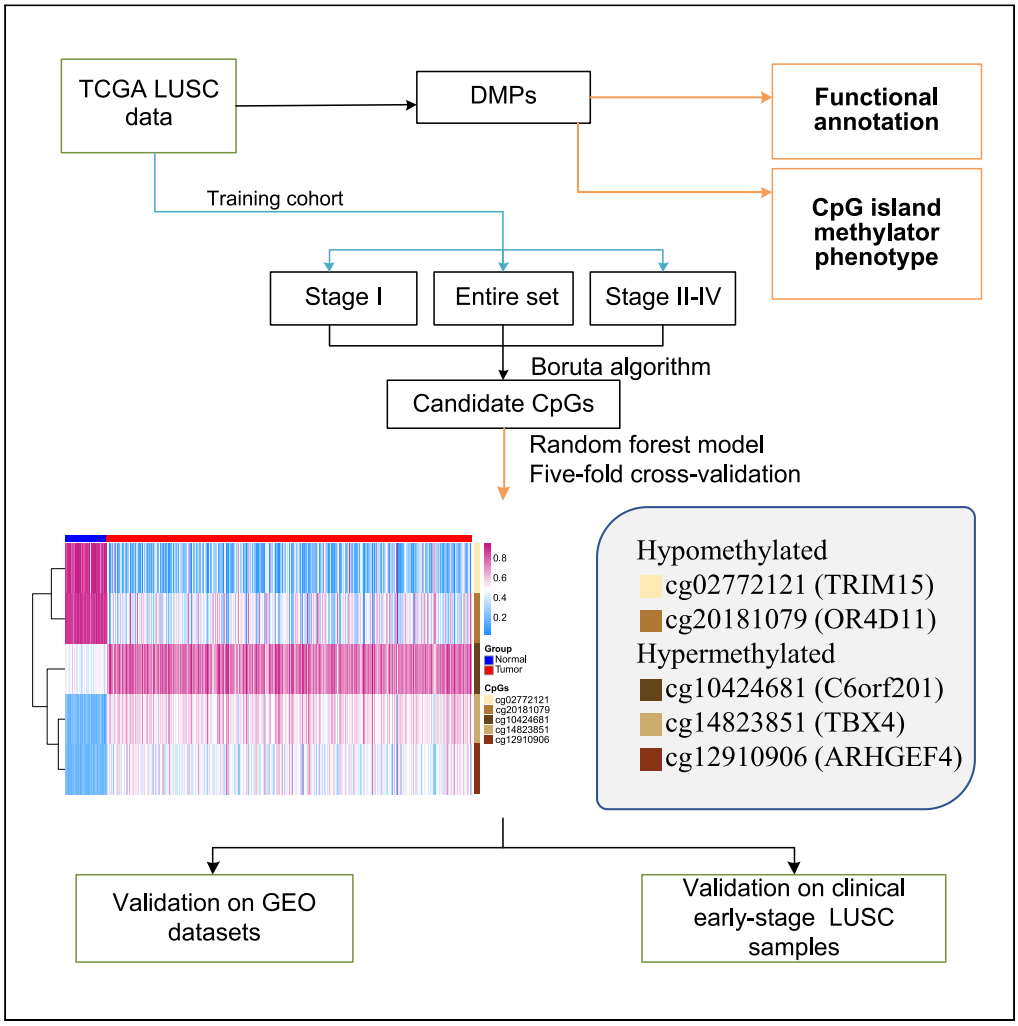
iScience

CellPress
OPEN ACCESS

Article

Whole-genome DNA methylation and DNA methylation-based biomarkers in lung squamous cell carcinoma



Qidong Cai, Boxue He, Guangxu Tu, ..., Shaoliang Peng, Yongguang Tao, Xiang Wang

wangxiang@csu.edu.cn

Highlights

Comprehensively evaluate DNA methylation for lung squamous cell carcinoma (LUSC)

Five methylation biomarkers along with mapped genes are identified in LUSC

Aberrant methylation biomarkers also show in lung progressive CIS lesions

Cai et al., iScience 26, 107013
July 21, 2023
https://doi.org/10.1016/j.isci.2023.107013

iScience

CellPress
OPEN ACCESS

Article

Whole-genome DNA methylation and DNA methylation-based biomarkers in lung squamous cell carcinoma

Qidong Cai,[1,2,9] Boxue He,[1,2,9] Guangxu Tu,[1,2] Weilin Peng,[1,2] Shuai Shi,[1,2] Banglun Qian,[1,2] Qingchun Liang,[3] Shaoliang Peng,[4,5,6] Yongguang Tao,[1,2,7,8] and Xiang Wang[1,2,10,*]

SUMMARY

Exploring early detection methods through comprehensive evaluation of DNA methylation for lung squamous cell carcinoma (LUSC) patients is of great significance. By using different machine learning algorithms for feature selection and model construction based on The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) databases, five methylation biomarkers in LUSC (along with mapped genes) were identified including cg14823851 (TBX4), cg02772121 (TRIM15), cg10424681 (C6orf201), cg12910906 (ARHGEF4), and cg20181079 (OR4D11), achieving extremely high sensitivity and specificity in distinguishing LUSC from normal samples in independent cohorts. Pyrosequencing assay verified DNA methylation levels, meanwhile qRT-PCR and immunohistochemistry results presented their accordant methylation-related gene expression statuses in paired LUSC and normal lung tissues. The five methylation-based biomarkers proposed in this study have great potential for the diagnosis of LUSC and could guide studies in methylation-regulated tumor development and progression.

INTRODUCTION

Lung cancer has over 2.2 million new cases and caused about 1.8 million deaths in 2020, taking a major part in the global cancer burden.[1] More than 80% of primary lung carcinoma is non-small cell lung cancer (NSCLC), which has two predominant subtypes pathologically—lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC).[2] Although a lot of significant advancements have been achieved for the treatment, most patients with LUSC still cannot get diagnosed at the early stage and have an unsatisfactory 5-year survival rate.[3] Nowadays, high-resolution computed tomography (HRCT) is the dominant screening method for lung cancer, which has dramatically reduced NSCLC-related mortality rates through early detection. However, many problems, including a high false-positive rate, patient fear of radiation, excessive cost, the requirement for long-term follow-up, and the risk of overdiagnosis,[4] have started to emerge. The development of dynamic determination of early tumor markers combined with imaging methods may bring new hope for early detection and early diagnosis of LUSC. The serum proteins carcinoembryonic antigen (CEA) and neuron-specific enolase (NSE) are widely used for assisting early NSCLC diagnosis. However, the diagnostic performances of CEA and NSE are unsatisfactory (sensitivities of 26% and 21–39%, respectively).[5] Advancing early detection methods, mining new diagnosis biomarkers, and understanding the progression mechanisms of LUSC are of great clinical significance.

DNA methylation is a kind of epigenetic modification which plays significant roles in gene regulation and genome stability.[6] In mammals, DNA methylation principally appears in the context of 5′—Cytosine—phosphate—Guanine—3′ (CpG) dinucleotides and was also often referred to as "CpG methylation". Research about DNA methylation and demethylation has provided some nuanced understanding of human diseases, like cancer.[7] It has been proved by methylome profiles that different cancer subtypes have nonrandom DNA methylation nature. In consideration of the advantages such as cell-type specificity, occurring frequently and stably, and being easily detected even by circulating tumor DNA, DNA methylation aroused a lot of interest with the potential to be cancer diagnostic biomarkers.[8] However, although thousands of scientific articles provided DNA methylation-based biomarkers for cancer diagnosis, only a few of them were eventually considered for clinical utilization, majorly because of obstacles methodologically or experimentally.[9] Hence, we are in sore need of easily implement

[1]Department of Thoracic Surgery, Second Xiangya Hospital, Central South University, Changsha 410011, China

[2]Hunan Key Laboratory of Early Diagnosis and Precise Treatment of Lung Cancer, Second Xiangya Hospital, Central South University, Changsha 410011, China

[3]Department of Pathology, Second Xiangya Hospital, Central South University, Changsha 410011, China

[4]College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China

[5]School of Computer Science, National University of Defense Technology, Changsha 410073, China

[6]Peng Cheng Lab, Shenzhen 518000, China

[7]Key Laboratory of Carcinogenesis and Cancer Invasion, Ministry of Education, Department of Pathology, Xiangya Hospital, Central South University, Hunan 410078, China

[8]NHC Key Laboratory of Carcinogenesis (Central South University), Cancer Research Institute and School of Basic Medicine, Central South University, Changsha, Hunan 410078, China

[9]These authors contributed equally

[10]Lead contact

*Correspondence: wangxiang@csu.edu.cn

https://doi.org/10.1016/j.isci.2023.107013

DNA methylation-based models with strong diagnostic performance. Besides, DNA methylation-correlated gene expression was often observed in cancer cells, like many silenced tumor-suppressor genes along with hypermethylation, as well as some activated repetitive elements with hypomethylation,[10,11] suggesting underlying regulation mechanisms to be found. Previous studies also expounded that the CpG island (CGI) methylator phenotype (CIMP), which refers to a subgroup of tumor patients with extensive hypermethylation of a large number of CGIs in the gene promoter region, has a close association with many clinicopathological characteristics in multiple cancers,[12] while the CIMP-related gene mutations and their impact on prognosis have rarely been studied in LUSC.

Actually, recently several publications have introduced diagnostic methylation biomarkers for NSCLC patients,[13,14] and the majority of them were concerned with LUAD,[15,16] while methylation-based diagnostic models specially for LUSC and comprehensive analysis for LUSC methylome data were rarely provided. In this study, we analyzed methylation profiles in the The Cancer Genome Atlas (TCGA)-LUSC dataset and explored differentially methylated positions (DMPs) of LUSC and normal lung tissues, based on which we made functional annotations and scanned CIMP of LUSC. Then, utilizing random forest-based approaches, we discriminated important CpGs as candidate biomarkers for LUSC diagnosis, followed by random forest-based approaches to determine a panel of CpGs with the smallest average error rate in the nested 5-fold cross-validation. Next, test cohorts from the Gene Expression Omnibus (GEO) databases and DNA pyrosequencing assays of patients in our department were introduced for verification of our diagnostic model. What's more, we identified the expression of mapped genes by qRT-PCR and protein immunohistochemistry (IHC).

RESULTS

Integrated methylation profiles of LUSC

After differential analysis, 12,958 hypermethylated DMPs were identified, 4,650 of which were located in the intergenic region (IGR), and the remaining 8,308 DMPs corresponded to 2,489 genes. 20,917 hypomethylated DMPs were identified, 7,672 of which were located in IGR, and the remaining 13,245 DMPs corresponded to 5,254 genes. As shown in Figure 1B, the number of hypermethylated DMPs was highest in chromosome 2 (1,311 CpGs, 5.7%), followed by chromosome 1 (1,239 CpGs, 5.4%), while it was lowest in chromosome 21 (91 CpGs, 0.4%) and chromosome 22 (121 CpGs, 0.5%); the number of hypomethylated DMPs was highest in chromosome 1 (1,969 CpGs, 7.7%) and chromosome 7 (1,781 CpGs, 6.9%), while it was lowest in chromosome 18 (228 CpGs, 0.9%) and chromosome 21 (155 CpGs, 0.6%). Besides, 81.9% DMPs were hypermethylated in the CGIs, and 93.7% DMPs were hypermethylated in the promoter area of CGI (Figure 1C). As the probes were far away from the CGI,[17,18] the percentage of hypermethylated DMPs gradually decreased (Figure 1D). Gene body and IGR were dominant by hypomethylation, while DMPs in transcription start site (TSS) 200 and the first exons were more frequently hypermethylated (Figure 1E). The genome-wide hypomethylation and promoter CGI hypermethylation have been recognized in a variety of tumors,[19] and the methylation characteristic in LUSC was consistent with that.

Functional annotation

Under the screening criteria, 303 hypermethylated and 463 hypomethylated genes were selected for Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome functional enrichment analyses, respectively. The results demonstrated that LUSC hypermethylated genes were mainly enriched in pathways including "regulation of TP53 activity through association with co-factors", "activation of anterior HOX genes in hindbrain development during early embryogenesis", "transcriptional misregulation in cancer", "SUMOylation of transcription factors", etc (Figure S1A). And, LUSC hypomethylated genes were mainly enriched in pathways including "cell-cell communication", "GPCR downstream signaling", "calcium signaling pathway", and "inflammatory mediator regulation of TRP channels" (Figure S1B).

CIMP

A total of 2,120 most variable CpGs were included in further analysis. After excluding the samples with missing data, a total of 337 patients were used for consistency clustering and 3 clusters were obtained (Figure 2A). Compared to cluster 1 (N = 156) and cluster 3 (N = 67), cluster 2 (N = 114) showed a global hypermethylated status, and its methylation levels were significantly higher than others, which was preliminarily identified as CIMP (Figures 2B and 2C). We next analyzed the association between the clusters and clinical characteristics, in which American Joint Committee on Cancer (AJCC) stage, TP53 mutation, KEAP1 mutation, CDKN2A mutation, and KMT2D mutation were significantly associated with clusters (Table 1).

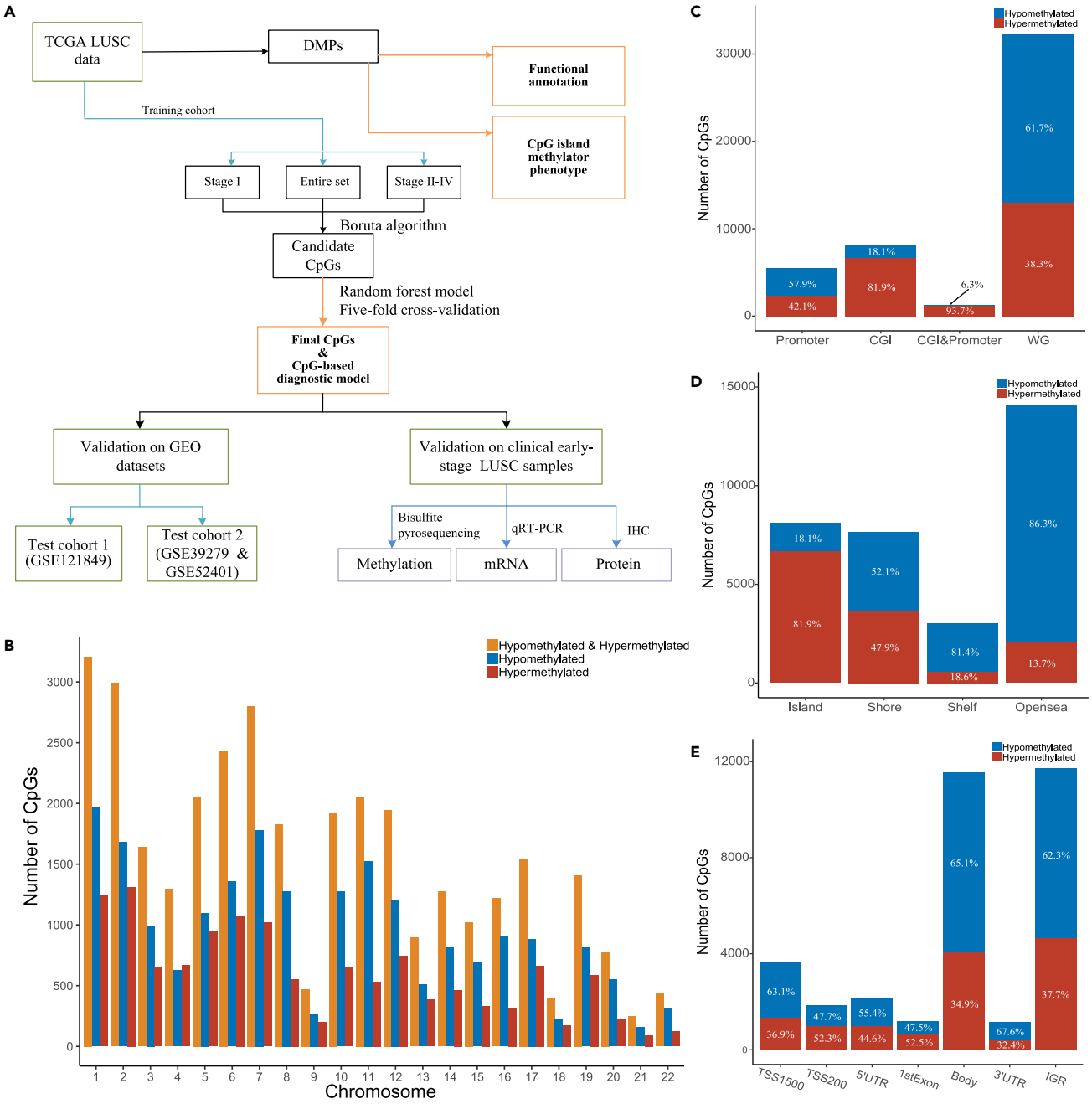


Figure 1. Scheme of this article and distribution characteristics of DMPs in LUSC

(A) Scheme showing the bioinformatic analysis and the experiment process.

(B) Number distribution of DMPs on autosomes.

(C) Distribution of DMPs across regions of the genome. CGI, CpG islands; WG, whole genome.

(D) Distribution of DMPs by distance from CGI. From left to right, the shore (regions up to 2 kb from CpG island), shelf (regions up to 2 to 4 kb from CpG island), and opensea areas (the rest of the genome) are with increasing distance from the CGI.

(E) Distribution of DMPs by distance from TSS. The 5′UTR, 1stExon, Body, 3′UTR, and IGR areas have increasing distances from the TSS. TSS, transcription start site.

Furthermore, we compared the overall survival (OS) of CIMP patients (cluster 2) with others. Different from the clinical outcomes of the CIMP group in hepatocellular carcinoma,[20] the CIMP group in LUSC was unlikely to be associated with shorter OS (Figure 2D).

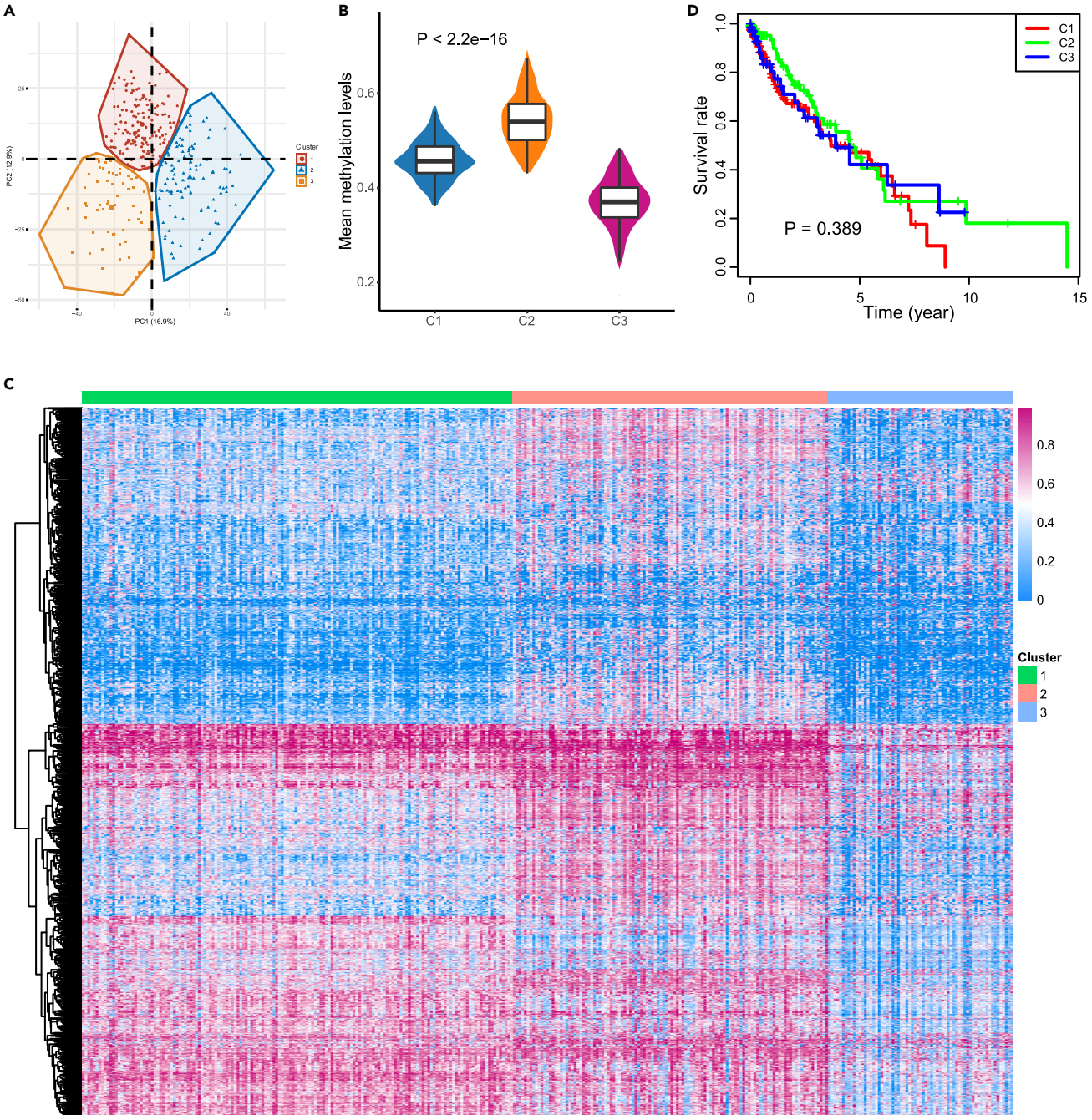


Figure 2. CpG island methylator phenotype of LUSC

(A) Principal-component analysis (PCA) showing that the 3 clusters divided can be clearly distinguished. The percent sign in the title of the horizontal and vertical axes indicates the proportion of variance that the corresponding principal component can explain. Each point presents one patient involved in PCA.

(B) The mean CpG methylation levels of the 3 clusters. Data were presented as mean ± SEM.

(C) Heatmap reflecting the overall methylation levels of the 3 methylation subtypes. A higher number on the color scale means a higher hypermethylation level. A total of 2,120 most variable CpGs were involved.

(D) Survival curves of the 3 clusters. C1, cluster 1; C2, cluster 2; C3, cluster 3.

Using the transcriptomic data of these patients, gene set enrichment analysis (GSEA) was performed by comparing gene expression profiles between each cluster and the remaining two clusters. The result showed cluster 1 was mainly enriched in immune response-related pathways and cluster 2 was mainly

Table 1. Clinical features and somatic mutation differences of three clusters

Characteristics	Classes	Cluster 1	Cluster 2	Cluster 3	Adjusted p value
n		156	114	67	
Gender (%)	Female	51 (32.7)	24 (21.1)	14 (20.9)	0.057
	Male	105 (67.3)	90 (78.9)	53 (79.1)	
Age at diagnosis (%)	<65	45 (28.8)	43 (37.7)	20 (29.9)	0.301
	≥65	111 (71.2)	71 (62.3)	47 (70.1)	
Smoking history (%)	Nonsmoker	6 (3.8)	5 (4.4)	2 (3.0)	1.000
	Current or past smoker	150 (96.2)	109 (95.6)	65 (97.0)	
Stage (%)	High (Stage I, Stage II)	17 (10.9)	19 (16.7)	18 (26.9)	0.014
	Low (Stage III, Stage IV)	139 (89.1)	95 (83.3)	49 (73.1)	
TP53 (%)	Mutation	113 (72.4)	98 (86.0)	56 (83.6)	0.018
	Wild-type	43 (27.6)	16 (14.0)	11 (16.4)	
KEAP1 (%)	Mutation	10 (6.4)	18 (15.8)	3 (4.5)	0.014
	Wild-type	146 (93.6)	96 (84.2)	64 (95.5)	
CDKN2A (%)	Mutation	18 (11.5)	27 (23.7)	11 (16.4)	0.031
	Wild-type	138 (88.5)	87 (76.3)	56 (83.6)	
KMT2D (%)	Mutation	25 (16.0)	30 (26.3)	21 (31.3)	0.020
	Wild-type	131 (84.0)	84 (73.7)	46 (68.7)	
EGFR (%)	Mutation	8 (5.1)	3 (2.6)	2 (3.0)	0.660
	Wild-type	148 (94.9)	111 (97.4)	65 (97.0)	
KRAS (%)	Mutation	3 (1.9)	0 (0.0)	0 (0.0)	0.306
	Wild-type	153 (98.1)	114 (100.0)	67 (100.0)	
ROS1 (%)	Mutation	12 (7.7)	10 (8.8)	9 (13.4)	0.388
	Wild-type	144 (92.3)	104 (91.2)	58 (86.6)	
ALK (%)	Mutation	7 (4.5)	4 (3.5)	2 (3.0)	0.873
	Wild-type	149 (95.5)	110 (96.5)	65 (97.0)	
STK11 (%)	Mutation	2 (1.3)	0 (0.0)	0 (0.0)	0.686
	Wild-type	154 (98.7)	114 (100.0)	67 (100.0)	

enriched in "KEAP1-NFE2L2 pathway" and multiple cell cycle-related pathways; meanwhile cluster 3 was mainly enriched in pathways including "Meiotic chromosome segregation", "RNA polyadenylation", "protein DNA complex subunit organization", etc (Figure S2). Interestingly, both the SNP data and GSEA results indicated the dysregulated KEAP1-NFE2L2 signaling pathway is closely associated with cluster 2 (namely, LUSC CIMP). Therapies targeting this pathway may lead to more benefits for LUSC CIMP patients.

Methylation-based diagnostic model

To reduce the noise from the data as much as possible and filter out the most important CpGs to identify normal or LUSC samples, the Boruta algorithm was applied to the stage I subgroup, stage II-IV subgroup, and the entire training cohort. Then overlapped 32 hypermethylated CpGs and 37 hypomethylated CpGs in all three subgroups were obtained (Figure 3A, Table S2), which were used for the latter nested 5-fold cross-validation. And the best classification performance was achieved at the lowest cross-validation error with 5 CpGs (cg14823851 of T-box transcription factor 4 (TBX4), cg02772121 of tripartite motif-containing 15 (TRIM15), cg10424681 of C6orf201, cg12910906 of Rho guanine nucleotide exchange factor 4 (ARHGEF4), and cg20181079 of OR4D11) (Figure 3B). Internal verification using the 5-fold cross-validation method showed the classification model was highly fitted to the training data (area under the receiver operating characteristic (ROC) curve (AUC) = 1 in 5-folds, Figure 3C).

The model was further imported into 2 test cohorts. Distinct methylation differences of the 5 CpGs were shown between normal and LUSC samples. Noticeably, apart from the consistent methylation differences

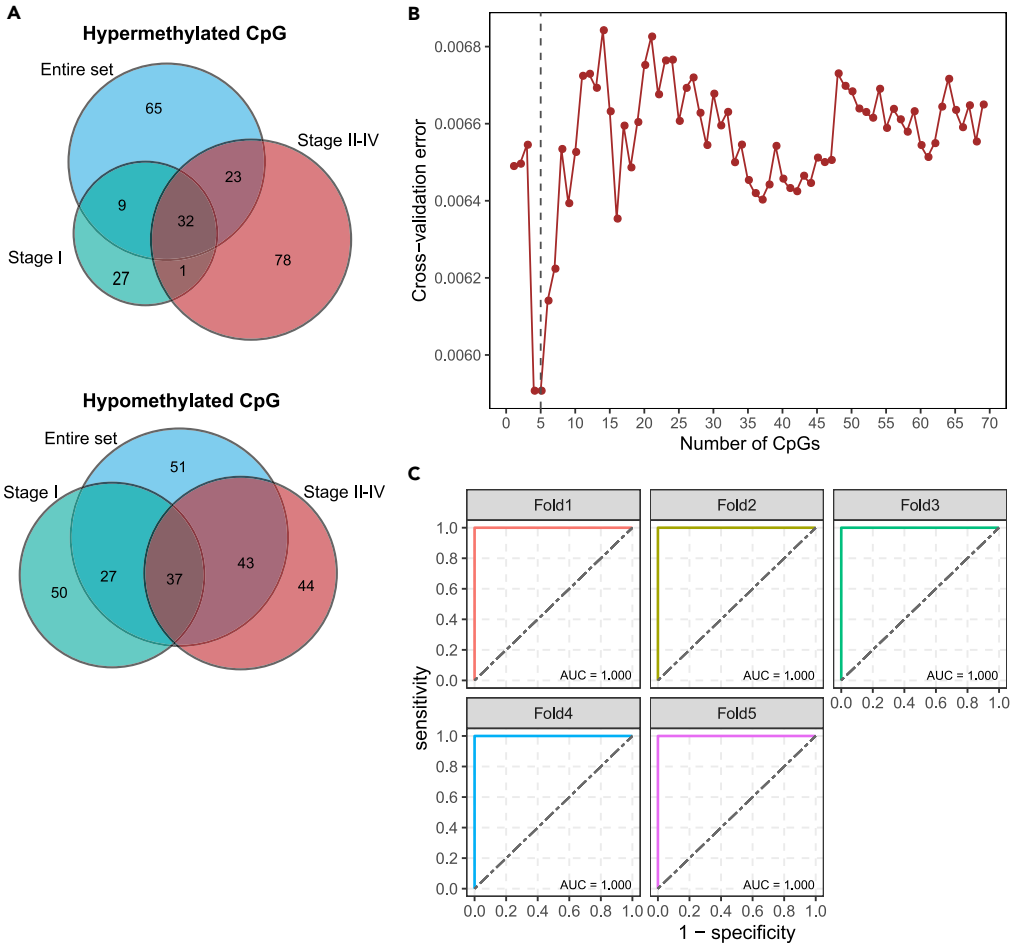


Figure 3. Construction of a CpG-based diagnostic model for LUSC

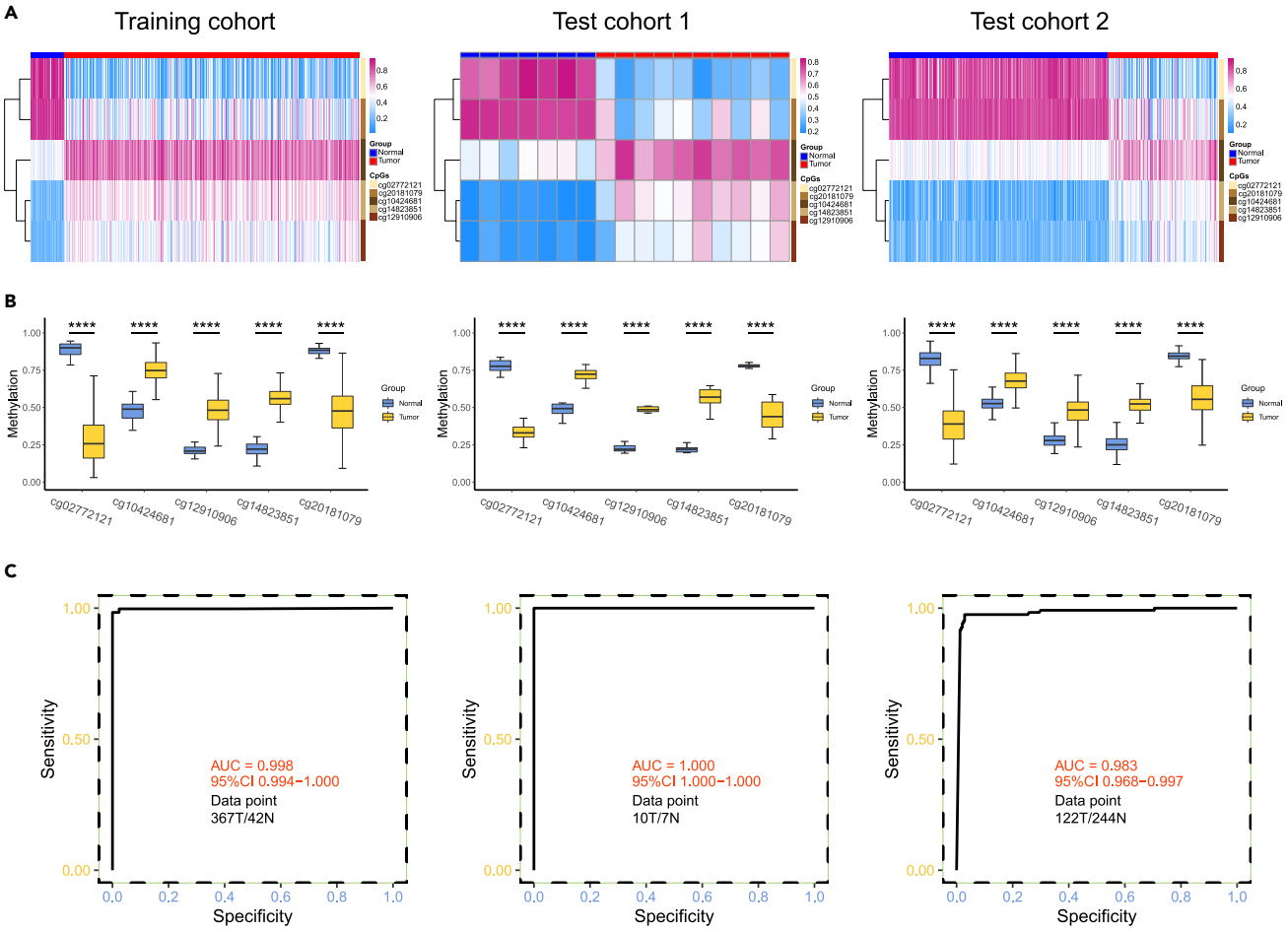
(A) Venn diagrams of overlapped hypermethylated CpGs (upper) and hypomethylated CpGs (down) obtained after feature selection for the full training set, Stage I set, and Stage II-IV training set, respectively.

(B) Average cross-validation errors for models built with different numbers of CpGs.

(C) The ROC curves of the resulting model consist of 5 CpGs from 5-fold cross-validation in the training set.

of the candidate CpGs in 3 cohorts, unsupervised hierarchical clustering with dendrogram showed a consistent clustering relationship between CpGs in different cohorts (Figures 4A and 4B), both of which indicated the robustness of the results.

As presented in Figures 4C and 4D, the ROC curves and confusion matrices showed great discriminative ability of the model in differentiating groups (AUC = 0.998, sensitivity = 97.6%, and specificity = 99.7% in training cohort; AUC = 1.000, sensitivity = 100.0%, and specificity = 100.0% in test cohort 1; and AUC = 0.983, sensitivity = 97.5%, and specificity = 82.4% in test cohort 2). Since test cohort 2 had well-established AJCC staging data, we further analyzed the performance for early-stage LUSC. 56 stage I samples were analyzed, and the results showed its classification performance (AUC = 0.983, sensitivity = 96.4%, and specificity = 82.4%) was barely changed for samples from early-stage LUSC (Figure S3). We also compared the performance in discriminating between LUSC and normal samples with a previous study.[21] After being trained by the same training cohort, a similar random forest-based model was generated using the CpGs identified by Shi et al. However, compared with our 5-CpG signature, their 6-CpG signature does not perform so satisfactorily in test cohorts (AUC = 0.813, sensitivity = 100.0%, and specificity = 50.0% in test cohort 1; and AUC = 0.451, sensitivity = 100.0%, and specificity = 17.6% in test cohort 2) (Table S3), indicating the robustness of our model in different datasets.



D

Training dataset	Real Normal	Real LUSC	
Predict normal	41	1	
Predict LUSC	1	366	Totals
Totals	42	367	409
Correct	41	366	407
Specificity (%)	99.7		
Sensitivity (%)		97.6	

Test dataset 1	Real Normal	Real LUSC	
Predict normal	7	0	
Predict LUSC	0	10	Totals
Totals	7	10	17
Correct	7	10	17
Specificity (%)	100.0		
Sensitivity (%)		100.0	

Test dataset 2	Real Normal	Real LUSC	
Predict normal	201	3	
Predict LUSC	43	119	Totals
Totals	244	122	366
Correct	201	119	320
Specificity (%)	82.4		
Sensitivity (%)		97.5	

Figure 4. Model validation and the diagnostic performance of the constructed model in the training set (left), test cohort 1 (medium), and test cohort 2 (right)

(A) Heatmaps showing methylation levels of 5 CpGs in the model.

(B) Boxplots reflecting methylation differences of the 5 CpGs. Data were presented as mean ± SEM. Student's *t* test was used. ****, $p < 0.0001$.

(C) ROC curves illustrating the diagnostic performance of the model.

(D) Confusion matrices presenting model-specific prediction results, sensitivity, and specificity.

Further analyses revealed the aberrant DNA methylations of these CpGs developed early and were found in lung progressive carcinoma *in situ* (CIS) lesions. Interestingly, compared to regressive CIS samples, the progressive ones showed significantly more severe dysregulated methylation, suggesting the potential of these biomarkers in discriminating between benign and malignant lesions. In addition, although not statistically significant in all comparisons, the results showed incremental trends in the levels of abnormal methylations of the 5 CpGs from normal samples to regressive CIS samples and from regressive to progressive CIS samples (Figure S4A), indicating their progression-specific methylation changes in these pathological processes. However, the methylation statuses of the CpGs in LUAD samples showed consistent trends with those in LUSC samples (Figure S4B).

In addition, the methylation levels of the adjacent CpGs of the candidate CpGs in promoter regions were analyzed. Results exhibited these CpGs had universally the same methylation statuses as corresponding candidate CpGs (Figure 5). Interestingly, cg11838898 of ARHGEF4 presented significant hypomethylation statuses in LUSC samples among all datasets, revealing a diametrically opposed DNA methylation trend compared to other CpGs from the ARHGEF4 promoter region.

Validation of methylation signatures in LUSC

The information of patients participating in this study was listed in Table S4. To further identify methylation markers for LUSC diagnosis, 10 pairs of LUSC tissues and normal lung samples were utilized for pyrosequencing for the validation of selected CpG sites. Four CpG sites (cg14823851, cg02772121, cg10424681, and cg12910906) all revealed the same trends in methylation levels as these public databases along with extremely significant differences (p values equal to 1.51e-04, 1.10e-05, 5.86e-04, and 2.76e-04, respectively; Figure 6A). Simultaneously, the pyrosequencing revealed the methylation levels of other two CpGs after cg14823851, four CpGs around cg10424681, and two CpGs surrounding cg12910906, all presenting corresponding results ($p < 0.001$, each; Figures 6B–6D). As for cg20181079, however, the primer specificity score was too low to complete the PCR amplification, so we did not verify its methylation by experiments here. Shortly, the pyrosequencing results illustrated the perfect diagnostic effect of our methylation model.

Expression status of associated genes

Considering the central dogma of molecular biology,[22] we conducted qRT-PCR and IHC experiments to explore the expression genes corresponding to candidate CpGs at the mRNA and protein levels. What should be mentioned is, after the removal of surgical specimens for clinicopathologic diagnosis, the remaining samples for our study were occasionally too small to extract both DNA and RNA. Totally, we gathered RNA of LUSC tissues and paired normal lung tissues from 30 patients, in which 7 pairs of specimens were also involved in DNA pyrosequencing (Table S4). Compared with normal lung tissues, the mRNA expression of TBX4 is obviously lower in LUSC tissues ($p = 1.90e-04$), and so is the level of C6orf201 ($p = 1.98e-02$), while the mRNA levels of TRIM15 and ARHGEF4 presented no significant difference between the normal tissues and LUSC tissues ($p > 0.05$, both), which may also be reflected by their relatively minor differences in TCGA-LUSC data (Figure S5). Otherwise, the complexity and heterogeneity of clinically derived mRNA samples should also be considered; therefore, 30 pairs of data may still not be enough to understand the trends of some genes.[23] Accordingly, we successfully measured the low expression of TBX4 and C6orf201 mRNAs in LUSC tissues, matching with the hypermethylation of mapped methylation sites in the diagnostic model.

Thanks to the acquisition and reuse of pathological specimens, we detected the protein levels of TBX4, TRIM15, C6orf201, and ARHGEF4 by IHC with the same patients involved in DNA methylation validation (Figures 7A–7D). Quantitative analysis according to mean integrated optical density (IOD) showed that the expression of TBX4, C6orf201, and ARHGEF4 in LUSC was lower compared to the paired adjacent normal lung specimens (p values equal to 2.22e-04, 1.39e-04, and 4.01e-04, respectively), suggesting the possible gene silencing function of DNA methylation in LUSC. Besides, the protein of TRIM15 was over-expressed in LUSC tissues ($p = 2.52e-04$), mapped to the hypomethylation of cg02772121 as previously identified (Figure 7E). The mRNA and protein expression patterns of 4 genes in our surgically resectable LUSC patients (mostly in the I-II stage) were consistent with their methylation profiles, which confirmed that these CpG sites—corresponded with mapped genes—are novel promising early-diagnosis biomarkers for human LUSC.

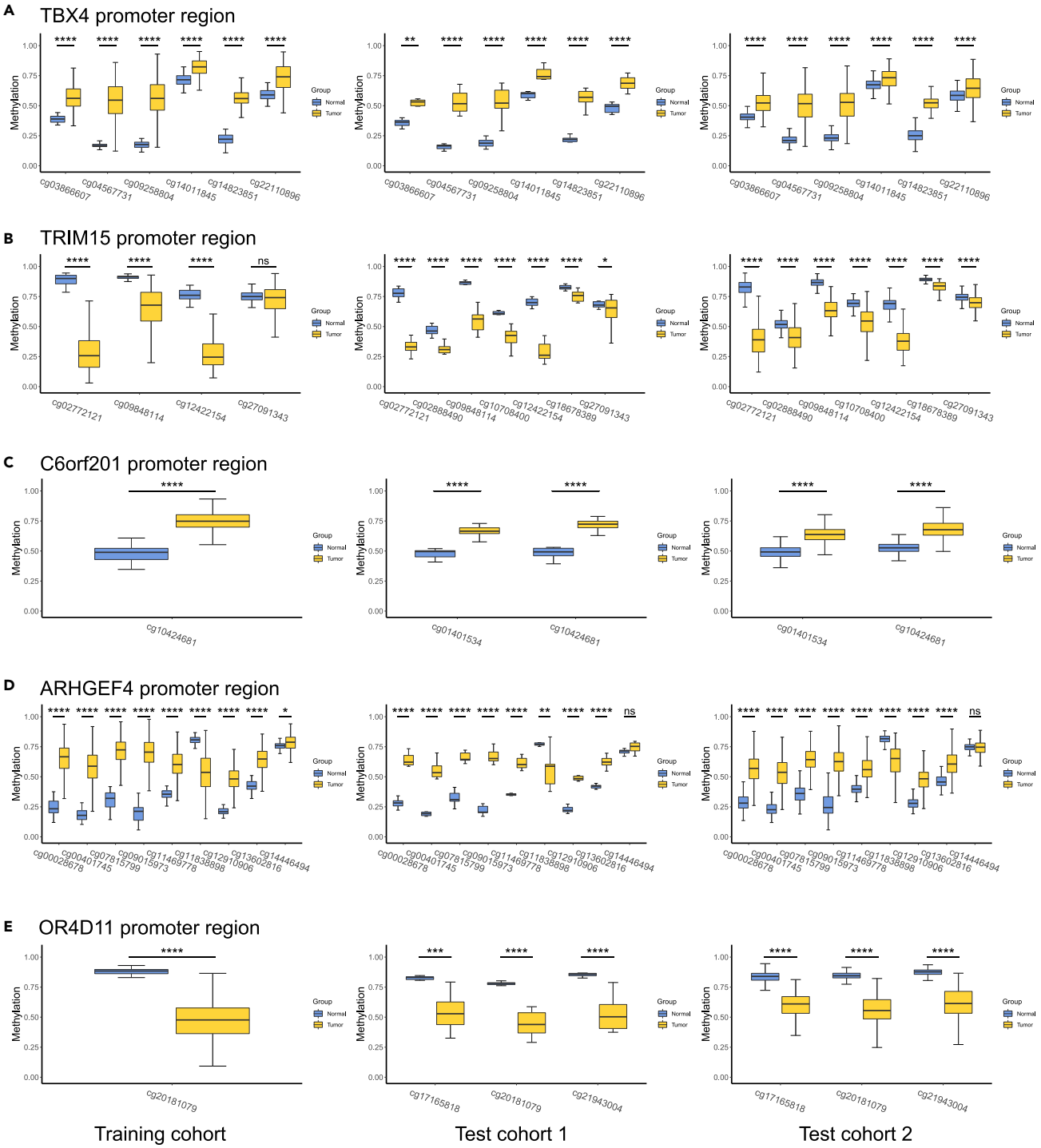


Figure 5. The methylation of all CpGs in the promoter regions of five genes

(A–E) Promoter region methylation of TBX4 (A), TRIM15 (B), C6orf201 (C), ARHGEF4 (D), and OR4D11 (E) in the training set (left), test cohort 1 (medium), and test cohort 2 (right). Data were presented as mean ± SEM. Student's *t* test was used. ns, not significant; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$.

DISCUSSION

The present study demonstrates a promising diagnostic CpG biomarker panel for LUSC patients and unveils the existence of underlying methylation-based gene regulation in tumor development. As far as we

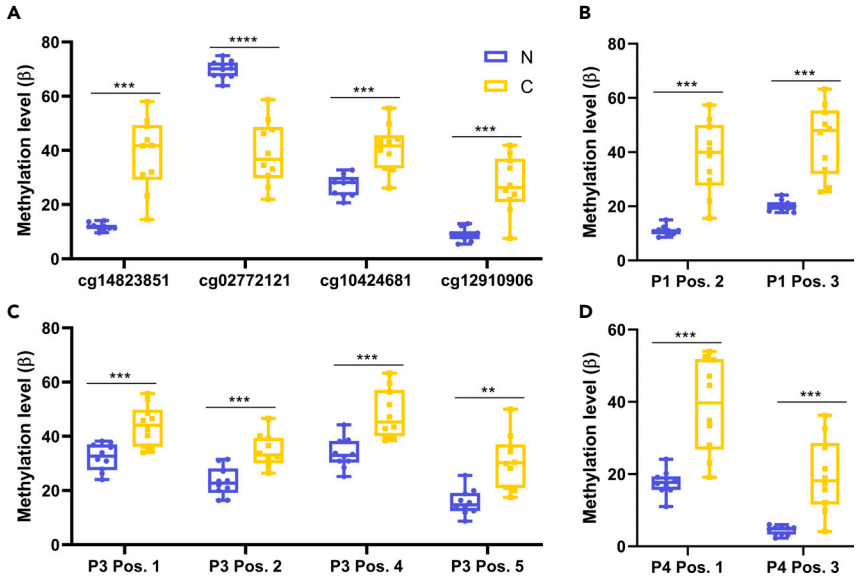


Figure 6. The pyrosequencing results of 10 pairs of normal and LUSC tissues

(A) Verification of four CpG sites in the diagnostic model. (cg14823851 from TBX4 matched P1 Pos. 1, cg02772121 from TRIM15 matched P2, cg10424681 from C6orf201 matched P3 Pos. 3, and cg12910906 from ARHGEF4 matched P4 Pos. 2). (B–D) The methylation levels of two CpGs after cg14823851 (B), four CpGs around cg10424681 (C), and two CpGs surrounding cg12910906 (D). Results are expressed as mean ± SEM. N, normal (in blue); C, cancer (in red); P1, primer 1; Pos. 1, position 1. Data were presented as mean ± SEM. Student's *t* test was used. **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$.

searched, this article is the first one to explore the DNA methylation-based biomarkers for the diagnosis of LUSC. Recent technological advances in high-throughput genome-wide detection and a large improvement in computational tools have provided the perspective for DNA methylation analysis.[24] A lot of scientific papers have focused on the clinical value of DNA methylation and illustrated nice models for the early detection of cancers such as esophageal adenocarcinoma[25] and hepatocellular carcinoma.[26] This study provides outstanding predictive biomarkers with the following advantages. Firstly, the methylation data all came from Illumina Infinium HumanMethylation450 BeadChips rather than Illumina Infinium HumanMethylation27 BeadChips, providing a much wider selection scope for candidate CpGs.[16] Secondly, in the establishment process, we took full consideration of the methylation alteration in early or advanced LUSC and combined several machine learning algorithms, which aimed to get a reliable and feasible prediction classifier. Thirdly, most of the similar studies only focused on the hypermethylated CpGs,[20,27] perhaps because the importance of decreases in DNA methylation associated with carcinogenesis was overshadowed by the emphasis on cancer-linked hypermethylation of tumor-suppressor genes. However, hypomethylation of DNA sequences is often observed during the early stages of tumorigenesis,[28] and our study also valued this neglected methylation alternation. Last but not least, to confirm the reliability of the CpGs, we performed integrated analyses from internal cross-validation of training cohort to external validation of multiple test cohorts and pyrosequencing validation of clinical specimens.

Here, we further analyzed the expression changes of the corresponding genes in the mRNA and protein levels along the central dogma of molecular biology, demonstrating their feasibility as diagnostic markers for LUSC. Previous studies based on proteomics and transcriptomics have reported higher stability of proteins than mRNA in cancer tissues such as prostate cancer[29] and breast cancer.[30] It is also illustrated that DNA methylation status possessed a stronger connection with protein than mRNA level in curable prostate cancer.[31] Moreover, the source of mRNA we detected was not so consistent with the other two data groups. Statistically, although the mRNA levels of TRIM15 and ARHGEF4 did not show significant changes in qRT-PCR, the trends of TBX4 and C6orf201 were negatively related to the methylated degree in LUSC. Meanwhile, the protein expression of four genes, when compared with DNA methylation status one to one, presented a significant negative correlation. The results obtained basically conform to the law of methylation-regulated gene expression, suggesting potential biomarkers at different levels.

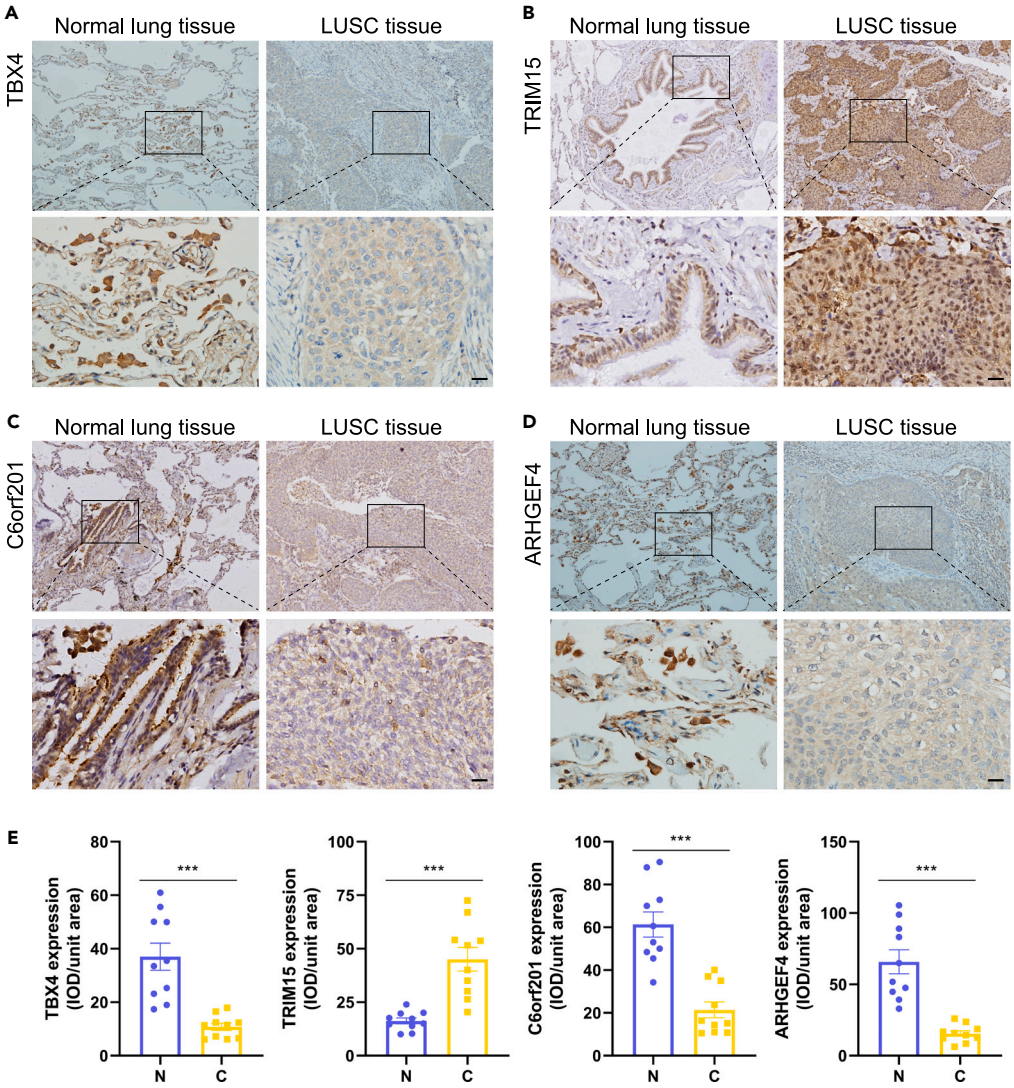


Figure 7. The protein levels of TBX4, TRIM15, C6orf201, and ARHGEF4 in LUSC

(A–D) Representative images of human normal adjacent lung tissues and LUSC tissues stained with TBX4 (A), TRIM15 (B), C6orf201 (C), or ARHGEF4 (D) antibodies as indicated. Scale bar, 20 μm.

(E) Immunohistochemical (IHC) staining of 4 antibodies in 10 pairs of normal and LUSC tissues was quantified by mean IOD. N, normal; C, cancer. Scale bar, 20 μm. Data were presented as mean ± SEM. Paired Student's *t* test was used. ***, $p < 0.001$.

TBX4 is a member of the evolutionarily conserved T-box family, whose main function is to act as a transcription factor to regulate gene expression patterns during embryonic development.[32] Mutations or deletions in the TBX4 gene could result in developmental organ disorders like pulmonary hypertension.[33,34] It has been well demonstrated that TBX4 works as a methylation marker to differentiate normal and osteoarthritic cartilage,[35] identify the stage of bladder cancers,[36] predict the survival for some pancreatic ductal adenocarcinoma patients,[37] and even distinguish LUAD from normal lung tissues—mapped to cg14823851—which is consistent with our results.[27] A mechanistic study suggested TBX4 inhibits the migration and invasiveness in two kinds of LUAD cell lines,[38] and we further verified its importance for LUSC. TRIM15 is a focal adhesion protein regulating focal adhesion disassembly, which could promote cell migration by reducing adhesion between cells.[39] Han X et al. analyzed TCGA and GEO databases and IHC verification results, finding that NSCLC patients with high expression of TRIM15 have a significantly poorer prognosis.[40] Recent studies defined the ubiquitin ligase function of TRIM15 in the development of melanomas[41] and illustrated its pivotal roles to promote NSCLC progression by promoting Keap1

ubiquitination and degradation in the Keap1-Nrf2 signaling pathway.[42] We found that its expression level increased in LUSC through the detection of mRNA and protein levels matching with hypomethylation in the promoter region, which provided certain evidence for further research on the role of TRIM15 in tumorigenesis and development. ARHGEF4 is a Rho guanine nucleotide exchange factor, which can stimulate ERK1/2 and GSK-3α/β and predict an unsatisfied prognosis for pancreatic cancer patients.[43] However, Qin Y et al. have found that ARHGEF4 is an extremely sensitive hypermethylated marker for esophageal cancer through the methylation analysis of a large number of blood and tissue samples.[44] Consistent with that, our study presented most CpGs of ARHGEF4 in the promoter region are hypermethylated and further tested the lower expression of its protein in LUSC tissues, suggesting latent methylation regulation. We also found that C6orf201 has significant changes in methylation as well as gene expression in LUSC. There remains a blank for research about it. What role C6orf201 played in cells and tumors is still needed to be studied, and understanding what impact C6orf201 has on the biological functions of LUSC has also become a center of our further research.

In summary, our study integrated genome-wide DNA methylation and relevant transcriptome data and identified five methylation markers including cg14823851 (TBX4), cg02772121 (TRIM15), cg10424681 (C6orf201), cg12910906 (ARHGEF4), and cg20181079 (OR4D11) for early diagnosis of LUSC. These markers showed robust classification performance in distinguishing between LUSC and normal samples with extremely high sensitivity and specificity. Corresponding changes in gene expression levels were also confirmed in LUSC tissues. Methylation biomarkers identified in this study may act as an important basis for further investigations regarding early LUSC.

Limitations of the study

Although the diagnostic model presented high accuracy and was verified well in our patients, some limitations of our study still exist and should be mentioned. First, besides verification in lung tissues, we have not validated our model via patient blood or sputum. Previous studies have illustrated the comparability of hypermethylation in tumor tissues and blood, as well as discussed the significance of circulating tumor DNA (ctDNA) methylation analysis working as an early, noninvasive detection method for lung cancer.[16,45] It would reflect a much stronger clinical value if our diagnostic methylation sites got confirmed by blood samples. Next, our verification cohort is a little small. More validation sets could offer more reliable results and may provide information on some clinicopathologic subgroups. At last, notwithstanding that we explored the connection of our methylation model with related genes and discussed with other studies, the mechanism by which these methylation sites regulate tumor progression still needs to be expounded in future studies, and methods like the usage of DNA-demethylating agents, such as 5-azacytidine and 5-azadeoxycytidine, in preclinical models[46,47] are potential to guide the clinical treatment effect.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Data collection and processing
  - Identification of DMPs, functional enrichment analysis, and exploration of CIMP
  - Model construction and evaluation in public databases
  - DNA pyrosequencing assay
  - Quantitative real-time PCR (qRT-PCR)
  - Protein immunohistochemistry (IHC) analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.107013.

ACKNOWLEDGMENTS

The authors want to give their sincerest thanks to Dr. Weiwei Lai and Dr. Rui Yang for their helpful comments and suggestions. This work was supported by the Fundamental Research Funds for the Central Universities of Central South University (2021zzts0383, B. He). Hunan Provincial Key Area R&D Programmes (2019SK2253 and 2020SK53424, X. Wang), the Scientific Research Program of Hunan Provincial Health Commission (20201047, X. Wang), the Nature Science Foundation of Hunan Province (2021JJ30957, X. Wang), the Science and Technology Innovation Program of Hunan Province (2022RC3072, Y. Tao), Central South University Research Programme of Advanced Interdisciplinary Studies (2023QYJC030, Y. Tao and X. Wang), and the Scientific Research Launch Project for employees of the Second Xiangya Hospital (Q. Cai).

AUTHOR CONTRIBUTIONS

XW and YGT conceived and designed the project. QDC and BXH carried out data analysis, experimental implementation, and manuscript writing. SLP helped in software coding and checking. QCL helped in immunohistochemistry. GXT, SS, WLP, and BLQ collected and dealt with information and specimens. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 14, 2022
Revised: March 11, 2023
Accepted: May 29, 2023
Published: June 5, 2023

REFERENCES

1. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. Ca - Cancer J. Clin. *71*, 209–249. https://doi.org/10.3322/caac.21660.

2. Zappa, C., and Mousa, S.A. (2016). Non-small cell lung cancer: current treatment and future advances. Transl. Lung Cancer Res. *5*, 288–300. https://doi.org/10.21037/tlcr.2016.06.07.

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

4. Duma, N., Santana-Davila, R., and Molina, J.R. (2019). Non-small cell lung cancer: epidemiology, screening, diagnosis, and treatment. Mayo Clin. Proc. *94*, 1623–1640. https://doi.org/10.1016/j.mayocp.2019.01.013.

5. Peled, N., and Ilouze, M. (2015). Screening for lung cancer: what comes next? J. Clin. Oncol. *33*, 3847–3848. https://doi.org/10.1200/jco.2015.63.1713.

6. Fitz-James, M.H., and Cavalli, G. (2022). Molecular mechanisms of transgenerational epigenetic inheritance. Nat. Rev. Genet. *23*, 325–341.

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

8. Michalak, E.M., Burr, M.L., Bannister, A.J., and Dawson, M.A. (2019). The roles of DNA, RNA and histone methylation in ageing and cancer. Nat. Rev. Mol. Cell Biol. *20*, 573–589.

9. Koch, A., Joosten, S.C., Feng, Z., de Ruijter, T.C., Draht, M.X., Melotte, V., Smits, K.M., Veeck, J., Herman, J.G., Van Neste, L., et al. (2018). Analysis of DNA methylation in cancer: location revisited. Nat. Rev. Clin. Oncol. *15*, 459–466. https://doi.org/10.1038/s41571-018-0004-4.

10. Schübeler, D. (2015). Function and information content of DNA methylation. Nature *517*, 321–326. https://doi.org/10.1038/nature14192.

11. Papin, C., Ibrahim, A., Gras, S.L., Velt, A., Stoll, I., Jost, B., Menoni, H., Bronner, C., Dimitrov, S., and Hamiche, A. (2017). Combinatorial DNA methylation codes at repetitive elements. Genome Res. *27*, 934–946. https://doi.org/10.1101/gr.213983.116.

12. Ushijima, T., and Suzuki, H. (2019). The origin of CIMP, at last. Cancer Cell *35*, 165–167. https://doi.org/10.1016/j.ccell.2019.01.015.

13. Ooki, A., Maleki, Z., Tsay, J.C.J., Goparaju, C., Brait, M., Turaga, N., Nam, H.S., Rom, W.N., Pass, H.I., Sidransky, D., et al. (2017). A panel of novel detection and prognostic methylated DNA markers in primary non-small cell lung cancer and serum DNA. Clin. Cancer Res. *23*, 7141–7152. https://doi.org/10.1158/1078-0432.Ccr-17-1222.

14. Liang, W., Zhao, Y., Huang, W., Gao, Y., Xu, W., Tao, J., Yang, M., Li, L., Ping, W., Shen, H., et al. (2019). Non-invasive diagnosis of early-stage lung cancer using high-throughput targeted DNA methylation sequencing of circulating tumor DNA (ctDNA). Theranostics *9*, 2056–2070. https://doi.org/10.7150/thno.28119.

15. Shen, N., Du, J., Zhou, H., Chen, N., Pan, Y., Hoheisel, J.D., Jiang, Z., Xiao, L., Tao, Y., and Mo, X. (2019). A diagnostic panel of DNA methylation biomarkers for lung adenocarcinoma. Front. Oncol. *9*, 1281. https://doi.org/10.3389/fonc.2019.01281.

16. Cai, Q., Zhang, P., He, B., Zhao, Z., Zhang, Y., Peng, X., Xie, H., and Wang, X. (2020). Identification of diagnostic DNA methylation biomarkers specific for early-stage lung adenocarcinoma. Cancer Genet. *246–247*, 1–11. https://doi.org/10.1016/j.cancergen.2020.08.002.

17. Irizarry, R.A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., et al. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat. Genet. *41*, 178–186. https://doi.org/10.1038/ng.298.

18. Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M.A., Bibikova, M., and Esteller, M. (2011). Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics *6*, 692–702. https://doi.org/10.4161/epi.6.6.16196.

19. Das, P.M., and Singal, R. (2004). DNA methylation and cancer. J. Clin. Oncol. *22*,

4632–4642. https://doi.org/10.1200/jco.2004.07.151.

20. Cheng, J., Wei, D., Ji, Y., Chen, L., Yang, L., Li, G., Wu, L., Hou, T., Xie, L., Ding, G., et al. (2018). Integrative analysis of DNA methylation and gene expression reveals hepatocellular carcinoma-specific diagnostic biomarkers. Genome Med. *10*, 42. https://doi.org/10.1186/s13073-018-0548-z.

21. Shi, Y.-X., Wang, Y., Li, X., Zhang, W., Zhou, H.-H., Yin, J.-Y., and Liu, Z.-Q. (2017). Genome-wide DNA methylation profiling reveals novel epigenetic signatures in squamous cell lung cancer. BMC Genom. *18*, 901. https://doi.org/10.1186/s12864-017-4223-3.

22. Schneider-Poetsch, T., and Yoshida, M. (2018). Along the central dogma-controlling gene expression with small molecules. Annu. Rev. Biochem. *87*, 391–420. https://doi.org/10.1146/annurev-biochem-060614-033923.

23. He, B., Wei, C., Cai, Q., Zhang, P., Shi, S., Peng, X., Zhao, Z., Yin, W., Tu, G., Peng, W., et al. (2022). Switched alternative splicing events as attractive features in lung squamous cell carcinoma. Cancer Cell Int. *22*, 5. https://doi.org/10.1186/s12935-021-02429-2.

24. Merkel, A., and Esteller, M. (2022). Experimental and bioinformatic approaches to studying DNA methylation in cancer. Cancers *14*, 349. https://doi.org/10.3390/cancers14020349.

25. Peng, W., Tu, G., Zhao, Z., He, B., Cai, Q., Zhang, P., Peng, X., Shi, S., and Wang, X. (2021). DNA methylome and transcriptome analysis established a model of four differentially methylated positions (DMPs) as a diagnostic marker in esophageal adenocarcinoma early detection. PeerJ *9*, e11355. https://doi.org/10.7717/peerj.11355.

26. Luo, B., Ma, F., Liu, H., Hu, J., Rao, L., Liu, C., Jiang, Y., Kuangzeng, S., Lin, X., Wang, C., et al. (2022). Cell-free DNA methylation markers for differential diagnosis of hepatocellular carcinoma. BMC Med. *20*, 8.

27. Li, M., Zhang, C., Zhou, L., Li, S., Cao, Y.J., Wang, L., Xiang, R., Shi, Y., and Piao, Y. (2020). Identification and validation of novel DNA methylation markers for early diagnosis of lung adenocarcinoma. Mol. Oncol. *14*, 2744–2758. https://doi.org/10.1002/1878-0261.12767.

28. Ehrlich, M. (2009). DNA hypomethylation in cancer cells. Epigenomics *1*, 239–259. https://doi.org/10.2217/epi.09.33.

29. Shao, W., Guo, T., Toussaint, N.C., Xue, P., Wagner, U., Li, L., Charmpi, K., Zhu, Y., Wu, J., Buljan, M., et al. (2019). Comparative analysis of mRNA and protein degradation in prostate tissues indicates high stability of proteins. Nat. Commun. *10*, 2524. https://doi.org/10.1038/s41467-019-10513-5.

30. Mertins, P., Mani, D.R., Ruggles, K.V., Gillette, M.A., Clauser, K.R., Wang, P., Wang, X., Qiao, J.W., Cao, S., Petralia, F., et al. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. Nature *534*, 55–62. https://doi.org/10.1038/nature18003.

31. Sinha, A., Huang, V., Livingstone, J., Wang, J., Fox, N.S., Kurganovs, N., Ignatchenko, V., Fritsch, K., Donmez, N., Heisler, L.E., et al. (2019). The proteogenomic landscape of curable prostate cancer. Cancer Cell *35*, 414–427.e6. https://doi.org/10.1016/j.ccell.2019.02.005.

32. Haarman, M.G., Kerstjens-Frederikse, W.S., and Berger, R.M.F. (2019). The ever-expanding phenotypical spectrum of human TBX4 mutations: from toe to lung. Eur. Respir. J. *54*, 1901504. https://doi.org/10.1183/13993003.01504-2019.

33. Galambos, C., Mullen, M.P., Shieh, J.T., Schwerk, N., Kielt, M.J., Ullmann, N., Boldrini, R., Stucin-Gantar, I., Haass, C., Bansal, M., et al. (2019). Phenotype characterisation of TBX4 mutation and deletion carriers with neonatal and paediatric pulmonary hypertension. Eur. Respir. J. *54*, 1801965. https://doi.org/10.1183/13993003.01965-2018.

34. Southgate, L., Machado, R.D., Gräf, S., and Morrell, N.W. (2020). Molecular genetic framework underlying pulmonary arterial hypertension. Nat. Rev. Cardiol. *17*, 85–95. https://doi.org/10.1038/s41569-019-0242-x.

35. Alvarez-Garcia, O., Fisch, K.M., Wineinger, N.E., Akagi, R., Saito, M., Sasho, T., Su, A.I., and Lotz, M.K. (2016). Increased DNA methylation and reduced expression of transcription factors in human osteoarthritis cartilage. Arthritis Rheumatol. *68*, 1876–1886. https://doi.org/10.1002/art.39643.

36. Reinert, T., Modin, C., Castano, F.M., Lamy, P., Wojdacz, T.K., Hansen, L.L., Wiuf, C., Borre, M., Dyrskjøt, L., and Orntoft, T.F. (2011). Comprehensive genome methylation analysis in bladder cancer: identification and validation of novel methylated genes and application of these as urinary tumor markers. Clin. Cancer Res. *17*, 5582–5592. https://doi.org/10.1158/1078-0432.Ccr-10-2659.

37. Zong, M., Meng, M., and Li, L. (2011). Low expression of TBX4 predicts poor prognosis in patients with stage II pancreatic ductal adenocarcinoma. Int. J. Mol. Sci. *12*, 4953–4963. https://doi.org/10.3390/ijms12084953.

38. Lai, I.L., Chang, Y.S., Chan, W.L., Lee, Y.T., Yen, J.C., Yang, C.A., Hung, S.Y., and Chang, J.G. (2019). Male-specific long noncoding RNA TTTY15 inhibits non-small cell lung cancer proliferation and metastasis via TBX4. Int. J. Mol. Sci. *20*, 3473. https://doi.org/10.3390/ijms20143473.

39. Uchil, P.D., Pawliczek, T., Reynolds, T.D., Ding, S., Hinz, A., Munro, J.B., Huang, F., Floyd, R.W., Yang, H., Hamilton, W.L., et al. (2014). TRIM15 is a focal adhesion protein that regulates focal adhesion disassembly. J. Cell Sci. *127*, 3928–3942. https://doi.org/10.1242/jcs.143537.

40. Ma, A., Tang, M., Zhang, L., Wang, B., Yang, Z., Liu, Y., Xu, G., Wu, L., Jing, T., Xu, X., et al. (2019). USP1 inhibition destabilizes KPNA2 and suppresses breast cancer metastasis. Oncogene *38*, 2405–2419. https://doi.org/10.1038/s41388-018-0590-8.

41. Zhu, G., Herlyn, M., and Yang, X. (2021). TRIM15 and CYLD regulate ERK activation via lysine-63-linked polyubiquitination. Nat. Cell Biol. *23*, 978–991. https://doi.org/10.1038/s41556-021-00732-8.

42. Liang, M., Wang, L., Sun, Z., Chen, X., Wang, H., Qin, L., Zhao, W., and Geng, B. (2022). E3 ligase TRIM15 facilitates non-small cell lung cancer progression through mediating Keap1-Nrf2 signaling pathway. Cell Commun. Signal. *20*, 62. https://doi.org/10.1186/s12964-022-00875-7.

43. Taniuchi, K., Furihata, M., Naganuma, S., and Saibara, T. (2018). ARHGEF4 predicts poor prognosis and promotes cell invasion by influencing ERK1/2 and GSK-3α/β signaling in pancreatic cancer. Int. J. Oncol. *53*, 2224–2240. https://doi.org/10.3892/ijo.2018.4549.

44. Qin, Y., Wu, C.W., Taylor, W.R., Sawas, T., Burger, K.N., Mahoney, D.W., Sun, Z., Yab, T.C., Lidgard, G.P., Allawi, H.T., et al. (2019). Discovery, validation, and application of novel methylated DNA markers for detection of esophageal cancer in plasma. Clin. Cancer Res. *25*, 7396–7404. https://doi.org/10.1158/1078-0432.Ccr-19-0740.

45. Xu, R.H., Wei, W., Krawczyk, M., Wang, W., Luo, H., Flagg, K., Yi, S., Shi, W., Quan, Q., Li, K., et al. (2017). Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. Nat. Mater. *16*, 1155–1161. https://doi.org/10.1038/nmat4997.

46. Biktasova, A., Hajek, M., Sewell, A., Gary, C., Bellinger, G., Deshpande, H.A., Bhatia, A., Burtness, B., Judson, B., Mehra, S., et al. (2017). Demethylation therapy as a targeted treatment for human papillomavirus-associated head and neck cancer. Clin. Cancer Res. *23*, 7276–7287. https://doi.org/10.1158/1078-0432.Ccr-17-1438.

47. Vernier, M., McGuirk, S., Dufour, C.R., Wan, L., Audet-Walsh, E., St-Pierre, J., and Giguère, V. (2020). Inhibition of DNMT1 and ERRα crosstalk suppresses breast cancer via derepression of IRF4. Oncogene *39*, 6406–6420. https://doi.org/10.1038/s41388-020-01438-1.

48. Hata, A., Nakajima, T., Matsusaka, K., Fukuyo, M., Morimoto, J., Yamamoto, T., Sakairi, Y., Rahmutulla, B., Ota, S., Wada, H., et al. (2020). A low DNA methylation epigenotype in lung squamous cell carcinoma and its association with idiopathic pulmonary fibrosis and poorer prognosis. Int J Cancer *146*, 388–399. https://doi.org/10.1002/ijc.32532.

49. Sandoval, J., Mendez-Gonzalez, J., Nadal, E., Chen, G., Carmona, F.J., Sayols, S., Moran, S., Heyn, H., Vizoso, M., Gomez, A., et al. (2013). A prognostic DNA methylation signature for stage I non-small-cell lung cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology *31*, 4140–4147. https://doi.org/10.1200/jco.2012.48.5516.

50. Tian, Y., Morris, T.J., Webster, A.P., Yang, Z., Beck, S., Feber, A., and Teschendorff, A.E.

(2017). ChAMP: updated methylation analysis pipeline for Illumina BeadChips. Bioinformatics (Oxford, England) *33*, 3982–3984. https://doi.org/10.1093/bioinformatics/btx513.

51. Lunardon, N., Menardi, G., and Torelli, N.J.R.j. (2014). ROSE: a package for binary imbalanced learning, p. 6.

52. Kursa, M.B. (2014). Robustness of Random Forest-based gene selection methods. BMC bioinformatics *15*, 8. https://doi.org/10.1186/1471-2105-15-8.

53. Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. R News *2*, 18–22. https://CRAN.R-project.org/doc/Rnews/.

54. Kuhn, Max (2008). "Building Predictive Models in R Using the caret Package." Journal of Statistical Software, 28, 1–26. doi:10.18637/jss.v028.i05, https://www.jstatsoft.org/index.php/jss/article/view/v028i05.

55. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (New York: Springer-Verlag). https://ggplot2.tidyverse.org.

56. Terry M. Therneau, Patricia M. Grambsch (2000). Modeling Survival Data: Extending the Cox Model. Springer, New York. ISBN 0-387-98784-3.

57. Otasek, D., Morris, J.H., Bouças, J., Pico, A.R., and Demchak, B. (2019). Cytoscape Automation: empowering workflow-based network analysis. Genome Biol. *20*, 185. https://doi.org/10.1186/s13059-019-1758-4.

58. Goldman, M., Craft, B., Swatloski, T., Cline, M., Morozova, O., Diekhans, M., Haussler, D., and Zhu, J. (2015). The UCSC cancer genomics browser: update 2015. Nucleic Acids Res. *43*, D812–D817. https://doi.org/10.1093/nar/gku1073.

59. Kursa, M.B., and Rudnicki, W.R. (2010). Feature selection with the Boruta package. J. Stat. Software *36*, 1–13.

60. Baturynska, I., and Martinsen, K. (2021). Prediction of geometry deviations in additive manufactured parts: comparison of linear regression with machine learning algorithms. J. Intell. Manuf. *32*, 179–200. https://doi.org/10.1007/s10845-020-01567-0.

61. Martisova, A., Holcakova, J., Izadi, N., Sebuyoya, R., Hrstka, R., and Bartosik, M. (2021). DNA methylation in solid tumors: functions and methods of detection. Int. J. Mol. Sci. *22*, 4247. https://doi.org/10.3390/ijms22084247.

62. Šestáková, Š., Šálek, C., and Remešová, H. (2019). DNA methylation validation methods: a coherent review with practical comparison. Biol. Proced. Online *21*, 19. https://doi.org/10.1186/s12575-019-0107-z.

63. Guo, L., Cui, C., Zhang, K., Wang, J., Wang, Y., Lu, Y., Chen, K., Yuan, J., Xiao, G., Tang, B., et al. (2019). Kindlin-2 links mechano-environment to proline synthesis and tumor growth. Nat. Commun. *10*, 845. https://doi.org/10.1038/s41467-019-08772-3.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
TBX4	Santa Cruz Biotechnology	Cat# sc-515196; RRID: AB_2938539
TRIM15	ProteinTech	Cat# 13623-1-AP; RRID: AB_2303892
C6orf201	Santa Cruz Biotechnology	Cat# sc-514971; RRID: AB_2938538
ARHGEF4	ProteinTech	Cat# 55213-1-AP; RRID: AB_2721023
Biological samples		
LUSC patient tumor tissues and paired normal lung specimens	the Second Xiangya Hospital of Central South University	N/A
Critical commercial assays		
TIANamp Genomic DNA Kit	TIANamp	Cat#4992254
EpiTect Fast DNA Bisulfite Kit	Qiagen	Cat#59104
PyroMark PCR Kit	Qiagen	Cat#978703, #978705
Trizol reagent	Invitrogen	Cat#15596018
SuperScript First Strand cDNA system	Invitrogen	Cat#11904018
SYBR Green PCR Master Mix	Roche	Cat#04707516001
MaxVision kit	Fuzhou Maixin Biotechnology Development	Cat#KIT-5004, Cat#KIT-5001
Deposited data		
LUSC and LUAD DNA methylation data from The Cancer Genome Atlas (TCGA)	The UCSC Xena database	https://xenabrowser.net/datapages/
Relevant transcriptomic data, clinical data, and somatic mutation data	the TCGA data portal	https://portal.gdc.cancer.gov/
methylation data	Hata et al.[48]	GEO: GSE121849
methylation data	Sandoval et al.[18]	GEO: GSE39279
methylation data	Shi et al.[21]	GEO: GSE52401
methylation data	Teixeira et al.[49]	GEO: GSE108124
Oligonucleotides		
primers in qPCR and DNA pyrosequencing assay, see Table S1	This paper	N/A
Software and algorithms		
PyroMark CpG Software (1.0.11)	Qiagen	N/A
GraphPad Prism 8.0	GraphPad Software Co., Ltd.	N/A
ImageJ	VILBER Co., Ltd.	N/A
R (version 4.2)	the R Core Team and the R Foundation for Statistical Computing	https://www.r-project.org/
ChAMP (The Chip Analysis Methylation Pipeline)	R package, Tian et al.[50]	N/A
ROSE (Random Over-Sampling Examples)	R package, Lunardon et al.[51]	N/A
Boruta	R package, Kursa et al.[52]	N/A
randomForest	Liaw and Wiener [53]	N/A
Caret	R package, Kuhn[54]	N/A
ggplot2	R package, Wickham[55]	N/A
Survival	R package, Terry et al.[56]	N/A
Cytoscape	Otasek et al.[57]	http://www.cytoscape.org/
GSEA software (v4.3.2)	UC San Diego and Broad Institute	http://www.broadinstitute.org/gsea/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Xiang Wang (wangxiang@csu.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table. Data reported in this paper from in-house cohorts will be shared by the lead contact upon request.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Human subjects were involved in this study. The study conformed to the ethical guidelines of the 1975 Declaration of Helsinki and was approved by the Ethics Committees of Second Xiangya Hospital of Central South University (2020-609). All patients involved signed informed consent before participating in this study. During February 2020 and September 2021, 33 primary lung cancer patients (including 9 females and 24 males, among 32 to 68 years old, details in Table S4) who intended for surgical removal at our department and were demonstrated to be LUSC pathologically were included. Fresh LUSC tissues and paired normal lung specimens were preserved by liquid nitrogen and then subjected to the analysis of genomic DNA or total RNA.

METHOD DETAILS

Data collection and processing

The schematic of this study is presented in Figure 1A. LUSC and LUAD DNA methylation data from The Cancer Genome Atlas (TCGA) was downloaded from The UCSC Xena database (https://xenabrowser.net/datapages/).[58] Relevant transcriptomic data, clinical data, and somatic mutation data (processed using the VarScan2 software) were extracted from the TCGA data portal (https://portal.gdc.cancer.gov/). To be noted, the methylation at each CpG site was represented with a β value ranging from 0 (no methylation) to 1 (100% methylation). Besides, four independent methylation datasets—GSE121849 (10 LUSC samples, 7 normal lung samples), GSE39279 (122 LUSC samples), GSE52401 (244 nontumor lung samples), and GSE108124 (36 progressive carcinoma *in situ* (CIS) samples, 18 regressive CIS samples, 33 normal lung samples)—were collected from the Gene Expression Omnibus (GEO) database. For diagnostic model development and validation, we assigned the TCGA-LUSC dataset, GSE121849 dataset, or the dataset merged by GSE39279 and GSE52401 as the training cohort, test cohort 1, or test cohort 2, respectively. All these methylation datasets were detected by the Illumina Human Methylation 450 platform.

The ChAMP software package was used for the data preprocessing. Probes that met the following criteria were excluded from our study group: (1) non-CpG probes; (2) multi-hit probes; (3) probes have single nucleotide polymorphism (SNP); (4) probes located in sex chromosomes; (5) have available methylation data in less than 90% of samples. Samples without information on American Joint Committee on Cancer (AJCC) were also excluded. The remaining missing data were imputed using the k-nearest neighbors imputation method.

Identification of DMPs, functional enrichment analysis, and exploration of CIMP

DMPs were identified by the ChAMP package based on β values of LUSC and adjacent samples in the training cohort, with the criteria of false discovery rate (FDR) < 0.05 and $|\Delta\beta| > 0.2$. To investigate the functions of DMPs, we used the Cytoscape (http://www.cytoscape.org/) plug-in ClueGO (http://apps.cytoscape.org/apps/cluego) to perform functional enrichment analysis. KEGG and Reactome databases

were selected. Mapped genes corresponding to the hypermethylated DMPs and hypomethylated DMPs with top 1000 |Δβ| were imported to the analysis, respectively.

Then, we chose the CpGs which located in the promoter region, had a large standard deviation (SD) in tumor tissues (SD > 0.2), and were hypomethylated in para-cancerous tissues (average β value < 0.05) to find the CIMP of LUSC. To identify CIMP, data dimensionality reduction and clustering methods according to a previous study were utilized.[20] Removing samples lacking sex, age, smoking history, somatic mutation, survival status, or over survival, CpGs which were located in the promoter region, had a large SD in tumor tissues (SD > 0.2), and were hypomethylated in para-cancerous tissues (average β value <0.05) were chosen for further analysis. Kmeans consensus clustering was performed on these CpGs. Principal Component Analysis (PCA) was utilized to re-verify the classification effect of clustering. Analysis of variance (ANOVA) was used to analyze the difference among clusters. The Kaplan Meier method and Log Rank test were performed to evaluate the OS of clusters. Fisher's exact test was used to compare clinical characteristics. To explore biological mechanisms behind each cluster, gene set enrichment analysis (GSEA) for corresponding TCGA transcriptomic data was performed using GSEA software (v4.3.2) with 1,000 permutations against C2.CP.KEGG, C2.CP.REACTOME, and C5.GO.BP genesets.

Model construction and evaluation in public databases

Since methylation of promoter regions significantly influences gene expression, CpG loci located 200 and 200–1500 bp upstream of the transcription start sites (TSS200 and TSS1500, respectively) were involved in subsequent feature selection. In order to identify CpGs of which aberrant methylation occurred in both the beginning occurrence phase and later phases of LUSC, we divided LUSC samples in the training cohort into the stage I subgroup and stage II-IV subgroup. All normal samples were added to both subgroups. In addition, to circumvent the problem caused by severely imbalanced data (normal samples << LUSC samples), normal samples were oversampled using the ROSE R package and the augmented subgroups with balanced class distribution were generated. Feature selection was carried out by the Boruta algorithm, a random-forest-based algorithm targeting the so-called "all-relevant problem".[59] Subsequently, CpGs selected by the algorithm were imported to nested five-fold cross-validation based on the random forest model, where the modeling process was performed five times for each selected CpG.[60] This process sequentially increased the number of CpGs (ranked by variable importances calculated by the Boruta algorithm) and it repeated five rounds. The mean cross-validation error was calculated and the classifier with the minimum error rate was chosen.

Test cohort 1 and test cohort 2 were used to verify the methylation levels of candidate CpGs and evaluate the performance of the random-forest-based classification model in LUSC. The sensitivity, specificity, and ROC curves were used to evaluate its classification performance. Subsequently, we analyzed the methylation differences of the identified CpGs among regressive CIS samples, progressive CIS samples, and normal samples using the GSE108124 dataset. TCGA-LUAD methylation data were also analyzed to evaluate their methylation statuses in LUAD samples.

DNA pyrosequencing assay

Bisulfite pyrosequencing provides an average methylated proportion of two alleles with great precision, which is one of the most reliable methods for DNA methylation validation, especially for shorter regions.[61,62] Here, we aimed to verify the methylation-based diagnostic model with our patients through pyrosequencing. First, genomic DNA was obtained from 10 paired LUSC tissues and adjacent nontumor tissues via the TIANamp Genomic DNA Kit (DP304-03, TIANGEN, Beijing, China). Then, DNA integrity was checked by the agarose gel electrophoresis technique. Bisulphite conversions were performed by utilizing the EpiTect Fast DNA Bisulfite Kit (DP304-0359826, Qiagen, Hilden, Germany) and polymerase chain reaction (PCR) was conducted through PyroMark PCR Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocols, where specific primers for PCR amplification (see Table S1) were designed and synthesized by Ouyi Biotech (Shanghai, China). Agarose gel electrophoresis was performed again to verify the quality of the PCR product and exclude the formation of primer dimers. A pyrosequencing assay was used to measure the methylation levels of selected methylated sites on the PyroMark Gold Q96 system (Qiagen, Hilden, Germany). CpG site quantification was analyzed with the PyroMark CpG Software 1.0.11.

Quantitative real-time PCR (qRT-PCR)

Total RNA was isolated from lung tissue samples by Trizol reagent (Invitrogen) and gotten concentration or quality checked on a NanoDrop UV-Vis Spectrophotometer (Thermo Scientific), then reversely transcribed into cDNA via the SuperScript First Strand cDNA system (Invitrogen). Primers applied are listed in Table S1. The qPCR amplifications were performed with an SYBR Green PCR Master Mix (Roche) in an Applied Biosystems Stepone Plus System (Applied Biosystems). The Delta Delta Ct ($\Delta\Delta$Ct) method was utilized for relative quantification and paired T-test was performed to obtain p values between two groups. The expression data of selected genes in TCGA were collected and analyzed by unpaired T-test to expound and prove the qRT-PCR results. GraphPad Prism 8.0 software (La Jolla, CA, USA) was utilized in plotting.

Protein immunohistochemistry (IHC) analysis

For the same patient involved in the DNA pyrosequencing assay, their paired LUSC samples and adjacent nontumor samples were obtained from the Department of Pathology of the Second Xiangya Hospital and validated by two pathologists. Paraffin sections from LUSC patient samples were firstly dewaxed and hydrated by xylene and ethanol. Then antigen retrieval was conducted in citrate buffer using a microwave for 10 min. After cooling to room temperature, the sections were incubated with normal goat serum blocking solution for 20 min, and subsequently stained with anti-TBX4 antibody (#sc-515196; Santa Cruz Biotechnology, Shanghai, China) at a dilution of 1/10, anti-TRIM15 antibody (#13623-1-AP; ProteinTech, Wuhan, China) at a dilution of 1/200, anti-C6orf201 antibody (#sc-514971; Santa Cruz Biotechnology, Shanghai, China) at a dilution of 1/500, and anti-ARHGEF4 antibody (#55213-1-AP; ProteinTech, Wuhan, China) at a dilution of 1/50. One hour later, the slides were thoroughly washed three times with PBS solution for 2 min each and then incubated for 15 min with the corresponding *MaxVision* kit (#KIT-5001 or #KIT-5004; Maixin Biol, Fuzhou, China) at room temperature. Next, the slides were thoroughly washed with PBS and dehydrated with ethanol of gradient concentrations. Finally, the whole slides were controlled with xylene and coverslipped. The images were surveyed and captured through a CX41 microscope (OLYMPUS, Tokyo, Japan) with the Microscope Digital Camera System DP-72 (OLYMPUS, Tokyo, Japan). Protein expression levels were presented by the mean integrated optical density (IOD/area).[63]

QUANTIFICATION AND STATISTICAL ANALYSIS

Graphs were created and analyzed using R software or GraphPad Prism software. Unless otherwise mentioned, statistical difference was determined by Student's *t* test and data were presented as mean $\pm$ standard error of mean (SEM). The p value <0.05 is considered significant.