


RESEARCH ARTICLE

Open Access



# A methylation-based nomogram for predicting survival in patients with lung adenocarcinoma

Xuelong Wang<sup>1†</sup>, Bin Zhou<sup>1†</sup>, Yuxin Xia<sup>2†</sup>, Jianxin Zuo<sup>1</sup>, Yanchao Liu<sup>1</sup>, Xin Bi<sup>1</sup>, Xiong Luo<sup>3</sup> and Chengwei Zhang<sup>1\*</sup> 

## Abstract

**Background:** DNA methylation alteration is frequently observed in Lung adenocarcinoma (LUAD) and may play important roles in carcinogenesis, diagnosis, and prognosis. Thus, this study aimed to construct a reliable methylation-based nomogram, guiding prognostic classification screening and personalized medicine for LUAD patients.

**Method:** The DNA methylation data, gene expression data and corresponding clinical information of lung adenocarcinoma samples were extracted from The Cancer Genome Atlas (TCGA) database. Differentially methylated sites (DMSs) and differentially expressed genes (DEGs) were obtained and then calculated correlation by Pearson correlation coefficient. Functional enrichment analysis and Protein-protein interaction network were used to explore the biological roles of aberrant methylation genes. A prognostic risk score model was constructed using univariate Cox and LASSO analysis and was assessed in an independent cohort. A methylation-based nomogram that included the risk score and the clinical risk factors was developed, which was evaluated by concordance index and calibration curves.

**Result:** We identified a total of 1362 DMSs corresponding to 471 DEGs with significant negative correlation, including 752 hypermethylation sites and 610 hypomethylation sites. Univariate Cox regression analysis showed that 59 DMSs were significantly associated with overall survival. Using LASSO method, we constructed a three-DMSs signature that was independent predictive of prognosis in the training cohort. Patients in high-risk group had a significant shorter overall survival than patients in low-risk group classified by three-DMSs signature (log-rank  $p = 1.9E-04$ ). Multivariate Cox regression analysis proved that the three-DMSs signature was an independent prognostic factor for LUAD in TCGA-LUAD cohort (HR = 2.29, 95%CI: 1.47–3.57,  $P = 2.36E-04$ ) and GSE56044 cohort (HR = 2.16, 95%CI: 1.19–3.91,  $P = 0.011$ ). Furthermore, a nomogram, combining the risk score with clinical risk factors, was developed with C-indexes of 0.71 and 0.70 in TCGA-LUAD and GSE56044 respectively.

**Conclusions:** The present study established a robust three-DMSs signature for the prediction of overall survival and further developed a nomogram that could be a clinically available guide for personalized treatment of LUAD patients.

**Keywords:** Lung adenocarcinoma, DNA methylation, Differentially methylated sites, Prognosis, Signature

\* Correspondence: [bjdlyxwk@163.com](mailto:bjdlyxwk@163.com)

<sup>†</sup>Xuelong Wang, Bin Zhou and Yuxin Xia contributed equally to this work.

<sup>1</sup>Department of Thoracic Surgery, Capital Medical University Electric Power Teaching Hospital, Beijing 100073, China

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

Lung cancer is the leading cause of cancer-related deaths worldwide [1], including two main types known as small-cell lung carcinoma (SCLC) and non-small-cell lung carcinoma (NSCLC). Lung adenocarcinoma (LUAD) is the most predominant subtype of NSCLC, with increased incidence over the past decades worldwide [2]. Despite recent advances in surgical techniques, radiotherapeutic interventions and combined chemotherapy strategies, the long-term survival rate of patients diagnosed with LUAD has not significantly improved [3]. Thus, it is indeed urgent to identify specific details regarding characteristic molecules in LUAD tissue to evaluate the prognosis of LUAD and develop strategies for personalized therapy.

DNA methylation, as the key element in epigenetic modifications, plays a significant role in the regulation of cellular functions and carcinogenesis. Increasing studies demonstrated that epigenetic alterations in DNA methylation were relevant to the progression and metastasis of LUAD [4–7]. Shen et al. demonstrates that the methylation status of homeobox A9 (*HOXA9*), keratin-associated protein 8–1 (*KRTAP8–1*), cyclin D1 (*CCND1*), and tubby-like protein 2 (*TULP2*) has great potential for the early recognition of LUAD in the undetermined lung nodules [8]. Seok et al. found that TGFBI promoter methylation is associated with poor prognosis in lung adenocarcinoma patients [9]. Furthermore, a prognostic DNA methylation signature was established by Sandoval et al. to distinguished patients with high- and low-risk early stage NSCLC, guiding the adjuvant chemotherapy [10]. Additionally, researchers suggested an internal CpG-based signature for survival prediction of lung adenocarcinoma patients. These researches demonstrated that the methylation level is deemed a crucial molecular biomarker for the diagnosis and prognosis of LUAD patients [11–13]. However, limited by either the current expertise on the association between the epigenetic modifications and clinical outcomes or lack of independent validation as small sample size, the identification of a robust prognostic DNA methylation signature is of considerable importance for LUAD patients.

In the present study, we extracted the DNA methylation data, gene expression data and corresponding clinical information of lung adenocarcinoma samples from The Cancer Genome Atlas (TCGA) database to select the differentially methylated sites (DMSs) corresponding to dysregulated genes and further explore the biological processes in which the aberrant methylation genes might be involved. Moreover, performing univariate Cox and LASSO analysis, we constructed a robust DMSs-based prognostic signature and validated the prognostic performance in an independent cohort extracted from Gene

Expression Omnibus (GEO). Furthermore, combing DMSs-based prognostic signature with clinical risk factors, we constructed a nomogram that could provide insight into regarding survival prediction and serve as a clinically available guide for personalized treatment of LUAD patients.

## Methods

### Data processing

All datasets and clinical information were described in Table 1 and Supplementary Table S1. The DNA methylation data (459 LUAD tissues and 30 normal tissues) and gene expression data (513 LUAD tissues and 59 normal tissues) of lung adenocarcinoma samples were extracted from TCGA (<https://cancergenome.nih.gov/>). Methylation beta-values derived from Illumina Infinium Human Methylation 450 BeadChip platform were extracted as site methylation measurements. The normalized count values of level 3 gene expression data derived from Illumina HiSeqV2 were extracted as gene expression measurements. Clinical information of 513 LUAD patients was obtained from TCGA. After corresponding patients with both methylation data and expression data, ninety-six LUAD patients were excluded because of unknown survival time, age, and tumor stage. Ultimately, 417 patients were retained in our study. An independent dataset (GSE56044 [14]) collected from GEO (<https://www.ncbi.nlm.nih.gov/geo/>) was used to test the prognostic ability, containing 82 LUAD patients with both methylation data and clinical information.

### Identification of differentially methylated sites

The differentially expressed genes (DEGs) were firstly selected between tumor and normal tissues using edgeR package in R. Multiple test corrections was performed using Benjamini & Hochberg's method and the cutoff values were set at the  $FDR < 0.05$  and  $|\log_2FC| > 2$ . Then, the methylation sites corresponding to these DEGs were selected. For each methylation site, we test the difference in methylation level between tumor and normal tissues to select differentially methylated site (DMS) by T-test with  $p < 0.05$ . More importantly, Pearson correlation analysis was performed to calculate the correlation between the methylation level of DMS and expression level of corresponding DEG. Such DMSs with significant negative correlation, which were thought to deeply influence the expression of corresponding DEGs, were selected for subsequent analysis.

### Functional enrichment analysis

Functional annotations of DEGs containing DMSs were performed using The Database for Annotation, Visualization and Integrated Discovery (DAVID, <https://david.ncifcrf.gov/>), which enriched gene oncology and

**Table 1** Cohorts analyzed in present study

	Training cohort (TCGA-LUAD)			Validation cohort (GSE56044)	
	Methylation data	Expression data	Clinical data	Methylation data	Clinical data
Normal	30	59	–	–	–
Tumor	417	417	417	82	82
Platform	Illumina HM450	Illumina HiSeqV2	–	Illumina HM450	–

pathways. Three categories, including biological processes, molecular function and cellular components, were involved in Gene ontology (GO). Kyoto Encyclopedia of Genes and Genomes (KEGG, <https://www.kegg.jp/>) was used to carry out the pathway enrichment, which is an essential database resource for a deep understanding of functions and biological process from large-scale molecular cohorts produced by high-throughput experimental technology. The criterion for significant enrichment was  $p < 0.05$ .

#### Protein-protein interaction (PPI) network

To further explore the interaction among the DEGs, the Search Tool for the Retrieval of Interacting Genes (STRING, <http://string-db.org/>), a database containing all known and predicted protein interactions, was used to identify a PPI network of DEGs. Each interaction was evaluated by combined score ranged from 0 to 1. The higher the combined score, the more reliable the interaction. In present study, we used a strict combined score  $> 0.7$  as the cut-off criterion to identify reliable interactions among the DEGs. The PPI network was visualized by Cytoscape software (version 3.7.0; [www.cytoscape.org](http://www.cytoscape.org)). Furthermore, the hub genes in PPI network were extracted using the cytoHubba application.

#### Construction of DMSs-based prognostic signature

The univariate Cox regression analysis was firstly performed to calculate the association between the methylation level of each DMS and patient's overall survival (OS) in training cohort. Those sites with  $P$ -values less than 0.05 were identified as prognosis-related DMSs. Then, using LASSO method to screen the prognosis-related DMSs and obtain an optimal model subsequently, the prognosis-related DMSs with coefficient not equal to 0 were retained as significant variables and a risk scoring model was established using the combination of weighted methylation values. The risk scores were calculated as shown in the following equation: Risk score = methylation of site 1 \*  $\beta_1$  + methylation of site 2 \*  $\beta_2$  + ...methylation of site n \*  $\beta_n$ .  $\beta_i$  is the regression coefficient of site  $i$ , which represents the contribution of site  $i$  to the prognostic risk score. Based on the equation, risk scores were calculated for LUAD patients in each cohort. Using the median risk score as the cutoff point,

patients were divided into low-risk (risk score below the median value) or high-risk (risk score above the median value) group correspondingly.

#### Development of DMSs-based nomogram

To translate the prognostic value of DMSs-based signature into clinical application, a nomogram, including the risk score and the clinical risk factors of LUAD patients evaluated by multivariate Cox proportional-hazards regression, was developed for predicting the 3- and 5-years OS in TCGA-LUAD cohort. The discriminatory ability of the nomogram was evaluated by calculating the concordance index (C-index), which is a measure of discrimination. Calibration plots were plotted to compare the observed and predicted probabilities for the nomogram.

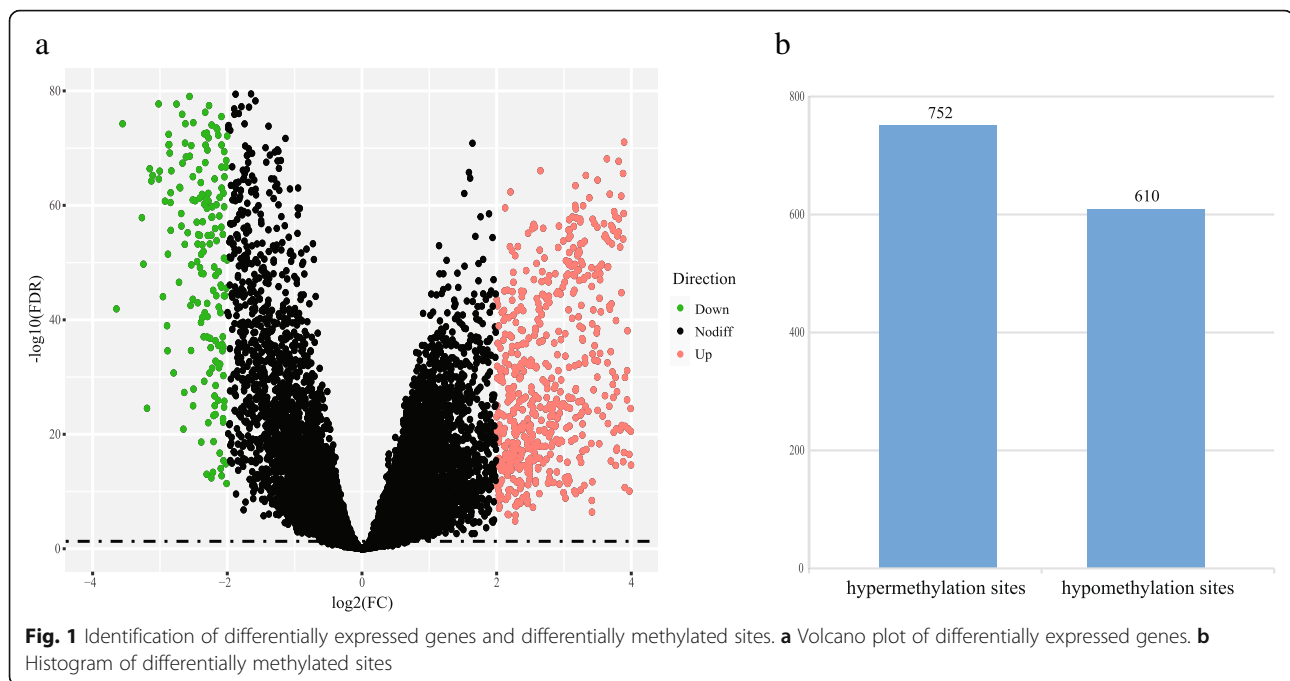
#### Statistical analysis

The multivariate Cox proportional-hazards regression model was used to evaluate the independent prognostic value of the signature after adjusting for age, sex and stage. Hazard ratios (HRs) and 95% confidence intervals (CIs) were computed based on the Cox regression analysis. Survival curves were estimated using the Kaplan–Meier method and were compared using the log-rank test. Fisher's exact test was used to observe the differences in mortality rate and lymph node metastasis rate between different risk groups. Values of  $p < 0.05$  were considered significant. All statistical analysis was performed using the R3.4.0.

## Results

#### Identification of differentially methylated sites in LUAD

We initially performed differential expression analysis to select DEGs between LUAD and normal lung tissues in TCGA-LUAD dataset. With cut-off criteria of  $FDR < 0.05$  and  $|\log_2FC| > 2.0$ , a total of 960 DEGs were identified, including 653 up-regulated DEGs and 307 down-regulated DEGs (Fig. 1A). We then selected the methylation sites which were differentially methylated between LUAD and normal lung tissues and significantly negatively correlated with the expression of corresponding DEGs. We thought that such methylation sites could influence the gene expression and further participate in tumor progression. The results showed that a total of 1362 DMSs



corresponding to 471 DEGs were identified, including 752 hypermethylation sites and 610 hypomethylation sites (Fig. 1B).

#### Functional enrichment of DMGs

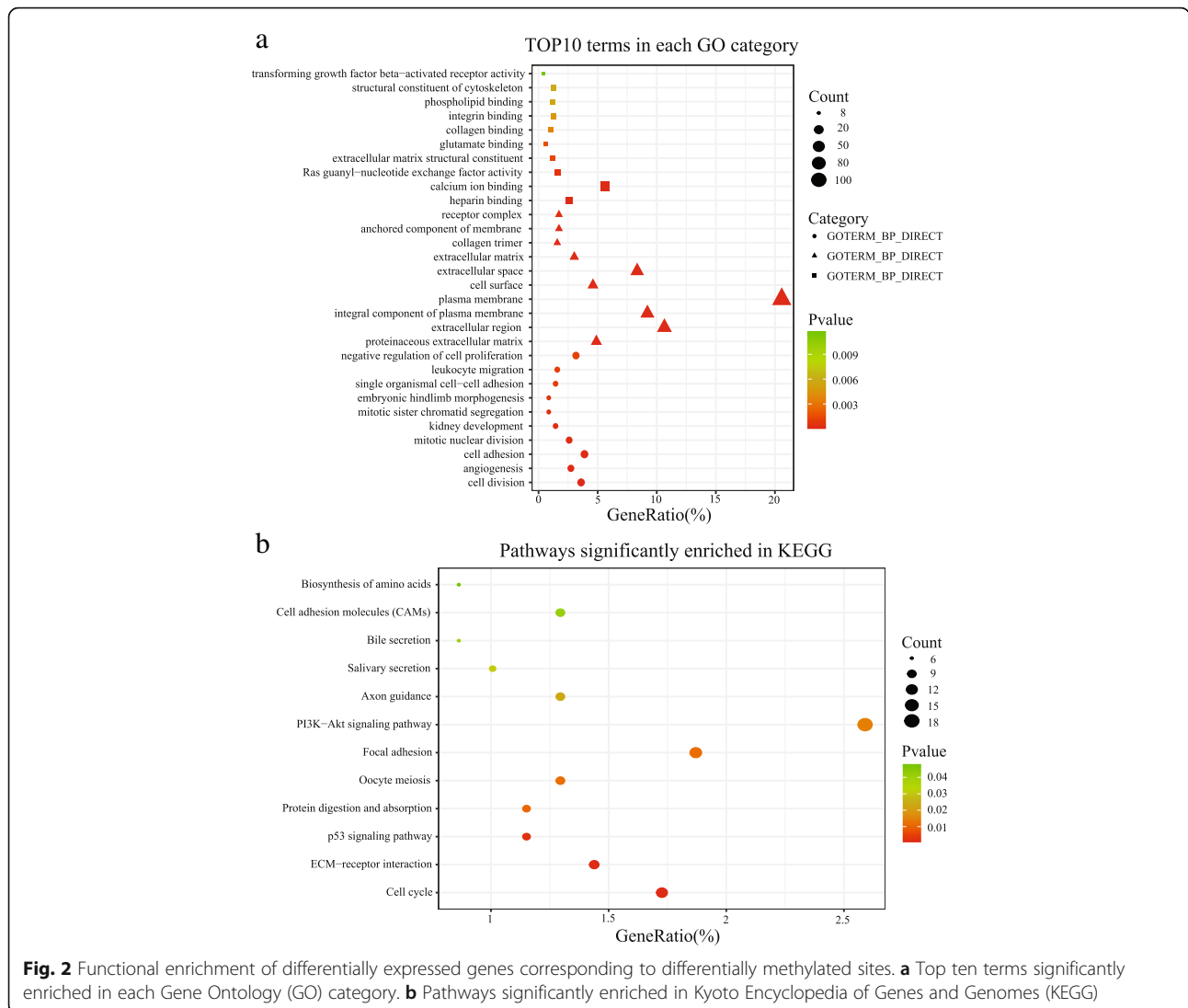
To further investigate the biological processes which the DMSs might be involved in, we performed GO annotation and KEGG pathway enrichment using DAVID database for the 471 corresponding DEGs. The DEGs were significantly enriched in many cancer-related pathways. The top significant terms emerging from the gene ontology enrichment analysis were shown in Fig. 2A. For instance, the most significant GO term, cell division, has been reported in multiple articles related to the progression and metastasis of cancer [15–17]. We also found that DEGs were significantly enriched in angiogenesis, which is a core hallmark of advanced cancers, especially in LUAD [18–20]. Besides, other significant GO terms, such as regulation of cell cycle and regulation of small GTPase mediated signal transduction were also related to cancer progression and chemoresistance reported in many studies [21, 22]. As shown in Fig. 2B, KEGG pathway enrichment analysis found twelve significantly enriched pathways related to cancer progression, such as PI3K-Akt signaling pathway [23, 24], ECM-receptor interaction [25, 26] and p53 signaling pathway [27, 28]. The results indicated that these DEGs played key roles in multiple cancer-related pathways, and further indicated that the DMSs might be involved in LUAD progression by regulating the corresponding gene expression.

#### Construction of PPI network

Using STRING database, a PPI network was constructed to further explore the interactions between the 471 DEMRNAs. After removing unconnected nodes, the PPI network of DEGs is consisted of 188 nodes and 888 edges when combined score > 0.7 was set as the cutoff criterion (Fig. 3A). Furthermore, the top 10 hub genes, including cyclin dependent kinase 1 (CDK1), cyclin A2 (CCNA2), cyclin B1 (CCNB1), cell division cycle 20 (CDC20), cell division cycle associated 8 (CDCA8), aurora kinase B (AURKB), assembly factor for spindle microtubules (ASPM), PDZ binding kinase (PBK), ribonucleotide reductase regulatory subunit M2 (RRM2) and centromere protein F (CENPF), were identified using the cytoHubba plugin for Cytoscape, with a higher degree of connectivity (Fig. 3B). Most of ten genes had been reported to be closely related to tumorigenesis and progression of LUAD.

#### Establishment of the DMSs-based prognostic signature

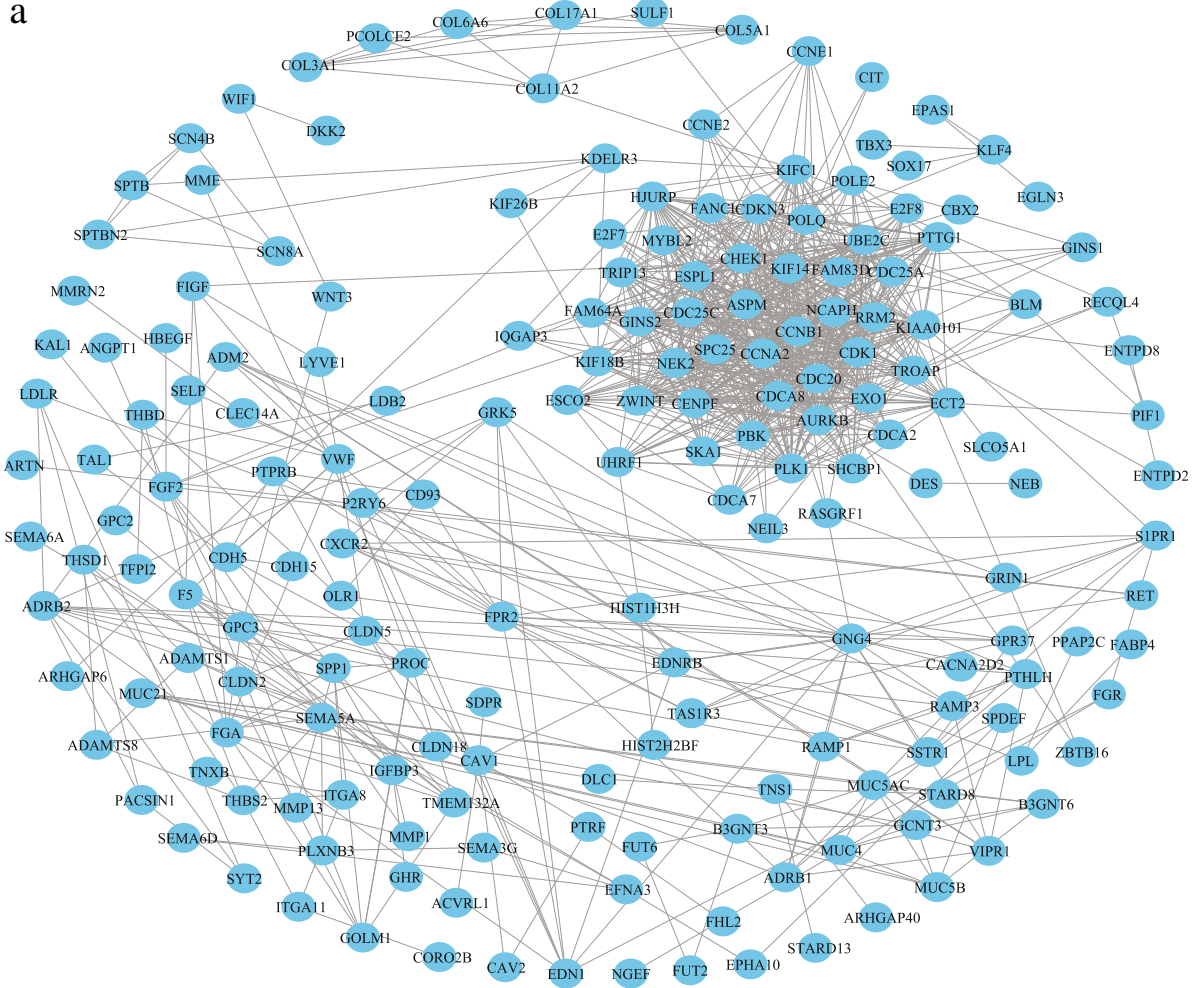
Performing the univariate Cox regression analysis, we identified DMSs with potential prognostic value in TCGA-LUAD cohort. Details of the clinical characteristics are presented in Supplementary Table S1. We found that 59 DMSs were significantly associated with overall survival, including 47 hypermethylation sites and 12 hypomethylation sites. The list of 59 DMSs is showed in Supplementary Table S2. Thus, these methylation sites were defined as prognosis-related DMSs to construct the prognostic signature. We used the glmnet package in R to perform LASSO regression analysis in TCGA-LUAD



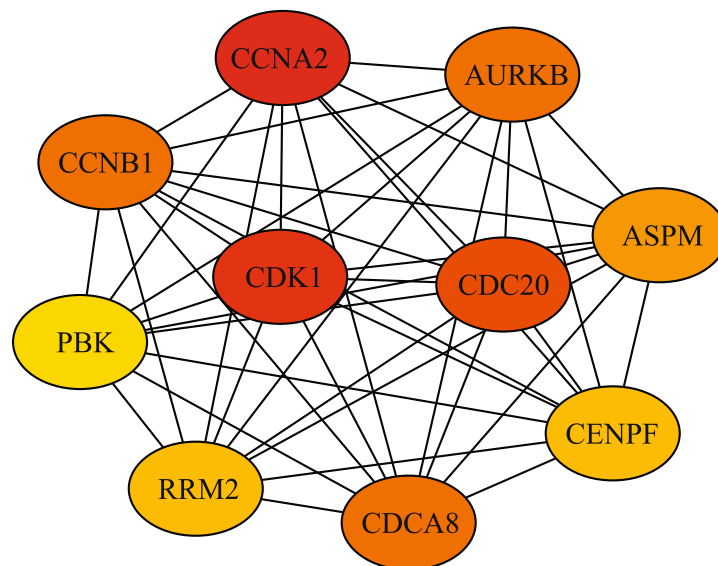
cohort. We obtained the optimal value of the parameter  $\lambda$ , which controlled the degree of LASSO regression complexity, and selected the significant variables through multiple cross-validation. We found that the parameter  $\lambda$  reached the optimal value, when the number of variables was three. Therefore, combining the regression coefficients of three DMSs under the optimal  $\lambda$  value, we constructed a three-DMSs risk score model to guide the prognosis of LUAD patients. The general information of the three DMSs is displayed in Table 2. The risk score formula was created as follows: Risk score = (1.0003\*methylation level of cg21339084) + (0.1484\*methylation level of cg07400091) + (- 0.2536\*methylation level of cg23843180). Calculating the risk scores for patients in TCGA-LUAD cohort, we classified patients into a high-risk or a low-risk group based on the median risk score. We found that the three-DMSs signature significantly stratified patients in terms of overall survival (log-rank  $p = 1.9E-04$ ; Fig. 4A). Patients with high risk scores had significantly shorter OS than those with low

risk scores. The mortality rate was 34.0% (71/209) in the high-risk group, significantly higher than 14.4% (30/208) in the low-risk group ( $p < 0.001$ , Fisher exact test; Fig. 4B). The risk score distribution, survival status, and methylation profile of the three prognostic DMSs are shown in Fig. 4C. As shown in Table 3, multivariate Cox regression analysis suggested that the three-DMSs signature was an independent prognostic factor, after adjusting for age, sex and stage (HR = 2.29, 95%CI: 1.47–3.57,  $P = 2.36E-04$ ). Furthermore, noticing the patients with lymph node metastasis status, we found that patients in the high-risk group had a higher lymph node metastasis rate than those in the low-risk group (26.2% vs. 15.8%,  $p = 0.018$ , Fisher exact test; Fig. 4B). From the three DMSs, two were associated with high risk (cg21339084 and cg07400091; HR > 1) and one appeared to be protective (cg23843180; HR < 1). The methylation level of the three prognostic DMSs was detected and the differences between high- and low-risk

**a**



**b**



**Fig. 3** Construction of protein-protein interaction (PPI) network and Identification of hub genes. **a** PPI network. **b** Ten hub genes extracted from PPI network

**Table 2** General information of the three DMSs

ProbeID	Gene	chrom	chromStart	chromEnd	coefficient
cg21339084	LIMS2	chr2	128,422,432	128,422,434	1.0003
cg07400091	S1PR1	chr1	101,704,472	101,704,474	0.1484
cg23843180	NGEF	chr2	233,852,838	233,852,840	-0.2536

groups were compared. We found that patients with high-risk scores tended to hypermethylation at risky sites, whereas patients in the low-risk group tended to hypomethylation at protective sites (Fig. 5A-C).

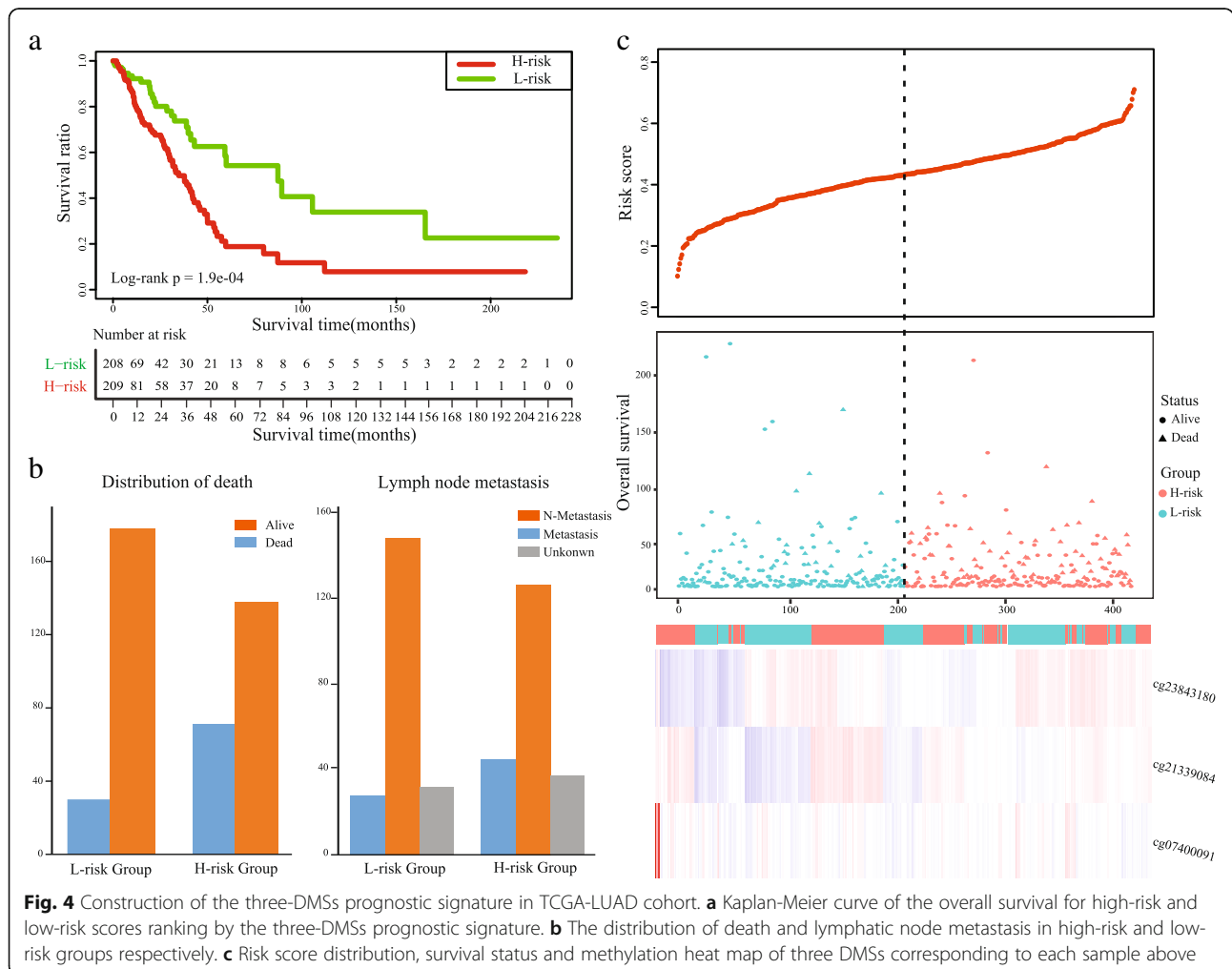
**Prognostic validation of the three-DMSs signature**

An independent cohort (GSE56044), containing 82 LUAD patients with both methylation data and clinical information, was used to validate the prognosis performance of the three-DMSs signature. Similarly, we calculated the risk score for each patient using the three-DMSs signature, after which patients were classified into a high-risk ( $n = 41$ ) or a low-risk ( $n = 41$ ) group based on the median risk score. We found that patients in high-

risk group had a shorter survival time than those in low-risk group (HR = 2.15, 95% CI: 1.20–3.85, log-rank  $p = 0.008$ , Fig. 6A). Furthermore, we calculated the mortality rate in each risk group. The result showed that the mortality rate in high-risk group was 32% higher than that in low-risk group ( $p = 0.006$ , Fisher exact test; Fig. 6B). The risk score distribution, survival status, and expression profile of the three prognostic DMSs are shown in Fig. 6C. As biased stage information, the stage variable is excluded when performed multivariate Cox regression analysis. In accordance with the result of training set, the multivariate Cox regression analysis confirmed that the three-DMSs signature was significantly correlated with overall survival as an independent prognostic factor (HR = 2.16, 95% CI: 1.19–3.91,  $P = 0.011$ , Table 3).

**Construction of three-DMSs signature-based nomogram**

Multivariate Cox analysis indicated that three variables (age, stage, and three-DMSs risk score) were independent risk factors for OS. Thus, a nomogram predicting 3- and 5-years OS was constructed based on the



**Fig. 4** Construction of the three-DMSs prognostic signature in TCGA-LUAD cohort. **a** Kaplan-Meier curve of the overall survival for high-risk and low-risk scores ranking by the three-DMSs prognostic signature. **b** The distribution of death and lymphatic node metastasis in high-risk and low-risk groups respectively. **c** Risk score distribution, survival status and methylation heat map of three DMSs corresponding to each sample above

**Table 3** Univariate and multivariate Cox regression analysis in TCGA-LUAD and GSE56044

Variables	Univariate analysis		Multivariate analysis	
	HR (95% CI)	P	HR (95% CI)	P
<b>TCGA-LUAD cohort</b>				
Age				
<= 60/> 60	0.96 (0.63–1.47)	0.852	1.19 (0.77–1.85)	0.437
Sex				
Male/Female	0.94 (0.64–1.40)	0.774	0.99 (0.66–1.49)	0.964
Stage				
I + II/III + IV	2.74 (1.82–4.13)	1.30e-06	2.79 (1.85–4.21)	1.06e-06
Risk score				
Low/High	2.22 (1.45–3.42)	2.77e-04	2.29 (1.47–3.57)	2.36e-04
<b>GSE56044 cohort</b>				
Age				
<= 60/> 60	2.83 (1.26–6.35)	0.012	2.88 (1.26–6.59)	0.012
Sex				
Male/Female	1.10 (0.63–1.93)	0.737	0.79 (0.44–1.43)	0.438
Risk score				
Low/High	2.15 (1.20–3.85)	0.010	2.16 (1.19–3.91)	0.011

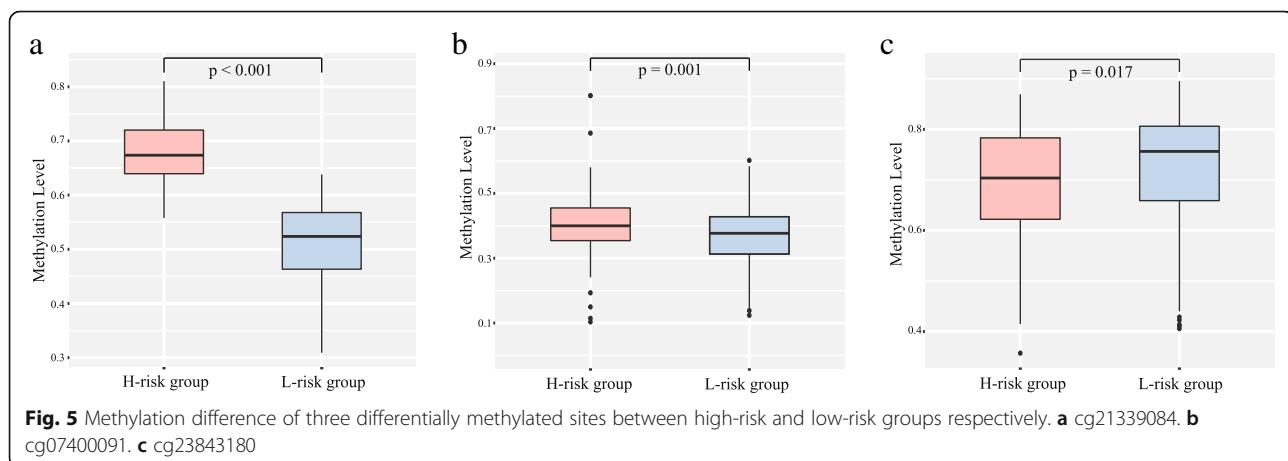
multivariate analysis data. As shown in Fig. 7, the total points for a patient can be obtained by adding the points from each independent prognostic factor listed in the nomogram. C-indexes for the nomogram were 0.71 (95%CI: 0.58–0.85) and 0.70 (95%CI: 0.52–0.88) in TCGA-LUAD and GSE56044 cohorts, respectively. The calibration plots for the probabilities of 3 and 5-year OS indicated no apparent departure from the ideal line, showing good agreement between the nomogram-predicted OS and actual OS of LUAD patients in both the training and validation cohorts (Fig. 8). Such results indicated that the three-DMSs signature-based nomogram could provide insight into regarding survival prediction and serve as a clinically available guide for personalized treatment of LUAD patients.

## Discussion

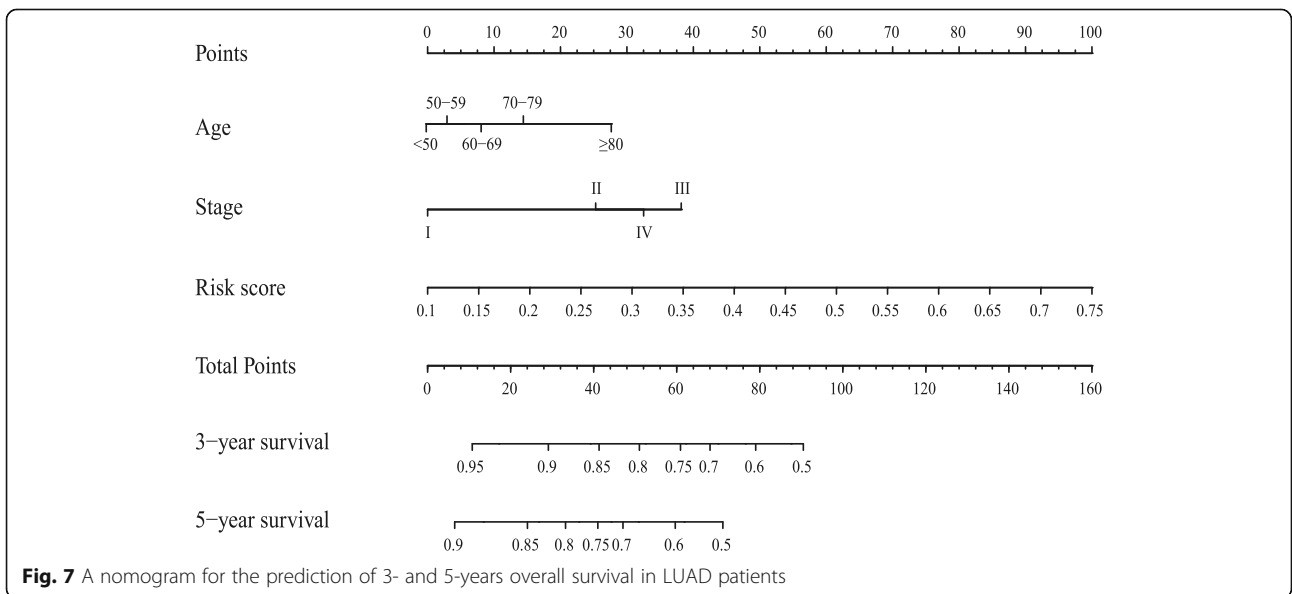
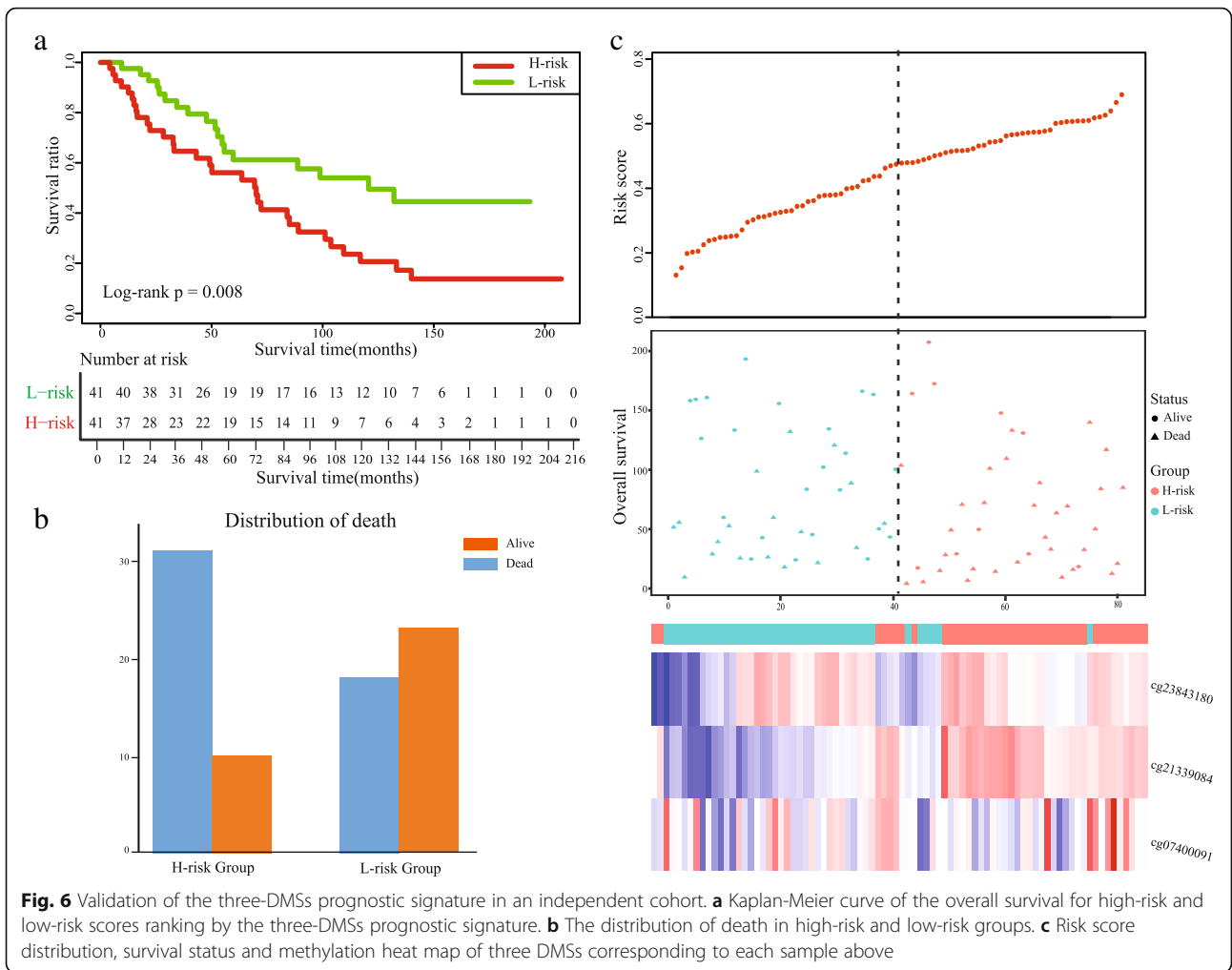
Due to the heterogeneity of LUAD, it is still a great challenge to develop successful individual-based treatment [29, 30]. Aberrant DNA methylation is of considerable importance in LUAD onset and progression [31, 32]. A special focus on DNA methylation alterations to develop the prognostic and predictive signatures for LUAD patients would be meaningful for survival prediction, guiding the personalized treatment decisions. Zheng et al. [11] constructed a CpG-based signature for survival prediction of lung adenocarcinoma patients based on TCGA database. However, such studies were limited by either small sample size or lack of validation of the signature as an independent prognostic factor. Therefore, in-depth studies on the LUAD progressive mechanisms, identification of specific methylation CpG sites and construction of the robust prognostic signatures are urgently required.

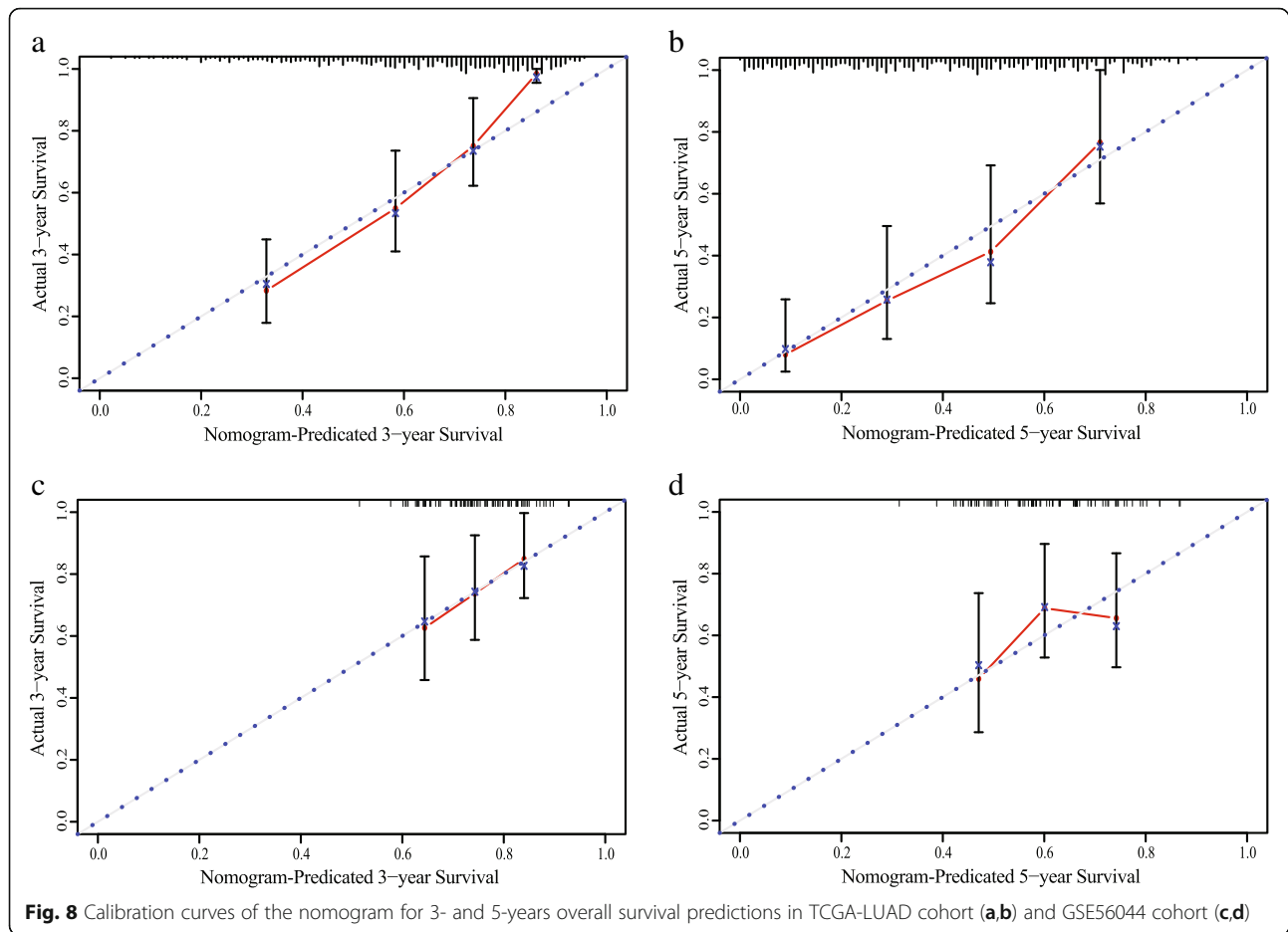
In the present study, we screened the DMSs that significantly correlated with corresponding gene expression, which may be involved in cancer progression by regulating the gene expression. Thus, a three-DMSs methylation signature significantly associated with the OS of LUAD patients was constructed based on genome-wide DNA methylation profiles using the Cox regression and LASSO analyses. The three-DMSs signature performed well in classifying patients into a high-risk or a low-risk group with significant survival difference. Furthermore, a nomogram was developed by combing the DMSs-based prognostic signature with clinical risk factors, which could provide a clinically available and robust guide for survival prediction and personalized treatment of LUAD patients.

Our study showed that three DMSs within prognostic signature had a critical role in progression and metastasis of LUAD. The three DMSs, including cg21339084, cg07400091 and cg23843180, correspond to LIMS2, S1PR1 and NGEF respectively (Table S3). We found that cg21339084 and cg07400091 were located in the S\_









Shore of CpG islands. The hypermethylation of cg21339084 and cg07400091 was significantly correlated with loss of expressions of LIMS2 and S1PR1. The cg23843180 was located in the 5'UTR of promoter, whose hypomethylation increased the expression of NGEF. Beside, we found that all three DMSs were located in DNase-I-hypersensitive sites (DHS) region, indicating the relationship between DNA methylation and DHS. Furthermore, we annotated all methylation probes of the three genes, and calculated the methylation difference and correlation with gene expression. The results showed that almost all 36 probes of LIMS2 were located in the promoter region and were hyper-methylated, indicating that loss of expressions of LIMS2 was significantly affected by promoter methylation. Many researches had demonstrated that frequent epigenetic silencing of LIMS2 could be important in GC progression events [33]. A total of 21 methylation probes of S1PR1 were located around the CpG islands. All probes were significantly hyper-methylated in LUAD samples except cg10020333, indicating that the hypermethylation of S1PR1 was closely related to LUAD progression. Previous study had shown that S1PR1 could act as methylation-driven genes to reveal prognostic biomarkers in LUAD [34]. Besides, 21

methylation probes were annotated within NGEF. NGEF is a novel member of the family of Dbl genes and functions as a guanine nucleotide exchange factor for the Rho-type GTPases. Few studies described its roles in carcinogenesis [35]. We found that the distributions and methylation levels of these probes were different, indicating that there might be multiple regulatory mechanisms in LUAD progression. These results indicated the aberrant methylation of three DMSs might play vital roles in promoting LUAD progression and metastasis, but the underlying mechanisms need further experimental verification.

In this study, we selected methylation sites that were significantly negatively correlated with the expression to ensure the regulatory effect on genes. However, several important methylation sites might be lost due to lack of significance. The expression of genes is regulated by lots of factors rather than methylation, such as mutation and copy number variation. For example, cg26500801 was located in CpG island of KEAP1. We found that cg26500801 was significantly hyper-methylated in LUAD samples. Previous research had confirmed the effect of methylation on KEAP1 transcription control across multiple histologies of lung cancer [36]. However, we found that the correlation between methylation level of

cg26500801 and expression level of KEAP1 is not significant. We observed that 17 % of samples had KEAP1 mutations. Recent studies demonstrated that KEAP1/NRF2 axis dysfunction is strongly related to tumor progression and chemo- and radiotherapy resistance of cancer cells [37]. Fabrizio et al. reported that epigenetic abnormalities were demonstrated as emerging mechanisms of KEAP1/NRF2 axis modulation in addition to the most frequently investigated point mutations in solid tumors [38]. Elshaer et al. also found that KEAP1 mutations were associated with DNA methylation changes capable of shaping regulatory network functions [39]. Similarly, cg00912625 is a methylation site that is located in CpG island of CNTN4. Our results showed that cg00912625 was also significantly hyper-methylated in LUAD samples. However, the correlation between methylation level of cg00912625 and expression level of CNTN4 is not significant. We found that CNTN4 loss accounted for distinctly higher proportion than its gain in LUAD samples, indicating that CNV might contribute to abnormal expression. Therefore, combining both epigenomic and transcriptomic changes along with genetic alterations may provide a better understanding of the molecular mechanisms associated with the progression of lung cancer and may help to provide better therapeutic approaches.

## Conclusion

Analyzing methylation and expression data comprehensively, our study identified a robust three-DMSs prognostic signature, which was significantly associated with the OS of LUAD patients. Furthermore, a nomogram was developed by combining the three-DMSs prognostic signature with clinical risk factors, which could provide a clinically available and robust guide for survival prediction and personalized treatment of LUAD patients. Further studies on the functional mechanism of the three DMSs could be carried out, which might provide helpful guidance for LUAD therapy as promising therapeutic targets in the near future.

## Abbreviations

LUAD: Lung adenocarcinoma; GEO: Gene expression omnibus; TCGA: The cancer genome atlas; DMSs: Differentially methylated sites; DEGs: Differentially expressed genes; GO: Gene ontology; KEGG: Kyoto encyclopaedia of genes and genomes; LASSO: The least absolute shrinkage and selection operator; OS: Overall survival; HR: Hazard ratios; CI: Confidence interval

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12885-021-08539-4>.

**Additional file 1: Table S1.** Clinical information analyzed in present study.

**Additional file 2: Table S2.** The list of prognosis-related DMSs.

**Additional file 3: Table S3.** Annotation of methylation probes corresponding to three target genes.

## Acknowledgments

We would like to thank TCGA and GEO databases for sharing the large amount of data.

## Authors' contributions

XW, BZ, YX and CZ conceived and designed the study. YL, XB and XL downloaded and screened the data. CZ and JZ edited the manuscript. All authors read and approved the final manuscript.

## Funding

This work was not funded by any grant.

## Availability of data and materials

All data generated or analyzed during this study are included in this published article.

## Declarations

### Ethics approval and consent to participate

Approval from the ethical board for this study was not required because of the public nature of all the data.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>Department of Thoracic Surgery, Capital Medical University Electric Power Teaching Hospital, Beijing 100073, China. <sup>2</sup>Department of emergency, Capital Medical University Electric Power Teaching Hospital, Beijing 100073, China. <sup>3</sup>Department of Internal Medicine, Beijing Nuclear Industry Hospital, Beijing 100822, China.

Received: 10 September 2020 Accepted: 28 June 2021

Published online: 12 July 2021

## References

- Hirsch FR, Scagliotti GV, Mulshine JL, Kwon R, Curran WJ Jr, Wu YL, et al. Lung cancer: current therapies and new targeted treatments. *Lancet*. 2017; 389(10066):299–311. [https://doi.org/10.1016/S0140-6736\(16\)30958-8](https://doi.org/10.1016/S0140-6736(16)30958-8).
- Zhou C. Lung cancer molecular epidemiology in China: recent trends. *Transl Lung Cancer Res*. 2014;3(5):270–9. <https://doi.org/10.3978/j.issn.2218-6751.2014.09.01>.
- Chansky K, Sculier JP, Crowley JJ, Giroux D, Van Meerbeeck J, Goldstraw P, et al. The International Association for the Study of Lung Cancer staging project. Prognostic factors and pathologic TNM stage in surgically managed non-small cell lung cancer. *Zhongguo Fei Ai Za Zhi*. 2010;13(1):9–18. <https://doi.org/10.3779/j.issn.1009-3419.2010.01.02>.
- Gao C, Zhuang J, Li H, Liu C, Zhou C, Liu L, et al. Exploration of methylation-driven genes for monitoring and prognosis of patients with lung adenocarcinoma. *Cancer Cell Int*. 2018;18(1):194. <https://doi.org/10.1186/s12935-018-0691-z>.
- Song X, Zhao C, Jiang L, Lin S, Bi J, Wei Q, et al. High PITX1 expression in lung adenocarcinoma patients is associated with DNA methylation and poor prognosis. *Pathol Res Pract*. 2018;214(12):2046–53. <https://doi.org/10.1016/j.prp.2018.09.025>.
- Su C, Shi K, Cheng X, Han Y, Li Y, Yu D, et al. Methylation of CLEC14A is associated with its expression and lung adenocarcinoma progression. *J Cell Physiol*. 2019;234(3):2954–62. <https://doi.org/10.1002/jcp.27112>.
- Zhang R, Lai L, Dong X, He J, You D, Chen C, et al. SIPA1L3 methylation modifies the benefit of smoking cessation on lung adenocarcinoma survival: an epigenomic-smoking interaction analysis. *Mol Oncol*. 2019;13(5): 1235–48. <https://doi.org/10.1002/1878-0261.12482>.

8. Shen N, Du J, Zhou H, Chen N, Pan Y, Hoheisel JD, et al. A diagnostic panel of DNA methylation biomarkers for lung adenocarcinoma. *Front Oncol*. 2019;9:1281. <https://doi.org/10.3389/fonc.2019.01281>.
9. Seok Y, Lee WK, Park JY, Kim DS. TGFBI promoter methylation is associated with poor prognosis in lung adenocarcinoma patients. *Mol Cells*. 2019;42(2): 161–5. <https://doi.org/10.14348/molcells.2018.0322>.
10. Sandoval J, Mendez-Gonzalez J, Nadal E, Chen G, Carmona FJ, Sayols S, et al. A prognostic DNA methylation signature for stage I non-small-cell lung cancer. *J Clin Oncol*. 2013;31(32):4140–7. <https://doi.org/10.1200/JCO.2012.4.85516>.
11. Zheng R, Xu H, Mao W, Du Z, Wang M, Hu M, et al. A novel CpG-based signature for survival prediction of lung adenocarcinoma patients. *Exp Ther Med*. 2020;19(1):280–6. <https://doi.org/10.3892/etm.2019.8200>.
12. Wang R, Zhu H, Yang M, Zhu C. DNA methylation profiling analysis identifies a DNA methylation signature for predicting prognosis and recurrence of lung adenocarcinoma. *Oncol Lett*. 2019;18(6):5831–42. <https://doi.org/10.3892/ol.2019.10931>.
13. Wang Y, Wang Y, Wang Y, Zhang Y. Identification of prognostic signature of non-small cell lung cancer based on TCGA methylation data. *Sci Rep*. 2020; 10(1):8575. <https://doi.org/10.1038/s41598-020-65479-y>.
14. Karlsson A, Jonsson M, Lauss M, Brunnstrom H, Jonsson P, Borg A, et al. Genome-wide DNA methylation analysis of lung carcinoma reveals one neuroendocrine and four adenocarcinoma epitypes associated with patient outcome. *Clin Cancer Res*. 2014;20(23):6127–40. <https://doi.org/10.1158/1078-0432.CCR-14-1087>.
15. Lopez-Lazaro M. The stem cell division theory of cancer. *Crit Rev Oncol Hematol*. 2018;123:95–113. <https://doi.org/10.1016/j.critrevonc.2018.01.010>.
16. Shostak A. Circadian Clock, Cell Division, and Cancer: From Molecules to Organism. *Int J Mol Sci*. 2017;18(4):873. <https://doi.org/10.3390/ijms18040873>.
17. Xia Z, Ou-Yang W, Hu T, Du K. Prognostic significance of CDC25C in lung adenocarcinoma: an analysis of TCGA data. *Cancer Genet*. 2019;233–234:67–74. <https://doi.org/10.1016/j.cancergen.2019.04.001>.
18. Cong Z, Diao Y, Li X, Jiang Z, Xu Y, Zhou H, et al. Long non-coding RNA linc00665 interacts with YB-1 and promotes angiogenesis in lung adenocarcinoma. *Biochem Biophys Res Commun*. 2020;527(2):545–52. <https://doi.org/10.1016/j.bbrc.2020.04.108>.
19. Frezzetti D, Gallo M, Maiello MR, D'Alessio A, Esposito C, Chicchinelli N, et al. VEGF as a potential target in lung cancer. *Expert Opin Ther Targets*. 2017; 21(10):959–66. <https://doi.org/10.1080/14728222.2017.1371137>.
20. Popper HH. Progression and metastasis of lung cancer. *Cancer Metastasis Rev*. 2016;35(1):75–91. <https://doi.org/10.1007/s10555-016-9618-0>.
21. Chen B, Huang S, Pisanic II TR, Stark A, Tao Y, Cheng B, et al. Rab8 GTPase regulates klotho-mediated inhibition of cell growth and progression by directly modulating its surface expression in human non-small cell lung cancer. *EBioMedicine*. 2019;49:118–32. <https://doi.org/10.1016/j.ebiom.2019.10.040>.
22. Liu M, Zhang H, Li Y, Wang R, Li Y, Zhang H, et al. HOTAIR, a long noncoding RNA, is a marker of abnormal cell cycle regulation in lung cancer. *Cancer Sci*. 2018;109(9):2717–33. <https://doi.org/10.1111/cas.13745>.
23. Fumarola C, Bonelli MA, Petronini PG, Alfieri RR. Targeting PI3K/AKT/mTOR pathway in non small cell lung cancer. *Biochem Pharmacol*. 2014;90(3):197–207. <https://doi.org/10.1016/j.bcp.2014.05.011>.
24. Zhu HE, Yin JY, Chen DX, He S, Chen H. Agmatinase promotes the lung adenocarcinoma tumorigenesis by activating the NO-MAPKs-PI3K/Akt pathway. *Cell Death Dis*. 2019;10(11):854. <https://doi.org/10.1038/s41419-019-2082-3>.
25. Bao Y, Wang L, Shi L, Yun F, Liu X, Chen Y, et al. Transcriptome profiling revealed multiple genes and ECM-receptor interaction pathways that may be associated with breast cancer. *Cell Mol Biol Lett*. 2019;24(1):38. <https://doi.org/10.1186/s11658-019-0162-0>.
26. Yeh MH, Tzeng YJ, Fu TY, You JJ, Chang HT, Ger LP, et al. Extracellular matrix-receptor interaction signaling genes associated with inferior breast Cancer survival. *Anticancer Res*. 2018;38(8):4593–605. <https://doi.org/10.21873/anticancer.12764>.
27. Zeng SG, Lin X, Liu JC, Zhou J. Hypoxia-induced internalization of connexin 26 and connexin 43 in pulmonary epithelial cells is involved in the occurrence of nonsmall cell lung cancer via the P53/MDM2 signaling pathway. *Int J Oncol*. 2019;55(4):845–59. <https://doi.org/10.3892/ijo.2019.4867>.
28. Zhong G, Chen X, Fang X, Wang D, Xie M, Chen Q. Fra-1 is upregulated in lung cancer tissues and inhibits the apoptosis of lung cancer cells by the P53 signaling pathway. *Oncol Rep*. 2016;35(1):447–53. <https://doi.org/10.3892/or.2015.4395>.
29. Chen Z, Fillmore CM, Hammerman PS, Kim CF, Wong KK. Non-small-cell lung cancers: a heterogeneous set of diseases. *Nat Rev Cancer*. 2014;14(8): 535–46. <https://doi.org/10.1038/nrc3775>.
30. Greulich H. The genomics of lung adenocarcinoma: opportunities for targeted therapies. *Genes Cancer*. 2010;1(12):1200–10. <https://doi.org/10.1177/1947601911407324>.
31. Duruisseaux M, Esteller M. Lung cancer epigenetics: from knowledge to applications. *Semin Cancer Biol*. 2018;51:116–28. <https://doi.org/10.1016/j.semcancer.2017.09.005>.
32. Mehta A, Dobersch S, Romero-Olmedo AJ, Barreto G. Epigenetics in lung cancer diagnosis and therapy. *Cancer Metastasis Rev*. 2015;34(2):229–41. <https://doi.org/10.1007/s10555-015-9563-3>.
33. Kim SK, Jang HR, Kim JH, Noh SM, Song KS, Kim MR, et al. The epigenetic silencing of LIMS2 in gastric cancer and its inhibitory effect on cell migration. *Biochem Biophys Res Commun*. 2006;349(3):1032–40. <https://doi.org/10.1016/j.bbrc.2006.08.128>.
34. Li R, Yang YE, Yin YH, Zhang MY, Li H, Qu YQ. Methylation and transcriptome analysis reveal lung adenocarcinoma-specific diagnostic biomarkers. *J Transl Med*. 2019;17(1):324. <https://doi.org/10.1186/s12967-019-2068-z>.
35. Rodrigues NR, Theodosiou AM, Nesbit MA, Campbell L, Tandle AT, Saranath D, et al. Characterization of Ngef, a novel member of the Dbl family of genes expressed predominantly in the caudate nucleus. *Genomics*. 2000; 65(1):53–61. <https://doi.org/10.1006/geno.2000.6138>.
36. Fabrizio FP, Sparaneo A, Centra F, Trombetta D, Storlazzi CT, Graziano P, et al. Methylation density pattern of KEAP1 gene in lung cancer cell lines detected by quantitative methylation specific PCR and pyrosequencing. *Int J Mol Sci*. 2019;20(11):2697. <https://doi.org/10.3390/ijms20112697>.
37. Taguchi K, Motohashi H, Yamamoto M. Molecular mechanisms of the Keap1-Nrf2 pathway in stress response and cancer evolution. *Genes Cells*. 2011;16(2):123–40. <https://doi.org/10.1111/j.1365-2443.2010.01473.x>.
38. Fabrizio FP, Sparaneo A, Trombetta D, Muscarella LA. Epigenetic versus genetic deregulation of the KEAP1/NRF2 Axis in solid tumors: focus on methylation and noncoding RNAs. *Oxidative Med Cell Longev*. 2018;2018: 2492063.
39. Elshaer M, ElManawy AI, Hammad A, Namani A, Wang XJ, Tang X. Integrated data analysis reveals significant associations of KEAP1 mutations with DNA methylation alterations in lung adenocarcinomas. *Aging (Albany NY)*. 2020;12(8):7183–206. <https://doi.org/10.18632/aging.103068>.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

