A risk score system based on DNA methylation levels and a nomogram survival model for lung squamous cell carcinoma

MING ZHANG, LIBING SUN, YI RU, SHASHA ZHANG, JUNJUN MIAO, PENGDA GUO, JINGHUAN LV, FENG GUO and BIAO LIU

Department of Oncology, The Affiliated Suzhou Municipal Hospital of Nanjing Medical University, Suzhou, Jiangsu 215001, P.R. China

Received September 14, 2019; Accepted January 30, 2020

DOI: 10.3892/ijmm.2020.4590

Correspondence to: Dr Feng Guo or Dr Biao Liu, Department of Oncology, The Affiliated Suzhou Municipal Hospital of Nanjing Medical University, 16 Baita West Road, Suzhou, Jiangsu 215001, P.R. China

E-mail: guofeng27@suda.edu.cn
E-mail: liuyrew1234@hotmail.com

Abbreviations: AUC, area under the receiver operating characteristic curve; DEG, differentially expressed gene; DMG, differentially methylated gene; DM-IncRNA, differentially methylated IncRNA; FDR, false discovery rate; FC, fold change; HNSC, head and neck squamous cell carcinoma; KM, Kaplan-Meier analysis; IncRNA, long non-coding RNA; LScc, lung squamous cell carcinoma; NSCLC, non-small cell lung cancer; LUAD, lung adenocarcinoma; PCC, Pearson correlation coefficient; RFS, recurrence-free survival; RS, risk score; ROC, receiver operating characteristic; TCGA, The Cancer Genome Atlas

Key words: lung squamous cell carcinoma, differential expression analysis, risk score system, nomogram survival model, pathway enrichment analysis

Abstract. Lung squamous cell carcinoma (LScc) is one of the primary types of non-small cell lung carcinoma, and patients with recurrent LScc usually have a poor prognosis. The present study was conducted to build a risk score (RS) system for LScc. Methylation data on LScc (training set) and on head and neck squamous cell carcinoma (validation set 2) were obtained from The cancer Genome Atlas database, and GSE39279 (validation set 1) was retrieved from the Gene Expression Omnibus database. differentially methylated protein-coding genes (dMGs)/long non-coding RNAs (dM-lncRNAs) between recurrence-associated samples and nonrecurrence samples were screened out using the limma package, and their correlation analysis was conducted using the cor.test() function. Following identification of the optimal combinations of dMGs or dM-lncRNAs using the penalized package in R, RS systems were built, and the system with optimal performance was selected. Using the rms package, a nomogram survival model was then constructed. For the differentially expressed genes (DEGs) between the high- and low-risk groups, pathway enrichment analysis was performed by Gene Set Enrichment Analysis. There were 335 DMGs and DM-IncRNAs in total. Following screening out of the top 10 genes (aldehyde dehydrogenase 7 family member A1, chromosome 8 open reading frame 48, cytokine-like 1, heat shock protein 90 alpha family class A member 1, isovaleryl-CoA dehydrogenase, phosphodiesterase 3A, PNMA family member 2, SAM domain, SH3 domain and nuclear localization signals 1, thyroid hormone receptor interactor 13 and zinc finger protein 878) and 6 top IncRNAs, RS systems were constructed. According to Kaplan-Meier analysis, the DNA methylation level-based RS system exhibited the best performance. In combination with independent clinical prognostic factors, a nomogram survival model was built and successfully predicted patient survival. Furthermore, 820 DEGs between the high- and low-risk groups were identified, and 3 pathways were identified to be enriched in this gene set. The 10-DMG methylation level-based RS system and the nomogram survival model may be applied for predicting the outcomes of patients with LScc.

Introduction

Lung cancer is a common cancer worldwide, with the second highest mortality rate among females and the highest mortality rate among males (1,2). Non-small cell lung cancer (NSCLC) is the primary lung cancer type that accounts for ~85% of lung cancer cases and includes the following common pathological subtypes: Lung squamous cell carcinoma (LScc), lung adenocarcinoma (LUAD), and lung large cell carcinoma (3,4). LScc is more common among males, potentially due to smoking habits (5). LScc has distinct clinical characteristics, and the prognosis of patients with recurrent LScc is poor (6). Therefore, it is urgent to identify new molecular factors predictive of LScc treatment outcomes.

Methylation changes in certain tumor-associated genes have been identified in previous studies examining tumorigenesis, suggesting that they are critical risk factors of tumorigenesis and molecular markers for early diagnosis (7). Promoter methylation of PR/SET domain 5 is significantly associated with lymph node metastasis and tumor differentiation status of LScc; therefore, this gene is a candidate target for the
diagnosis, prognosis and treatment of this cancer (8). The drought-repressed 4 gene has been proposed for methylation status analysis of LSCC cells, and its low expression correlates with a poor patient prognosis (9). Tripartite motif-containing 58/cg26157385 methylation is associated with expression of 8 prognosis-associated genes in LSCC, indicating its potential regulatory role in the progression of LSCC (10).

Long non-coding RNAs (lncRNAs) are a class of RNAs measuring >200 nucleotides and have pivotal functions in the progression of numerous cancer types (11). In addition, increasing numbers of lncRNAs with a prognostic value have been described, for example, urothelial cancer associated and long intergenic non-coding RNA-p21 (12-14). The lncRNAs cancer susceptibility 2 and surfactant associated 1, pseudogene participate in the regulation of tumor suppressor genes and oncogenes during the formation of LSCC and may be used for the diagnosis, prognosis and targeted treatment of this disease (15). Several lncRNA combination signatures have been identified to predict LSCC prognosis (16,17). In LUAD, DNA methylation has regulatory effects on the role of lncRNA (18); however, to the best of our knowledge, this type of regulation in LSCC has rarely been reported.

Despite these valuable data, the predictive performance of methylated DNA sequences or lncRNAs is rarely compared, and additional potential biomarkers should be investigated to improve cancer prognosis. In the present study, the methylation levels of DNA sequences and lncRNAs in patients with LSCC were compared, and the levels identified to be significantly correlated with LSCC prognosis were screened out. Furthermore, different risk score (RS) systems were built, and the one with the best predictive performance was selected.

Materials and methods

Data sources. Methylation data from patients with LSCC (downloaded on March 25, 2019; platform: Illumina Infinium Human Methylation 450 BeadChip) including 372 LSCC samples and 43 healthy tissue samples, were retrieved from The Cancer Genome Atlas (TCGA; https://cancergenome.nih.gov/) database. Following collation with clinical information, 293 LSCC samples with data recurrence were selected as the training set.

With ‘lung cancer’ and ‘Homo sapiens’ as key words, a search for methylation datasets was conducted in the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) database (19) using the following criteria: i) Available histological data that could differentiate adenocarcinoma from squamous cell carcinoma; ii) total sample size was not <150; and iii) the clinical information associated with the LSCC samples contained data on actual future recurrence. Following this search, an eligible methylation dataset (accession no., GSE39279 (20); platform: Illumina HumanMethylation450 BeadChip; validation set 1) was selected and downloaded. There were 444 samples in the dataset, including 43 LSCC samples with data concerning future recurrence.

In addition, the methylation data on head and neck squamous cell carcinoma (HNSC), including 530 HNSC samples and 50 healthy tissue samples, were also retrieved from TCGA to investigate whether the predictive model for LSCC was applicable to other squamous carcinoma types. Following collation with the clinical information downloaded from TCGA simultaneously, 382 HNSC samples with information on future recurrence were selected as validation set 2.

Differential methylation analysis. According to the data corresponding to probe locations and IDs provided in the downloaded annotation files that contained information about protein-coding genes and non-coding RNAs, like gene and lncRNA function, DNA location site and pathway information, corresponding lncRNAs and genes in the methylation data were annotated using the HUGO Gene Nomenclature Committee (http://www.genenames.org/) database (21), which contains records on 4,112 lncRNAs and 19,201 genes.

On the basis of the recurrence data, the LSCC samples in the training set were classified into recurrence-associated (the tumor recurred) and nonrecurrence (the tumor did not recur) groups. Using the limma package (22) (v.3.34.7, https://bioconductor.org/packages/release/bioc/html/limma.html) in the R software (23), differentially methylated genes (DMGs) between the recurrence-associated and nonrecurrence samples were identified. The thresholds for significance were defined as $|\log_{2}(\text{fold change})|>0.263$ and false discovery rate (FDR) <0.05. Following the addition of data on methylation levels of the DMGs in the training set, bidirectional hierarchical clustering was performed on the DMGs via the pheatmap package (v1.0.8, https://cran.r-project.org/web/packages/pheatmap/index.html) in R (24).

Analysis of correlation between the methylation levels and expression levels of the DMGs. Data on the methylation levels and the matched expression levels of the DMGs were extracted, and their Pearson correlation coefficients (PCCs) were calculated using the cor.test() function (https://stat.ethz.ch/R-manual/R-devel/library/stats/html/cor.test.html) with P<0.05 as the significance threshold. DMGs with negative PCCs were selected for further analyses.

Construction of RS systems. Using univariate and multivariate Cox regression analyses in R package survival (v3.1-8; https://www.rdocumentation.org/packages/survival/versions/3.1-8), differentially methylated lncRNAs (DM-lncRNAs) and DMGs significantly associated with overall survival and independent prognosis factors were next screened out. A log-rank $P<0.05$ was selected as the cutoff criterion. Based on the DMGs correlating with independent prognosis, an optimal gene combination and optimal lncRNA combination were identified using the LASSO Cox regression model (26) in penalized package in R (27) (v.0.9.50; https://cran.r-project.org/web/packages/penalized/index.html). Optimal parameter ‘lambda’ in the model was calculated by the 1,000-fold cross-validation likelihood (cvl) method.

In combination with prognostic coefficients of the optimal DM-lncRNAs and DMGs, methylation status-based or methylation level-based RS systems were built.

For the optimal combinations, cutoff values of methylation levels of the optimal lncRNAs and genes were computed via the X-Tile Bio-Informatics Tool (https://medicine.yale.edu/lab/rimm/research/software.aspx) (28). The cutoff values were determined based on the following criterion: Monte Carlo $P<0.05$. The status of each sample at an lncRNA or DNA methylation level was defined in accordance with the cutoff value of
expression analysis of the high- and low-risk groups was performed to identify differentially expressed genes (DEGs) according to FDR <0.05 and |log

RS systems based on the methylation levels of optimal lncRNAs or optimal genes were constructed. The formula for calculating the RS of each sample was as follows:

\[
\text{Methylation RS} = \sum \beta_{\text{RNAs}} \times \text{Status}_{\text{RNAs}}^a;
\]

where \( \beta \) and \( \text{Status} \) represent the regression coefficient of an lncRNA or gene and the methylation status variable, respectively. The status RS systems, based on the methylation status of the optimal genes and optimal lncRNAs, were constructed separately as two separate status RS systems.

Samples in the training set were subdivided into high- and low-risk groups based on the median of the RSs of the two status RS systems and the two methylation RS systems. Using the Kaplan-Meier (KM) survival curve constructed by means of R package survival (29), associations between the prognosis and RS systems were analyzed. Concomitantly, the two validation sets were used to validate the results obtained. Finally, from the four RS systems generated, the system with optimum performance was selected through comparison of the parameters in the training set and the two validation sets.

Construction of a nomogram survival model. In the training set, independent prognostic factors were selected by univariate and multivariate Cox regression analyses via the survival package (29). The cutoff criterion was log-rank P<0.05. According to the risk data determined by the RS system, the associations between the independent clinical prognostic factors and prognosis were then analyzed. Using the rms package (v.5.1-2, https://cran.r-project.org/web/packages/rms/index.html) (30) in R, a 5-year nomogram survival model involving the independent clinical prognostic factors was built.

Pathway enrichment analysis. RSs of the best RS system were used to classify the samples in the training set into high- and low-risk groups. Using the limma package (23), differential expression analysis of the high- and low-risk groups was performed to identify differentially expressed genes (DEGs) between the two risk groups. The DEGs were screened according to FDR <0.05 and log FC>0.263 cut-off values. Using Gene Set Enrichment Analysis (http://software.broadinstitute.org/gsea/index.jsp) (31), pathway enrichment analysis of the DEG set was conducted. The pathways with P<0.05 were identified as significant.

Results

Differential-methylation analysis. On the basis of the annotation files, 1,028 IncRNAs and 15,544 genes were selected from the methylation dataset. According to the recurrence data, samples in the training set were subdivided into the recurrence-associated (78 LSCC samples from patients who later experienced cancer recurrence) and nonrecurrence (215 LSCC samples) groups. There were 335 DMGs and DM-lncRNAs between the recurrence-associated and nonrecurrence groups, including 27 DM-IncRNAs, 4 hypomethylated and 23 hypermethylated IncRNAs, and 308 DMGs, 139 hypomethylated and 169 hypermethylated genes (Fig. 1A). The bidirectional hierarchical clustering heatmap for the DMGs and GM-IncRNAs is presented in Fig. 1B.

Analysis of the correlation between the methylation levels and expression levels of the DMGs or DM-IncRNAs. As aforementioned, the matched methylation samples from the LSCC mRNA-seq samples were first selected, and the methylation and expression data on the DMGs and DM-lncRNAs were then extracted. The overall correlation between the methylation levels and expression levels of the DMGs/DM-lncRNAs were analyzed among the matched samples, and a significant negative correlation (PCC=-0.3857; P=0.574x10⁻¹⁴) was identified (Fig. 2). A total of 181 DMGs and DM-IncRNAs with negative PCCs, including 25 DM-IncRNAs and 156 DMGs, were identified in the training dataset (Fig. 3; Table I).

Construction of the RS system. Among the 181 DMGs and DM-IncRNAs, 105 were identified to be associated with prognosis, including 91 DMGs and 14 DM-IncRNAs, by univariate Cox regression analysis. Then, 16 DMGs and 6 DM-IncRNAs independently associated with prognosis were identified from the 105 DMGs/DM-IncRNAs using multivariate Cox regression analysis. According to the LASSO Cox regression model, 10 optimal DMGs, including aldehyde dehydrogenase 7 family member A1 (ALDH7A1), chromosome 8 open reading frame 48 (C8orf48), cytokine-like 1 (CYTL1), heat shock protein 90 alpha family class A member 1 (HSP90AA1), isovaleryl-CoA dehydrogenase (IVD), phosphodiesterase 3A (PDE3A), PNMA family member 2 (PNMA2), SAM domain, SH3 domain and nuclear localization signals 1 (SAMSN1), thyroid hormone receptor interactor 13 (TRIP13) and zinc finger protein 878 (ZNF878), and 6 optimal DM-lncRNAs, including ATXN8 opposite-strand lncRNA (ATXN8OS), brain cytoplasmic RNA 1 (BCYRN1), family with sequence similarity 138 member D (FAM138D), HLA complex group 11 (HCG11), MIR155 host gene (MIR155HG) and testis-specific transcript, Y-linked 13 (TTTY13) were finally identified in the training dataset (Table I).

The cutoff values of the methylation levels of the IncRNAs and genes in the aforementioned optimal combinations were calculated (Table I). In combination with the regression coefficients of the optimal genes or IncRNAs, RS systems based on the methylation status of optimal genes or IncRNAs were constructed. The formulas were as follows:
Gene status RS = -1.1070 x Status_{ALDH1A1} -0.5721 x Status_{C8orf48} + 0.4962 x Status_{CytL1} + 1.0118 x Status_{HSP90AA1} + 0.5696 x Status_{IVD} + 0.8992 x Status_{pDE3A} + 0.0109 x Status_{PNMA2} + 0.1803 x Status_{SAMSN1} + 0.1612 x Status_{TRIP13} - 0.1688 x Status_{ZNF878}; and lncRNA status RS = -1.2274 x Status_{ATXN8OS} + 13.1984 x Status_{BCYRN1} + 2.6252 x Status_{FAM138d} + 2.9526 x Status_{HG11} + 1.2577 x Status_{MIR155HG} + 1.8308 x Status_{TTTY13}.

In addition, RS systems based on the methylation levels of optimal genes or lncRNAs were built. The formulas for calculating the RSs were:

**Gene methylation (methy) RS = -1.1070 x Methy_{ALDH1A1} -0.5721 x Methy_{C8orf48} + 0.4962 x Methy_{CytL1} + 1.0118 x Methy_{HSP90AA1} + 0.5696 x Methy_{IVD} + 0.8992 x Methy_{pDE3A} + 0.0109 x Methy_{PNMA2} + 0.1803 x Methy_{SAMSN1} + 0.1612 x Methy_{TRIP13} - 0.1688 x Methy_{ZNF878}; and lncRNA methy RS = -1.2274 x Methy_{ATXN8OS} + 13.1984 x Methy_{BCYRN1} + 2.6252 x Methy_{FAM138d} + 2.9526 x Methy_{HG11} + 1.2577 x Methy_{MIR155HG} + 1.8308 x Methy_{TTTY13}.**

According to the median RS values, samples in the training set, validation set 1 and validation set 2 were separately subdivided into high- and low-risk groups. The associations between prognosis and the results of RS systems were analyzed. The KM curves for the methylation status-based RS systems and those for the methylation level-based RS systems are presented in Figs. 4 and 5, respectively. These RS systems demonstrated high levels of specificity, with the area under the receiver operating characteristic curve >0.6 (Figs. 4 and 5), indicating accurate levels of risk prediction. Of all the systems generated, the RS system based on the methylation level of the 10 optimal DEGs exhibited the best predictive performance (Table II).

**Construction of the nomogram survival model.** Independent clinical factors associating with prognosis in the training set were identified by means of univariate and multivariate Cox regression analyses (Table III). For both the training and validation sets, pathological T stage, radiotherapy and DNA methylation level-based model RS status were selected as independent prognostic factors. The KM curves demonstrating the correlations between pathological T-/radiotherapy and prognosis are presented in Fig. 6. The patients with low pathological T staging and those who underwent radiotherapy exhibited improved prognoses, which is consistent with the outcomes observed in clinical practice.

The 5-year nomogram survival model was built to further analyze the correlation between the 3 independent prognostic factors and prognosis; this model integrated various clinical indicators by means of the ‘points’ axis in the first row to predict the survival of patients corresponding to the samples (Fig. 7A). In addition, the predicted 5-year survival probability was compared to the actual 5-year survival probability, and a high level of consistency was observed (Fig. 7B).
Table I. Genes and lncRNAs involved in the optimal combinations.

| Type | Gene    | Locus          | Coefficient | Hazard ratio | 95% CI    | P-value | Cutoff |
|------|---------|----------------|-------------|--------------|-----------|---------|--------|
| gene | ALDH7A1 | cg26327732     | -1.107      | 0.415        | 0.250-0.884 | 2.350x10^-5 | 0.34   |
|      | C8orf48 | cg24727311     | -0.572      | 0.191        | 0.180-0.457 | 4.610x10^-4 | 0.34   |
|      | CYTL1   | cg17563034     | 0.496       | 1.509        | 1.104-2.323 | 3.051x10^-3 | 0.16   |
|      | HSP90AA1| cg23904247     | 1.012       | 4.769        | 2.041-6.111 | 1.080x10^-4 | 0.53   |
|      | IVD     | cg27529930     | 0.570       | 1.925        | 1.244-3.191 | 3.500x10^-3 | 0.11   |
|      | PDE3A   | cg26571814     | 0.899       | 1.977        | 1.745-3.979 | 9.310x10^-3 | 0.34   |
|      | PNMA2   | cg26268277     | 0.011       | 1.123        | 1.043-2.705 | 7.295x10^-3 | 0.48   |
|      | SAMSNI  | cg13951664     | 0.180       | 1.208        | 1.079-2.546 | 1.490x10^-4 | 0.25   |
|      | TRIP13  | cg17510385     | 0.161       | 1.33         | 1.038-2.691 | 9.838x10^-3 | 0.65   |
|      | ZNF878  | cg26626525     | -0.169      | 0.597        | 0.336-0.906 | 1.630x10^-3 | 0.06   |
| lncRNA | ATXN8OS| cg25514273     | -1.227      | 0.312        | 0.088-0.704 | 7.080x10^-3 | 0.24   |
|      | BCYRN1  | ch.X.1084981F  | 13.198      | 5.399        | 1.028-10.38 | 1.229x10^-3 | 0.04   |
|      | FAM138D | cg26523196     | 2.625       | 2.854        | 1.681-4.425 | 9.330x10^-3 | 0.73   |
|      | HCG11   | cg27490387     | 2.953       | 3.164        | 1.550-6.734 | 2.010x10^-3 | 0.20   |
|      | MIR155HG| cg23433889     | 1.258       | 1.431        | 1.281-3.186 | 3.341x10^-2 | 0.04   |
|      | TTTY13  | cg25918849     | 1.831       | 1.595        | 1.492-3.413 | 6.200x10^-3 | 0.24   |

lncRNA, long non‑coding RNA; CI, confidence interval.

Figure 3. Graphs for selection of the ‘lambda’ parameter, and coefficient distribution diagrams of the optimal genes or lncRNAs. (A) The parameter selection diagram for lncRNAs. In the graph on the left, the horizontal and vertical axes represent ‘lambda’ and cross-validation likelihood (cvl) values, respectively; the intersection of the black dashed lines represents the ‘lambda’ value at the largest cvl value. The graph on the right demonstrates the coefficient distribution diagram of the optimal lncRNA. (B) The parameter selection diagram for genes. In the graph on the left, the horizontal and vertical axes respectively represent ‘lambda’ and cvl values; the intersection of the black dashed lines represents the ‘lambda’ value at the largest cvl value. Right: The graph on the right demonstrates the coefficient distribution diagram of the optimal gene. lncRNAs, long non‑coding RNAs.
Figure 4. KM curves and ROC curves for the methylation status-based RS systems. (A) The KM curves (black and red curves respectively represent low- and high-risk groups) and the ROC curve (black, red and green lines represent the training set, validation set 1 and validation set 2, respectively) for the IncRNA methylation status-based RS system. (B) The KM curves (blue and purple curves respectively represent low- and high-risk groups) and the ROC curve (black, red and green lines represent the training set, validation set 1 and validation set 2, respectively) for the DNA methylation status-based RS system. KM, Kaplan-Meier; ROC, receiver operating characteristic; RS, risk score; TCGA, The Cancer Genome Atlas; HNSC, head and neck squamous cell carcinoma; AUC, area under the receiver operating characteristic curve; IncRNA, long non-coding RNA; HR, hazard ratio.
Figure 5. KM curves and ROC curves for the methylation level-based RS systems. (A) The KM curves (black and red curves represent low- and high-risk groups, respectively) and the ROC curve (black, red and green lines represent the training set, validation set 1 and validation set 2, respectively) for the long noncoding RNA (lncRNA) methylation level-based RS system. (B) The KM curves (blue and purple curves represent low- and high-risk groups, respectively) and the ROC curve (black, green and red lines represent the training set, validation set 1 and validation set 2, respectively) for the DNA methylation level-based RS system. KM, Kaplan-Meier; ROC, receiver operating characteristic; RS, risk score; TCGA, The Cancer Genome Atlas; HNSC, head and neck squamous cell carcinoma; AUc, area under the receiver operating characteristic curve; lncRNA, long non-coding RNA; HR, hazard ratio.
Table II. Survival analysis of the methylation status-based RS system.

A. KM survival analysis for the methylation status-based RS system

| Gene product type | Status model (log-rank P-value) | Methylation model (log-rank P-value) |
|-------------------|---------------------------------|------------------------------------|
|                   | Training HR (95% CI) | Validation 1 (GSE39279) HR (95% CI) | Validation 2 (TCGA-HNSc) HR (95% CI) | Gene product type | Training HR (95% CI) | Validation 1 (GSE39279) HR (95% CI) | Validation 2 (TCGA-HNSc) HR (95% CI) |
| IncRNA            | 2.628x10^-4 | 1.446x10^-1 | 1.385x10^-1 | 1.397 (0.896-2.178) |
| DNA               | 2.004x10^-12 | 6.112 (3.452-10.82) | 2.213 (1.050-4.664) | 2.742x10^-2 | 1.638 (1.052-2.550) |

B. KM survival analysis for the methylation level-based RS system

| Gene product type | Status model (ROC) | Methylation model (ROC) |
|-------------------|--------------------|-------------------------|
|                   | Training | Validation 1 (GSE39279) | Validation 2 (TCGA-HNSc) | Training | Validation 1 (GSE39279) | Validation 2 (TCGA-HNSc) |
| IncRNA            | 0.963 | 0.908 | 0.967 |
| DNA               | 0.995 | 0.961 | 0.912 |

KM, Kaplan-Meier; RS, risk score; ROC, receiver operating characteristic; AUC, area under the ROC curve; IncRNA, long non-coding RNA; HR, Hazard Ratio; CI, confidence interval; TCGA, The Cancer Genome Atlas; HNSc, head and neck squamous cell carcinoma.
Pathway enrichment analysis. As aforementioned, following the addition of the RSs of the DNA methylation level-based RS system to the analysis, the samples in the training set were classified into either high- or low-risk groups. A total of 820 DEGs between the two groups were identified, including 412 upregulated genes and 408 downregulated genes. Following pathway enrichment analysis, 3 pathways were identified to be enriched in this set of DEGs: The calcium signaling pathway (P=0.040944); complement and coagulation cascades (P=0.0034123); and vascular smooth muscle contraction (P=0.0427049; Table IV).

Discussion

In the present study, a total of 335 DMGs and DM-IncRNAs were identified between the recurrence-associated samples and nonrecurrence samples of tumor tissues from patients with LSCC. Among them, 16 DMGs and 6 DM-IncRNAs were significantly associated with independent prognosis, and 10 optimal DMGs (ALDH7A1, C8orf48, CYTL1, HSP90AA1, IVD, PDE3A, PNMA2, SAMSNI, TRIP13 and ZNF878) and 6 optimal DM-IncRNAs (ATXN8OS, BCYRN1, FAM138D, HCG11, MIR155HG and TTTY13) were used to construct the methylation status-based or methylation level-based RS systems. In the KM analysis, the 10-DMG methylation level-based RS system exhibited the best performance. The set of DEGs between high- and low-risk groups according to the RS system was identified to be enriched in 3 pathways: The calcium signaling pathway; complement and coagulation cascades; and vascular smooth muscle contraction. As the RS system based on the 10 DMGs exhibited an improved predictive performance compared with that based on the 6 IncRNAs, the present study focused on the functions of these 10 DMGs in LSCC.

Elevated expression of ALDH7A1 in prostate cancer has been previously described, and was suggested to predict disease progression and metastasis (32), suggesting that alteration of the expression of this gene may control tumor progression. However, to the best of our knowledge, studies investigating the involvement of this gene in lung cancer are scarce. Low expression of ALDH7A1 in tumors of surgically treated patients with NSCLC is associated with a decreased incidence of tumor recurrence, indicating that decreased expression of this gene may predict a good prognosis for these patients (33). Patients with LSCC with idiopathic pulmonary fibrosis indicate a significantly decreased methylation level of ALDH7A1 compared with those without fibrosis (34). This observation suggests that methylation status of ALDH7A1 may affect other clinical factors. In the present study, ALDH7A1 was a component of the 10-DMG methylation level-based RS system that exhibited optimal performance; therefore, the methylation level of this gene may predict LSCC recurrence. Nevertheless, as the data concerning this gene in lung cancer are limited, the specific mechanisms of action of the ALDH7A1 protein and ALDH7A1 DNA methylation in LSCC should be studied further.

CYTL1 has important roles in certain types of cancer and is regulated by DNA methylation in LSCC (35). The methylation patterns of CYTL1 are evidently different between early and late stages of LSCC, and hypermethylation is more common in the advanced stages (36). Although CYTL1 hypermethylation...
Figure 6. KM curves. (A) The KM curve revealing the correlation between pathological T and prognosis (black, red, blue and purple lines represent pathological T1, pathological T2, pathological T3 and pathological T4 stage, respectively); (B) The KM curve demonstrating the correlation between radiotherapy and prognosis (black and red lines represent patients without radiotherapy and those with radiotherapy, respectively). KM, Kaplan-Meier.

Figure 7. Nomogram survival model. (A) The nomogram survival model based on the independent prognostic factors. (B) The line graph indicating the consistency between the predicted 5-year survival probability and actual 5-year survival probability. The horizontal and vertical axes represent the predicted probability of RFS and the actual probability of RFS, respectively. The black line represents the trend line connecting the start time point and the 5-year time point. RFS, recurrence-free survival.
does not affect the repression activity mediated by histone deacetylases (36), this repression activity may be subject to the regulation between CYTL1 hypermethylation and histone deacetylases, as the DNA methylation level is considered to affect the binding of a histone deacetylase to a promoter region (36,37). CYTL1 is suggested to be a risk indicator of smoking-associated impairment of metabolic health, as it is hypomethylated and upregulated in non-tumor tissues of smokers (38). According to the present study, the CYTL1 gene was often methylated in recurring LScc tumors, suggesting that it may be a predictive factor of LScc recurrence.

High expression levels of HSP90AA1, heat shock protein 90 alpha family class B member 1 and heat shock protein 90 beta family member 1 are associated with adverse outcomes among patients with NSCLC, and therefore may serve as promising prognostic markers and therapeutic targets in NSCLC (39). HSP90AA1 is differentially expressed between LUAd and LScc, and therefore its expression profile may be used to distinguish the two subtypes (40). Nevertheless, data concerning HSP90AA1 methylation or the effect of this gene on tumor recurrence are scarce. PDE3A expression is low in chemoresistant NSCLC cells due to DNA hypermethylation, and high PDE3A expression is associated with improved survival in patients with LUAd (41). According to the results of the present study, methylated PDE3A may be associated with the prognosis of LScc.

PNMA2 is aberrantly expressed in various types of tumor in patients with paraneoplastic syndromes; therefore, PNMA2 may be implicated in tumorigenesis (42). PNMA2 has been demonstrated to be a tissue marker of small intestine neuroendocrine tumors, and Ma2 autoantibodies in the blood are a valuable biomarker for the diagnosis and prediction of tumor recurrence (43). Conversely, to the best of our knowledge, no studies on the methylation of this gene have been published, and this parameter may be a new prognostic indicator in LScc, based on the results of the present study.

Downregulation of SAMSN1 is detectable in lung cancer cell lines and may be involved in the development of this disease (44). Elevated TRIP13 expression contributes to the progression of LUAd and may be a candidate biomarker or therapeutic target in LUAd (45). In lung cancer in Xuanwei, TRIP13, CAM-responsive element-binding protein 3-like 4 and cyclin E2 exhibit concordant upregulation and frequent copy number gains, and have been proposed as potential oncogenes in the pathogenesis of lung cancer in this region (46). In addition, silencing of TRIP13 can suppress cell growth and metastasis of hepatocellular carcinoma by activating TGF-β1/SMAD3 signaling (47). TRIP13 is 1 of the 7 hypomethylated genes in kidney renal cell carcinoma that have been suggested as a prognostic factor (48). Downregulation of ZNF878 has been suggested to be correlated with a poor prognosis of patients with LUAd (49). Although published data on SAMSN1, TRIP13, and ZNF878 methylation levels or their effects on LScc are limited, when taken into consideration with the results of the present study, we hypothesize that these 3 genes may perform important functions in LScc progression, and their methylation levels may help to determine an accurate prognosis of the disease.

Information on the participation of genes C8orf48 and IVD in LScc is scarce. According to the results of the present study, their methylation may be a prognostic indicator of LScc recurrence.

Despite the identification of the 10-DMG signature that may predict the recurrence of LScc, the present study contained certain limitations, such as the small sample size of the data retrieved from the database, and the lack of expression validation experiments. In addition, the clinical information of the LScc samples in the training set was not compared to the validation sets, due to the difference in platforms. Nevertheless, this did preclude the predictive performance evaluation on risks of the identified clinical factors, as they could be distinguished from the AUC values. Further experiments are required to confirm the results.

In conclusion, 335 DMGs and DM-IncRNAs, including 27 DM-IncRNAs and 308 DMGs, were identified between recurrence-associated and nonrecurrence samples. The RS system based on the methylation levels of the 10 optimal DMGs (ALDH7A1, C8orf48, CYTL1, HSP90AA1, IVD, PDE3A, PNMA2, SAMSN1, TRIP13 and ZNF878) may help to predict the outcomes of patients with LScc.

Table IV. Pathways enriched for the differentially expressed genes.

| Term                                      | Size | ES     | NES    | P-value  | Genes                        |
|-------------------------------------------|------|--------|--------|----------|------------------------------|
| hsa04020: Calcium signaling pathway       | 10   | -0.396503 | -1.114227 | 4.094x10^-2 | EDNRB, HRH1, CALML3, PLN, LTB4R2, RYR1, BDKRB2, CALML5, ITPR2, F2R |
| hsa04610: Complement and coagulation cascades | 10   | 0.5863498 | 2.0125692 | 3.412x10^-3 | C3AR1, F12, FGG, C5, F8, SERPINA1, C4BPA, C2, BDKRB2, F2R |
| hsa04270: Vascular smooth muscle contraction | 7    | -0.537324 | -1.341176 | 4.270x10^-2 | PTGIR, CALML3, CALML5, PRKG1, PLA2G3, PPP1R14A, ITPR2 |

ES, enrichment score; NES, normalized enrichment score.

Acknowledgements

The results shown here are in whole or part based upon data generated by The Cancer Genome Atlas Research Network: https://www.cancer.gov/tcga.

Funding

The present study was supported by Suzhou Science and Technology Development Project (grant no. SYS2018085) and...
the Construction Project of Suzhou Clinical Medical Center (grant no. Szzx201506).

Availability of data and materials
All the data used in the present study are available from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/; accession no., GSE39279) and The Cancer Genome Atlas public databases.

Authors' contributions
MZ, LS, YR, SZ, JM, PG and JL were responsible for data collection and statistical analysis. MZ, LS and YR participated in drafting and editing the manuscript. FG and BL were responsible for designing the work and reviewing the manuscript. BL obtained the funding. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Patient consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

References
1. McGuire S: World cancer report 2014. Geneva, Switzerland: World health organization, international agency for research on cancer, WHO press, 2015. Adv Nutr 7: 418–419, 2016.
2. Hirsch FR and Bunn PA Jr: Progress in research on screening for designing the work and reviewing the manuscript. BL obtained the funding. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Patient consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

References
1. McGuire S: World cancer report 2014. Geneva, Switzerland: World health organization, international agency for research on cancer, WHO press, 2015. Adv Nutr 7: 418–419, 2016.
2. Hirsch FR and Bunn PA Jr: Progress in research on screening for designing the work and reviewing the manuscript. BL obtained the funding. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Patient consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

References
1. McGuire S: World cancer report 2014. Geneva, Switzerland: World health organization, international agency for research on cancer, WHO press, 2015. Adv Nutr 7: 418–419, 2016.
35. Zhu S, Kuek V, Bennett S, Xu H, Rosen V and Xu J: Protein cyclin1: Its role in chondrogenesis, cartilage homeostasis, and disease. Cell Mol Life Sci 76: 3515-3523, 2019.
36. Kwon YJ, Lee SJ, Koh JS, Kim SH, Lee HW, Kang MC, Bae JB, Kim YJ and Park JH: Genome-wide analysis of DNA methylation and the gene expression change in lung cancer. J Thorac Oncol 7: 20-33, 2012.
37. Kim SN, Kim NH, Lee W, Seo DW and Kim YK: Histone deacetylase inhibitor induction of P-glycoprotein transcription requires both histone deacetylase 1 dissociation and recruitment of CAAT/enhancer binding protein beta and pCAF to the promoter region. Mol Cancer Res 7: 735-744, 2009.
38. Tsai PC, Glastonbury CA, Eliot MN, Bollepalli S, Yet I, Castillo-Fernandez JE, Carnero-Montoro E, Hardiman T, Martin TC, Vickers A, et al: Smoking induces coordinated DNA methylation and gene expression changes in adipose tissue with consequences for metabolic health. Clin Epigenetics 10: 126, 2018.
39. Liu K, Kang M, Li J, Qin W and Wang R: Prognostic value of the mRNA expression of members of the HSP90 family in non-small cell lung cancer. Exp Ther Med 17: 2657-2665, 2019.
40. Lu C, Chen H, Shan Z and Yang L: Identification of differentially expressed genes between lung adenocarcinoma and lung squamous cell carcinoma by gene expression profiling. Mol Med Rep 14: 1483-1490, 2016.
41. Tian FM, Zhong CY, Wang XN and Meng Y: PDE3A is hyper-methylated in cisplatin resistant non-small cell lung cancer cells and is a modulator of chemotherapy response. Eur Rev Med Pharmacol Sci 21: 2635-2641, 2017.
42. Lee YH, Pang SW and Tan KO: PNMA2 mediates heterodimeric interactions and antagonizes chemosensitizing activities mediated by members of PNMA family. Biochem Biophys Res Commun 473: 224-229, 2016.
43. Cui T, Hurtig M, Elgue G, Li SC, Veronesi G, Essaghir A, Demoulin JB, Pelosi G, Alimohammadi M, Oberk K and Giandomenico V: Paraneoplastic antigen Ma2 autoantibodies as specific blood biomarkers for detection of early recurrence of small intestine neuroendocrine tumors. PLoS One 5: e16010, 2010.
44. Yamada H, Yanagisawa K, Tokumaru S, Taguchi A, Nimura Y, Osada H, Nagino M and Takahashi T: Detailed characterization of a homozygously deleted region corresponding to a candidate tumor suppressor locus at 21q11-21 in human lung cancer. Genes Chromosomes Cancer 47: 810-818, 2008.
45. Li W, Zhang G, Li X, Wang X, Li Q, Hong L, Shen Y, Zhao C, Gong X, Chen Y and Zhou J: Thyroid hormone receptor interactor 13 (TRIP13) overexpression associated with tumor progression and poor prognosis in lung adenocarcinoma. Biochem Biophys Res Commun 499: 416-424, 2018.
46. Zhang Y, Xue Q, Pan G, Meng QH, Tuo X, Cai X, Chen Z, Li Y, Huang T, Duan X and Duan Y: Integrated analysis of genome-wide copy number alterations and gene expression profiling of lung cancer in Xuanwei, China. PLoS One 12: e0169098, 2017.
47. Yao J, Zhang X, Li J, Zhao D, Gao B, Zhou H, Gao S and Zhang L: Silencing TRIP13 inhibits cell growth and metastasis of hepatocellular carcinoma by activating of TGF-β1/smad3. Cancer Cell Int 18: 208, 2018.
48. Hu F, Zeng W and Liu X: A gene signature of survival prediction for kidney renal cell carcinoma by multi-omic data analysis. Int J Mol Sci 20: pii: E5720, 2019.
49. Liu H and Zhao H: Prognosis related miRNAs, DNA methylation, and epigenetic interactions in lung adenocarcinoma. Neoplasma 66: 487-493, 2019.