frontiers in Genetics

ORIGINAL RESEARCH
published: 24 April 2019
doi: 10.3389/fgene.2019.00349

Prognostic and Predictive Value of Three DNA Methylation Signatures in Lung Adenocarcinoma

Yanfang Wang [1†], *Haowen Deng* [2†], *Shan Xin* [1,3], *Kai Zhang* [4], *Run Shi* [1*] *and Xuanwen Bao* [5,6*]

[1] *Ludwig-Maximilians-Universität München, Munich, Germany,* [2] *Chair for Computer Aided Medical Procedures and Augmented Reality, Technical University Munich, Munich, Germany,* [3] *Institute of Molecular Toxicology and Pharmacology, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany,* [4] *Department of Cardiology, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China,* [5] *Institute of Radiation Biology, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany,* [6] *Technical University Munich (TUM), Munich, Germany*

OPEN ACCESS

Edited by:
Jiucun Wang,
Fudan University, China

Reviewed by:
Mariana Brait,
Johns Hopkins University,
United States
David D. Eisenstat,
University of Alberta, Canada
Nejat Dalay,
Istanbul University, Turkey
Yi Wang,
Fudan University, China

Correspondence:
Run Shi
shirun@outlook.com
Xuanwen Bao
xuanwen.bao@helmholtz-muenchen.de

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Epigenomics and Epigenetics, a section of the journal Frontiers in Genetics

Received: *20 November 2018*
Accepted: *01 April 2019*
Published: *24 April 2019*

Citation:
Wang Y, Deng H, Xin S, Zhang K, Shi R and Bao X (2019) Prognostic and Predictive Value of Three DNA Methylation Signatures in Lung Adenocarcinoma. Front. Genet. 10:349. doi: 10.3389/fgene.2019.00349

Background: Lung adenocarcinoma (LUAD) is the leading cause of cancer-related mortality worldwide. Molecular characterization-based methods hold great promise for improving the diagnostic accuracy and for predicting treatment response. The DNA methylation patterns of LUAD display a great potential as a specific biomarker that will complement invasive biopsy, thus improving early detection.

Method: In this study, based on the whole-genome methylation datasets from The Cancer Genome Atlas (TCGA) and several machine learning methods, we evaluated the possibility of DNA methylation signatures for identifying lymph node metastasis of LUAD, differentiating between tumor tissue and normal tissue, and predicting the overall survival (OS) of LUAD patients. Using the regularized logistic regression, we built a classifier based on the 3616 CpG sites to identify the lymph node metastasis of LUAD. Furthermore, a classifier based on 14 CpG sites was established to differentiate between tumor and normal tissues. Using the Least Absolute Shrinkage and Selection Operator (LASSO) Cox regression, we built a 16-CpG-based model to predict the OS of LUAD patients.

Results: With the aid of 3616-CpG-based classifier, we were able to identify the lymph node metastatic status of patients directly by the methylation signature from the primary tumor tissues. The 14-CpG-based classifier could differentiate between tumor and normal tissues. The area under the receiver operating characteristic (ROC) curve (AUC) for both classifiers achieved values close to 1, demonstrating the robust classifier effect. The 16-CpG-based model showed independent prognostic value in LUAD patients.

Interpretation: These findings will not only facilitate future treatment decisions based on the DNA methylation signatures but also enable additional investigations into the utilization of LUAD DNA methylation pattern by different machine learning methods.

Keywords: LUAD, DNA methylation, regularized logistic regression, recursive feature elimination, LASSO Cox regression, metastasis

INTRODUCTION

Lung cancer is the leading cause of cancer-related mortality globally, causing over a million deaths a year (Genome Atlas Research Network., 2014; Jemal et al., 2018). There are two clinical types, one is the aggressive subtype small cell lung cancer and the other is non-small cell lung cancer (Hankey et al., 1999). Non-small cell lung cancer is histologically classified into four major subtypes by pathological and molecular characteristics: adenocarcinoma, large cell lung cancer, squamous cell lung cancer, and other types (Ettinger et al., 2010). Adenocarcinoma is the most common histological subtype of non-small cell lung cancer. Tobacco smoking is the major cause of lung adenocarcinoma (Toh et al., 2006). However, with the decrease in the number of smokers in many countries, the occurrence of LUAD in non-smokers has increased (Genome Atlas Research Network., 2014).

An accurate diagnosis of LUAD is one precondition to achieve a better treatment effect. Although the Mayo Clinic stage, size, grade, and necrosis (SSIGN) score, as well as the University of California Integrated Staging System can help improve the accuracy of the prognosis (Travis et al., 2011), the outcomes of patients with similar clinical characteristics or integrated systems scores still differ. Molecular characteristics may provide an indication for predicting the LUAD prognosis and response to therapy, thus offering great potential for improving individual treatment. Moreover, molecular characterization-based methods do not generally require bulk tissue samples, which can improve the patients' tolerance and reduce unnecessary operation steps. Among all the molecular characteristics, DNA methylation of CpG sites plays a crucial role in epigenetic regulation by reducing the activity of a DNA segment and repressing gene transcription (Jones, 2012; Du et al., 2015; Schübeler, 2015). DNA methylation is associated with carcinogenesis by repressing the expression of the tumor suppressor gene and promoting the expression of oncogenes (Herman et al., 1995; Schübeler, 2015; Vizoso et al., 2015; Klutstein et al., 2016). Hence, the cancer tissues have a distinct DNA methylation pattern compared to normal tissues. More importantly, unlike somatic genetic mutations in tumor tissues, DNA methylation patterns are inherently reversible changes and can therefore be promising targets for drug treatments (Ramchandani et al., 1999). Using DNA methylation signatures can help us make a better prognosis and predict the treatment response, thus prolonging the patients' survival.

Machine learning is a novel method to learn concept from data, which will help researchers discover the hidden insights. Based on DNA methylation patterns, machine learning techniques are developed and used to design models for precise classification and accurate prediction in medicine. In this study, we evaluated the possibility of DNA methylation signatures in identifying LUAD lymph node metastasis, differentiating between tumor tissue and normal tissue and predicting the OS of LUAD patients by applying TCGA whole-genome methylation datasets to several machine learning methods. Our results showed robust classifier effects with the AUC of both classifiers achieving values close to one for identifying lymph node metastasis and differentiating between tumor tissue and normal tissue. Cross-validation was applied to prevent overfitting. The LASSO Cox regression model was used to evaluate the patients' OS. Risk scores from the LASSO Cox model were combined with other clinicopathological risk factors to generate a nomogram to predict the prognosis and help the doctors to manage LUAD patients.

METHODS

Data Source

The DNA methylation files and patients' information were obtained from Xena (https://xenabrowser.net/). Complete clinical, molecular, and histopathological data-sets are available at the TCGA website (https://portal.gdc.cancer.gov/).

Feature Selection for DNA CpG Sites

We formulate critical methylation identification as a feature selection problem. Each CpG site is treated as a feature here and our goal is to find out which features are important for different tasks.

Variance Based Filtering

Variance is the squared deviation of the data from its mean, showing the spread of numbers. It is an important characteristic that reflects the distribution and discriminability of a feature. The variance σ of an observed sample sequence of a given feature $\{x_1, x_2, \ldots, x_i, \ldots, x_N\}$ is computed by averaging across the squared difference of each value to the mean μ.

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

$$\mu = \frac{\sum x_i}{N}$$

In general, a larger variance σ means a wider distributed and more separable feature space, which facilitates training a classifier to find class boundaries. On the other hand, variance σ is positively correlated with information entropy E, meaning that more information could be obtained with a larger variance σ. When σ is small, all the data is compressed and provides insufficient information for a classifier, so that we would avoid features with a small variance by setting a minimum threshold to filter out the indiscriminate features.

$$E = -\int P(x) \log P(x) dx$$

Regularized Logistic Regression Model

Logistic regression is a widely applied and useful statistical, non-linear model for predicting a binarized outcome based on a sequence of independent features. Assuming we have a general linear regression model y, which satisfies

$$y = \sum_{i=0}^{N} \beta_i x_i$$

where x_i stands for the i-th feature and β_i is the correspondent coefficient. Since there is no constraint on the range of β_i and x_i, there is no maximum or minimum limit for y, i.e., $y \in [-\infty, \infty]$. Consider a standard logistic function

$$f(x) = \frac{1}{1 + e^{-x}}$$

This could map an input space from an infinite $[-\infty, \infty]$ to a finite $[0, 1]$. By combining the linear regression model with the logistic function, we obtain the logistic regression model

$$y = \frac{1}{1 + e^{-\sum_{i=1}^{N} \beta_i x_i}}$$

By thresholding y with threshold t, we obtain the binarized output.

$$o = \begin{cases} 1, y \geq t \\ 0, y < t \end{cases}$$

Note that, regularization term can be compounded with a logistic regression model, to force the learned coefficients to be sparser and more resistant to overfitting, which is highly beneficial for feature selection as well. We term the logistic regression model with regularization term as "Regularized logistic regression model."

Recursive Feature Elimination

Recursive feature elimination (RFE) adopts a brute-force and recursive way of undermining important features. Given a predefined model, which weighs all the features internally, RFE recursively uses the set of features to train the model and discard features that are the least important for the model (e.g., small weights) and repeats the training with the remaining features. This operation keeps recycling until certain expectations are reached, such as the maximum number of expected features N_{exp}. The process is described in **Algorithm 1**.

Algorithm 1: Recursive Feature Elimination.

```
INPUT: a set of features $S = \{f_1, f_2, \ldots, f_i, \ldots, f_n\}$, expected feature number $N_{\text{exp}}$
OUTPUT: a set of kept features $S_{best} = \{f_{s_1}, f_{s_2}, \ldots, f_{s_i}, \ldots, f_{s_n}\}$
WHILE size $(S) > N_{\text{exp}}$ DO
    Train a model with the set of features in $S$
    Get the coefficients for the features learned from the classification model
    Prune features with small coefficients $S_{non-important}$
    Update feature set $S = S - S_{non-important}$
END WHILE
Keep the final set of features as the set of most important features $S_{best} = S$
```

Cox Regression

Cox regression, also called Proportional Hazards Regression, is a survival analysis model. It can be used to analyze relationships between different features and the survival time. The Cox model is based on the proportional hazards condition, which assumes that features have a proportional relationship to the exponential change of hazard. Thus, the model is formulated as a multiplication of a baseline hazard function with a sole time variable t, and an exponential function of the linear combination of all of the features as an input. Given a set of n samples $\{(\boldsymbol{X}_i, Y_i, s_i) \mid 0 \leq i \leq n,\ i \in \boldsymbol{R}\}$, where $X_i = (x_{i0}, x_{i1}, \ldots, x_{ik})$ and stands for the i-th sample of all the k features, Y_i is the observation time and s_i is the survival status, the hazard function is

$$H_i(t) = H_0(t)\, e^{\boldsymbol{X}_i^T \beta}$$

$\boldsymbol{\beta} = (\beta_{\boldsymbol{0}}, \beta_1, \ldots, \beta_k)$ is the coefficient vector weighing the contribution of the features. The partial likelihood of all the samples is

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} L_i(\boldsymbol{\beta}) \\ &= \prod_{i=1}^{n} \frac{H_i(Y_i \mid \boldsymbol{X}_i)}{\sum_{j:Y_j \geq Y_i} H_i\left(Y_i \middle| \boldsymbol{X}_j\right)} \\ &= \prod_{i=1}^{n} \frac{e^{\boldsymbol{X}_i^T \boldsymbol{\beta}}}{\sum_{j:Y_j \geq Y_i} e^{\boldsymbol{X}_j^T \boldsymbol{\beta}}} \end{aligned}$$

By penalizing -log $(L(\boldsymbol{\beta}))$, the optimal $\boldsymbol{\beta}$ could be uncovered.

LASSO Regularization

LASSO (Least Absolute Shrinkage and Selection Operator) is an important regularization in many regression analysis methods. The concept behind LASSO is that an L1-norm is used to penalize the weight of the model parameters. Assuming a model has a set of parameters $\{w_0, w_1, \ldots, w_n\}$, the LASSO regularization can be written as

$$\lambda \cdot \sum_{i=0}^{n} \|w_i\|_1$$

It can be also expressed as a constraint to the targeted objective function

$$\sum \left\|\boldsymbol{Y} - \boldsymbol{Y}^*\right\|_2,\ s.t.\ \|w_i\|_1 < t$$

An important property of the LASSO regularization term is that it can force the parameter values to be 0, thus generating a sparse parameter space, which is a desirable character for feature selection.

Workflow of the Coding Process

When it came to selecting the methylation features for the metastasis and tumor identification problems, we first used variance-based filtering to eliminate some of the least important CpG sites, and to decrease the computation for the following Regularized Logistic Regression Model and RFE. To avoid model overfitting and bias in the feature selection, cross validation was used in the following stages. The dataset was evenly divided into 5-folds, and further feature selections were conducted by applying Logistic Regression Model and RFE, following the standard pipeline of cross validation.

When predicting the OS of LUAD patients, we built the Cox proportional hazard regression model with LASSO regularization. 5-fold cross-validation was applied to avoid the overfitting. We plotted the plots in R software (R Foundation for Statistical Computing, Vienna, Austria. Version 3.4.3) and Python (Python Software Foundation. Python Language Reference, version 3.7).

RESULTS

Preparation of LUAD DNA Methylation Datasets

LUAD DNA methylation data and corresponding clinical data were downloaded from Xena (https://xenabrowser.net/) (Cline et al., 2013). After removing samples without a survival status and normalization, a total of 478 samples were analyzed in the present study (**Supplementary File 1**). The datasets included 409 samples for the recognition of metastasis, 428 samples for the recognition of tumor from normal tissue, and 446 samples for the prediction of OS [(**Supplementary Files 2**–(**4**].

Identification of 3616-CpG-Based Signature for the Recognition of Metastasis

Variance-based selection was applied to filter features (methylation CpG sites). Features with small variances tend to be less discriminative, so we filtered out features with a standard variance smaller than 0.01 and 135,094 methylation signatures were selected. Regularized logistic regression and cross-validation were then applied to weigh the importance of each feature. The 428 LUAD samples were randomly assigned to a test set or a validation set by the cross-validation method. In short, five rounds of cross-validation were performed using different partitions and the validation results were combined over five rounds to overcome overfitting. By varying the value of the coefficient threshold, we obtained a different number of features that could be kept. When we used those kept features to regress the linear Logistic model by 5-fold cross-validation, the mean accuracy trend was as follows (**Figure 1A**). The number of kept features with regard to the different values of coefficient thresholds was shown in **Figure 1B**. The best performance was achieved at the threshold value 0.05 with 6,198 features kept with a 5-fold cross-validation. Recursive feature elimination with the same cross-validation configuration was tested and the result indicated that the kept features were the optimum minimal set of all the features (**Figure 1C**). The value of kept methylation CpG sites was shown in **Figure 1D**. We assessed the accuracy of the 3616-CpG-based classifier for detecting metastasis with a ROC analysis (**Figure 1E**) with the same cross-validation configuration, and averaged the weights of selected features across different set as the final coefficients. Furthermore, the metastatic probability of each sample were calculated by the coefficients of kept methylation CpG sites (**Figure 1F** and **Supplementary File 5**). The AUC for the classifier achieved values close to 1in all of the 5-fold cross-validation, indicating the robust classifier effect. The tumor tissues in total dataset were divided into high metastatic risk score and low metastatic risk score groups, respectively, using 0.5 as the cutoff. The patients in the low metastatic risk score group have a longer OS than those in the high metastatic risk score group in the total datasets ($p < 0.0001$, **Figure 1G**) as well as in the separated 5-fold training and validation sets ($p < 0.05$, **Supplementary Figure 1**). We assessed the prognostic accuracy of the 3616-CpG-based classifier metastatic classifier with a time-dependent ROC analysis at varying follow-up times (500, 1,000, 1,500, 2,000, 2,500, 3,000 days) (**Supplementary Figure 2**). The accuracy was all around 66%, indicating that the 3616-CpG-based classifier for identifying metastasis could also work well for predicting the OS of LUAD patients.

Identification of 14-CpG-Based Signature to Recognize Tumor and Normal Tissues

134,015 features were kept by variance thresholding (0.01). Regularized Logistic regression and cross-validation were applied to weigh the importance of each feature as mentioned above. An accuracy of 100% can easily be achieved for the number of features range from 14 to 43,246 (**Figure 2A**). The number of kept features with regard to the different thresholding values was shown in **Figure 2B**. Recursive feature elimination with cross-validation was tested and the result indicated that an accuracy of 100% can be achieved when the kept feature numbers reached 14 (**Figure 2C**). 14 CpG sites were kept: cg25774643, cg03502002, cg14789818, cg23479922, cg04864807, cg07915921, cg20146541, cg08862830, cg01016533, cg19191888, cg08094098, cg01912692, cg10707110, cg24103195. The value of kept methylation CpG sites was shown in **Figure 2D**. We then calculated the probability of being tumor for each sample by the coefficients of kept methylation CpG sites (**Supplementary File 6** and **Figure 2E**) in the same tradition as of the recognition of metastasis. The accuracy of the 14-CpG-based classifier was assessed by means of ROC analysis (**Figure 2F**). The results showed that the accuracy reached 100% in all 5-fold cross-validation, indicating the high sensitivity and specificity of the 14-CpG-based classifier in differentiating between LUAD tumor tissues and corresponding normal tissues. Furthermore, we applied the 14-CpG-based classifier on an external dataset to confirm the accuracy of the 14-CgG-based classifier (**Figure 2G**). The AUC value was 98.4% for differentiating the tumor and normal tissues (**Figure 2H**). The analysis before showed the regularized logistic model we applied worked well in different datasets.

Identification of 16-CpG-Based Signature to Predict the OS of LUAD Patients

We used a LASSO Cox regression to build a prognostic model, which selected 16 methylation CpG sites from the CpG sites identified by the DNA methylation 450 k chip: cg00161124, cg01105229, cg03923535, cg10976778, cg12141052, cg12240358,

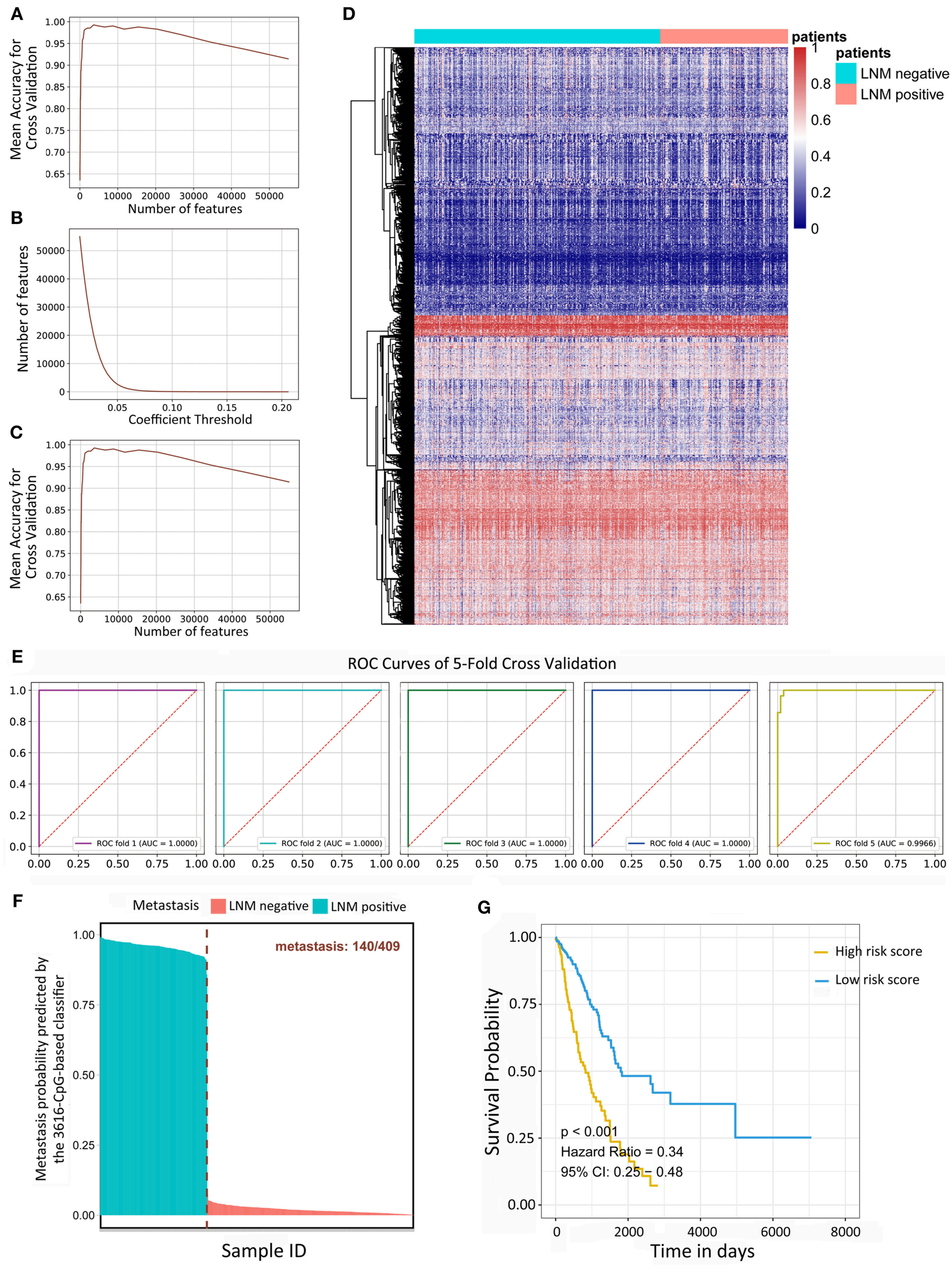

FIGURE 1 | Linear logistic classifier using a 5-fold cross-validation for the metastatic analysis and the efficiency for the 3616-CpG-based classifier. **(A)** The mean accuracy trend for the number of features and mean accuracy for the cross-validation. **(B)** The methylation features with regard to the different values of coefficient thresholds. **(C)** Recursive feature elimination with a cross-validation test. **(D)** Unsupervised hierarchical clustering and heat map associated with the methylation profile (according to the color scale shown) to recognize metastasis in LUAD. **(E)** ROC curves for 5-fold cross-validation showing the high sensitivity and specificity of the classifier in predicting metastasis. **(F)** The metastatic probability of each sample calculated by coefficients of the methylation signature. **(G)** Kaplan-Meier survival analysis for LUAD patients, which are divided into low-risk and high-risk groups using a cutoff value of 0.5. LNM: lymph node metastasis.

cg13297560, cg14139311, cg14184729, cg18140857, cg19410791, cg20268054, cg23146197, cg25229048, cg26709300, cg27018309 (**Figures 3A,B**). The values of the 16 methylation CpG sites for each patient were shown in **Figure 3B**. A formula was derived to calculate the risk score for every patient based on their individual 16 methylation β values (**Supplementary File 7**). The risk scores of tumor samples were calculated by the coefficients of the kept methylation CpG sites (**Figure 3C**). The patients were divided into high risk score and low risk score groups, respectively, with a cutoff of −0·54. Kaplan-Meier survival analysis (**Figure 3D**) showed that the survival probability of patients in lower risk score was significantly better than in high risk score group (log-rank test, all $p < 0.0001$). We assessed the prediction accuracy of the 16-CpG-based model by means of time-dependent ROC analysis at varying follow-up times. The AUC values for 500, 1,000, 1,500, 2,000, 2,500, and 3,000 days were 0.688, 0.681, 0.697, 0.685, 0.738, and 0.758, respectively, which confirmed the effectiveness of the 16-CpG-based model to predict the OS of LUAD patients (**Figure 3E**).

According to their clinicopathological conditions, like epidermal growth factor receptor (EGFR) mutation, K-ras or Ki-ras (KRAS) mutation, lymph node metastatic (LNM) condition, and AJCC stage, LUAD patients were divided up into several subgroups to validate the independent diagnostic value of the methylation signature. EGFR mutation showed a striking correlation with LUAD patient characteristics, which were correlated with the clinical treatment response and then affected the OS of LUAD patients. The Kaplan-Meier curves regarding EGFR mutation and wildtype groups were shown in **Figures 4A,B**. Patients with low risk scores generally had significantly better survival than those with high risk scores in both groups ($p < 0.0001$). Similarly, patients with low risk scores had a significantly longer OS than those with high risk scores in both KRAS mutation and wildtype subgroup and both LNM positive and negative groups (**Figures 4C–F**, $p < 0.0001$). For the patients in AJCC stage I and AJCC stage II-IV, the survival probability of patients with low risk scores was higher than those with low risk scores (**Figures 4G,H**). The stratification analysis above revealed that the 16-CpG-based model could effectively predict the OS of patients regardless of the patients' clinicopathological properties, and provide prognostic power to complement the clinical stage and SSIGN scores.

Lastly, the risk scores were applied to the Cox regression model with the clinicopathological risk factors to perform multivariable survival analysis, thereby generating a nomogram to predict patients' survival probability for 3 and 5 years (**Figure 5A**). In the multivariable survival analysis, we included age, gender, EGFR status, AJCC stage, and risk scores from 16-CpG-based model. The nomogram was further verified with calibration plots (**Figure 5B**). The results showed that the nomogram fared well with the ideal mode for 3 and 5 years, indicating the nomogram worked well in predicting the OS of LUAD patients. According to the risk scores from the nomogram, patients were divided into high risk and low risk group. Kaplan-Meier survival analysis showed that the survival probability of patients with low risk score was significantly higher than those with high risk score (**Figure 5C**, log-rank test, $p < 0.0001$). The prognostic accuracy of the nomogram was further accessed by time-dependent ROC curves (**Figure 5D**). The results showed that the AUC values were all around 0.7 at varying follow-up times (500, 1,000, 1,500, 2,000, 2,500, 3,000 days), indicating the high effectiveness of the nomogram in predicting the prognostic OS of LUAD patients.

DISCUSSION

The present study demonstrates the potential of using DNA methylation signatures to identify the lymph node metastasis of primary LUAD tissues, to differentiate between the LUAD tumor and normal tissues, and to predict the OS of LUAD patients. Invasive biopsy is the gold standard for the validation of tumor tissues and identification of histological subtypes. However, the collection of bulk tissue samples for immunohistochemical (IHC) staining may cause secondary damage to patients. An inadequate tissue yield or quality also creates barriers for the histological diagnosis. Besides, it may be difficult to identify lymph node metastasis during operation. Nowadays, molecular characterization methods provide new insights in pathological diagnosis (Tsou et al., 2007; Selamat et al., 2012; Zhang et al., 2013; Ogino et al., 2016). Since the global change of DNA methylation takes place at the beginning of carcinogenesis, DNA methylation has been considered a promising biomarker for the early detection and diagnosis of cancers (Franco et al., 2008; Hatano et al., 2015; Wu and Ni, 2015), which can complement the pathological IHC staining. Moreover, DNA methylation analysis does not require bulk tissue samples. Small amounts of tissue are enough for DNA extraction and methylation-chip or methylation-seq analysis, which will reduce the patients suffering. Hundreds of thousands of the DNA methylation CpG sites can be identified through genome-wide DNA methylation detection by DNA methylation chips or methylation-seq. Discovering a potential panel of DNA methylation-based biomarkers from the large DNA methylation files can be beneficial for the early diagnosis of cancer initiation and metastasis. Several research studies have shown the potential of utilizing DNA methylation profiles to help the diagnosis of different cancers (Diaz-Lagares et al., 2016; Zhang et al., 2017; Sandanger et al., 2018). One study applied an unsupervised clustering method on DNA methylation profiles to find potential subtypes of childhood B-cell acute lymphoblastic leukemia. The patients were allocated into two subgroups by the unsupervised hierarchical clustering of DNA methylation profiles, which showed a significant association between DNA methylation and disease-free survival (Sandoval et al., 2013a). Another study also utilized a similar strategy to find the association between DNA methylation signatures and the recurrence-free survival in non-small-cell lung cancer samples (Sandoval et al., 2013b). In our study, we applied supervised learning strategy (regularized logistic regression) to find the prognostic CpG cites in LUAD primary tissues. The RFE helped to eliminate the unnecessary features in regression, which constrained the numbers of key CpG sites for prognosis. Besides, LASSO Cox regression was useful to reduce the feature numbers in the COX survival analysis. One study built a

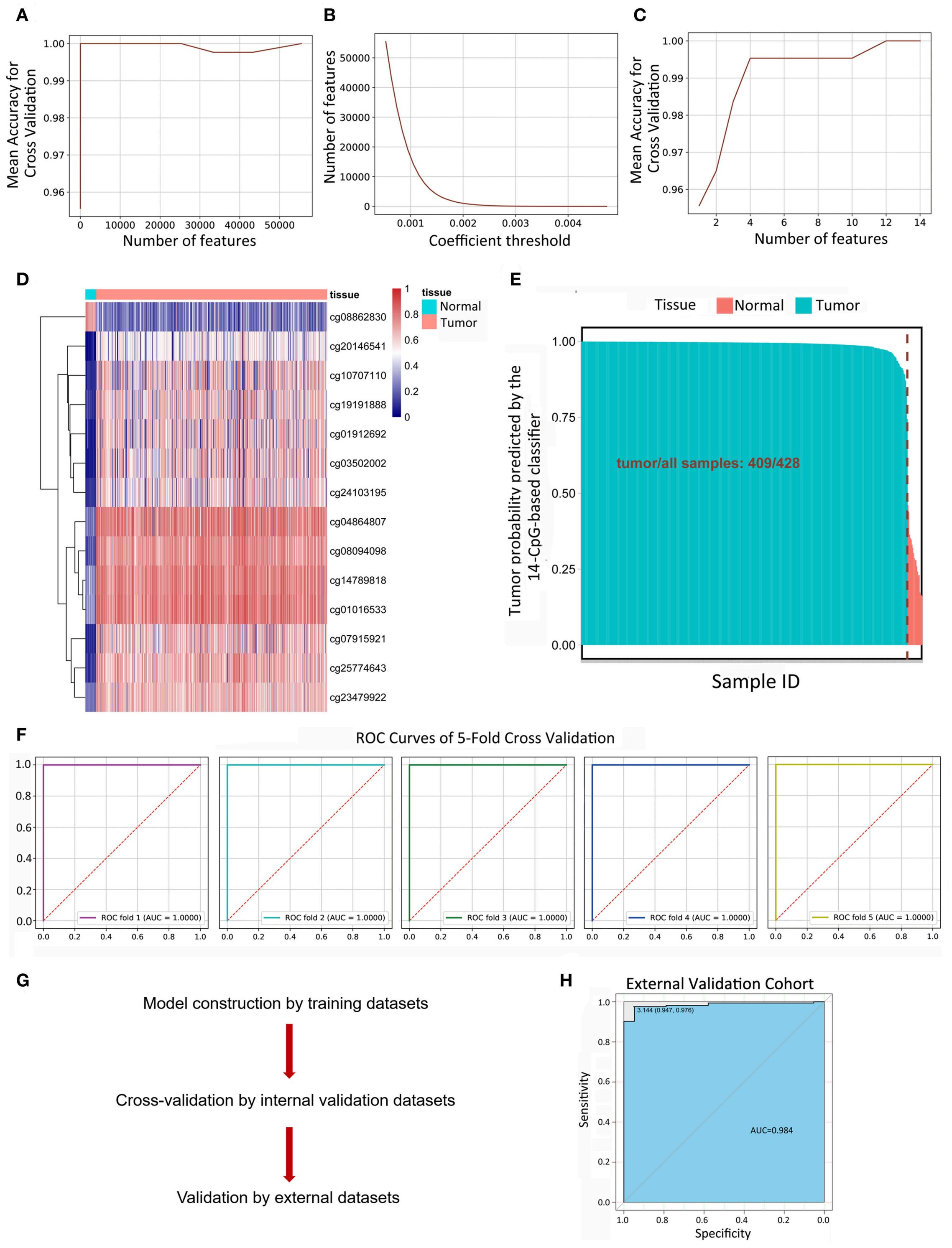

FIGURE 2 | Linear logistic classifier using a 5-fold cross-validation to differentiate between LUAD tumor tissues and corresponding normal tissues, and the efficiency for the 14-CpG-based classifier. **(A)** The mean accuracy trend for the number of features and mean accuracy for the cross-validation. **(B)** The methylation features with regard to the different values of coefficient thresholds. **(C)** Recursive feature elimination with a cross-validation test. **(D)** Unsupervised hierarchical clustering and heat map associated with the methylation profile (according to the color scale shown) to differentiate between LUAD tumor tissues and corresponding normal tissues. **(E)** The probability of being tumor for each sample calculated by the coefficients of methylation signatures. **(F)** ROC curves showing the high sensitivity and specificity in differentiating between LUAD tumor tissues and normal tissues. **(G)** The workflow of model construction, internal validation and external validation. **(H)** ROC curve showing the high sensitivity and specificity in differentiating between LUAD tumor tissues and corresponding normal tissues on an external dataset.

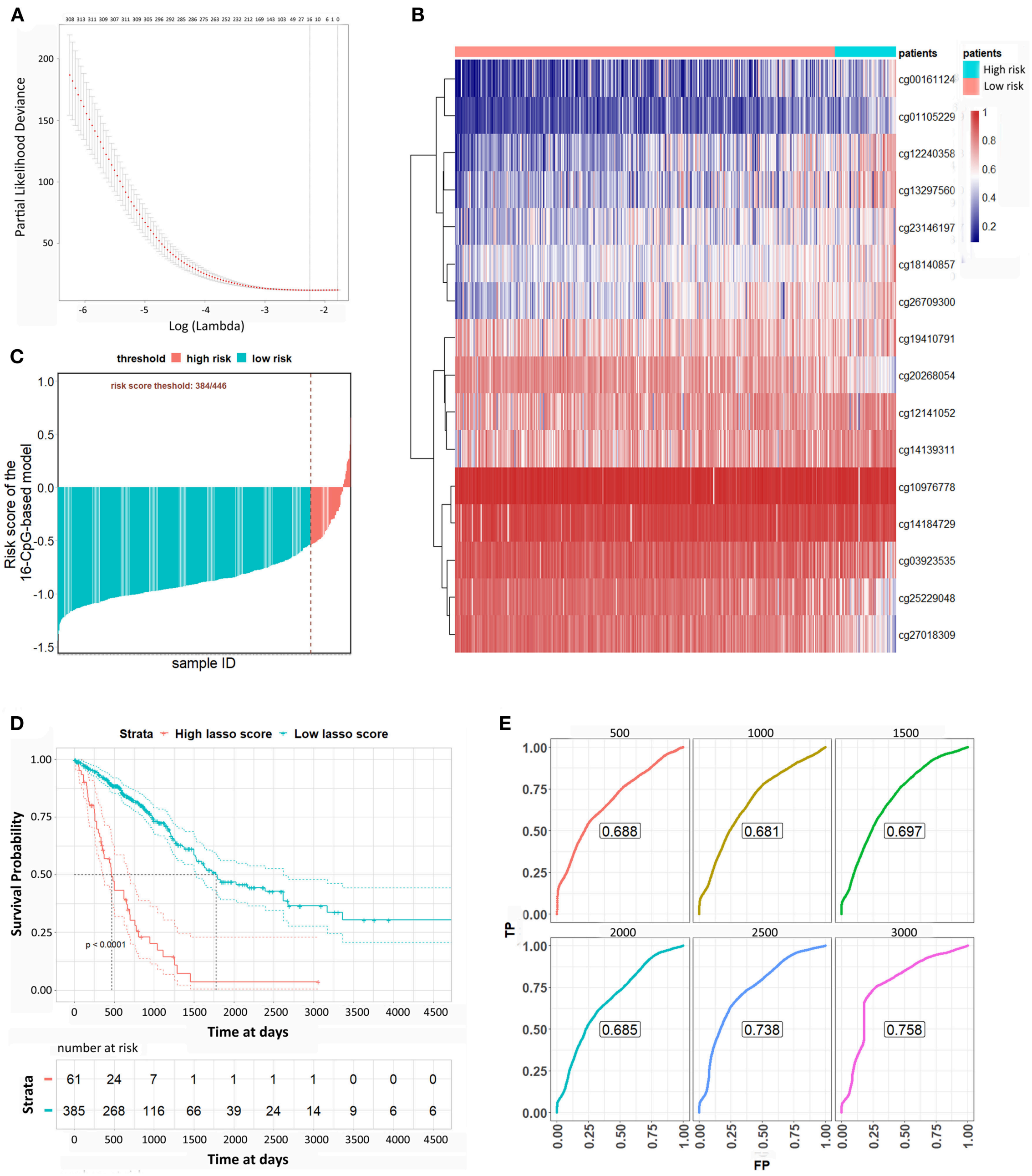

FIGURE 3 | Lasso Cox analysis to predict the OS of LUAD patients. **(A)** The selection of λ value for Lasso Cox regression. **(B)** Unsupervised hierarchical clustering and heat map associated with the methylation profile (according to the color scale shown) to predict the OS of LUAD patients. **(C)** The risk scores of each sample calculated by the coefficients of methylation signatures from Lasso Cox analysis. **(D)** Kaplan-Meier curves of LUAD patients with a low or high risk of death, according to risk scores from the 16-CpG-based classifier. **(E)** Time-dependent ROC analysis at varying follow-up times (500, 1,000, 1,500, 2,000, 2,500, 3,000 days). We used AUC values at 500, 1,000, 1,500, 2,000, 2,500, 3000 days to assess the prognostic accuracy.

prognostic signature by LASSO Cox regression to predict the progression-free survival of LUAD patients and demonstrated the potential biological significance of DNA methylation in the etiology of LUAD (Bjaanæs et al., 2016). Another study built a mortality risk score by LASSO Cox regression (Zhang et al., 2017). The signature based on ten selected CpG sites exhibited strong association with all-cause mortality. Moreover, one recent study used blood-derived DNA methylation and gene expression profiles to identify CpG lung cancer markers prior to diagnosis. They emphasized the difference of prognostic CpG sites in smoking and non-smoking lung cancer patients (Sandanger et al., 2018). In this study, based on the methylation

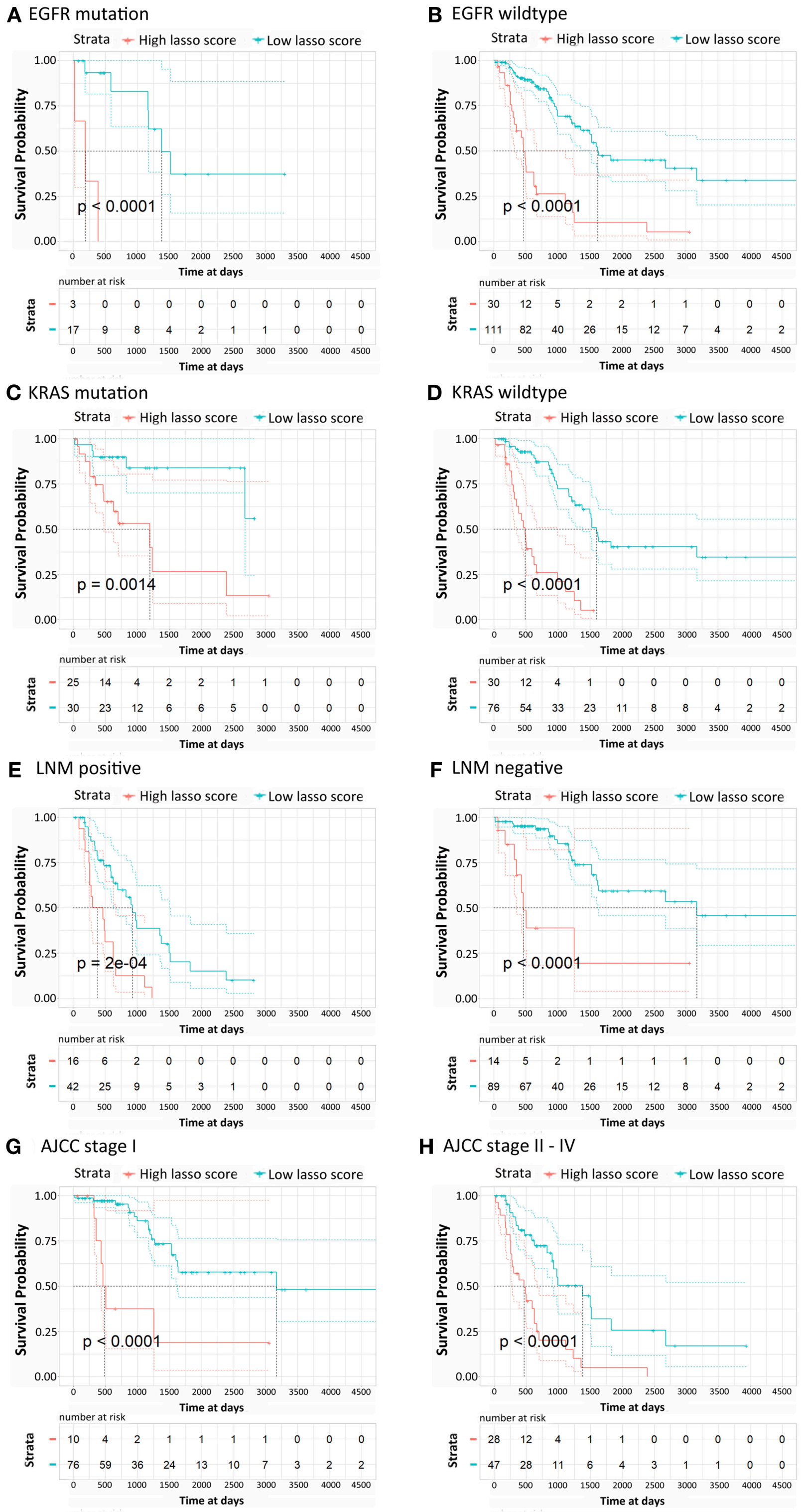

FIGURE 4 | Kaplan-Meier survival analysis for LUAD patients according to the 16-CpG-based classifier. Patients were classified according to clinicopathological risk factors. **(A,B)** EGFR status; **(C,D)** KRAS status; **(E,F)** lymph node metastatic (LNM) status; **(G,H)** AJCC stage I and II-IV. The patients were divided into low-risk and high-risk groups. *P*-values were calculated using the log-rank test.

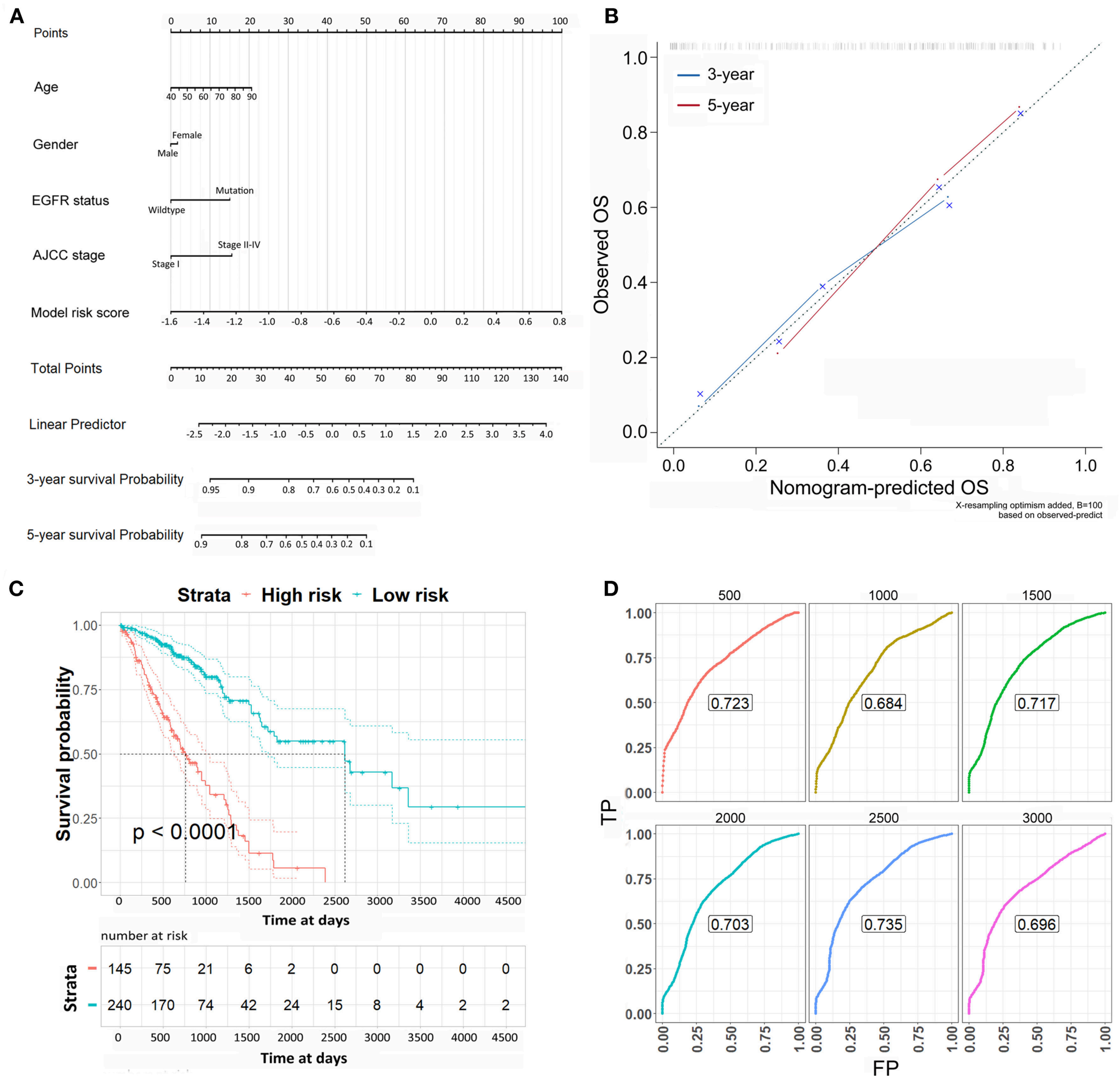

FIGURE 5 | The nomogram to predict the risk of death in 3 and 5 years of LUAD patients. **(A)** The nomogram to predict the risk of death in 3 and 5 years of LUAD patients. **(B)** Plots depict the calibration of each model in terms of agreement between predicted and observed 3- and 5-year outcomes. The dashed line represents the ideal line of a perfect match between nomogram-predicted and observed OS. The blue and red line indicate the performance of the proposed nomogram for 3 and 5 years, respectively. **(C)** Kaplan-Meier survival analysis for the OS of LUAD patients according to the risk scores from the nomogram. **(D)** Time-dependent ROC curves from the nomogram for overall survival in 3 and 5 years.

profiles of LUAD patients, we performed regularized logistic regression and LASSO Cox regression to identify the lymph node metastasis, to differentiate between tumor and normal tissues and to predict the OS of LUAD patients. From the primary LUAD tumor tissues, 3616 methylation CpG sites were kept to build a classifier to identify LUAD lymph node metastasis. ROC curves showed the high sensitivity and specificity of the 3616-CpG-based classifier in identifying lymph node metastasis from CpG sites of primary tumor tissues. All the samples came from the primary tumor tissues, which means that the metastatic behavior can be identified even without extracting tissues from lymph nodes. Therefore, it would work as a biomarker to predict the diagnosis of lymph node metastasis. Since the metastatic behavior of LUAD affects the OS of LUAD patients dramatically, we applied the metastatic classifier to check whether the model can be used to predict the OS of LUAD patients. The time-dependent ROC curves showed the effectiveness of the metastatic classifier in predicting the OS of LUAD patients at varying follow-up times. As expected, the patients in the high metastatic risk score group have a significantly worse OS than those in the low metastatic risk score group.

Tumor tissues are heterogeneous tissues that include cancer cells (epithelial cells), cancer stem cells, vascular epithelial cells and so on (Reya et al., 2001; Marusyk et al., 2012). More than 70% of the tumor tissues are cancer cells. The heterogeneity of tumor tissues may influence the accuracy of the diagnosis. We compared the heterogeneous tumor tissues with the normal tissues, which also include the vascular epithelial cells and other cell types. Considering the heterogeneous tumor tissues and heterogeneous normal tissues as a whole for each, we tried to eliminate the influence brought about by heterogeneity (Li et al., 2014). In this study, to differentiate between tumor and normal tissues, we concluded that 14 CpG methylation sites were enough for the diagnosis. To check the overfitting potential, we applied 5-fold cross-validation. The efficiency of the model above was tested by a ROC curves in five different training and validation datasets, which showed the high efficiency and specificity of the 14-CpG-based classifier in differentiating between LUAD tumor tissues and the normal tissues. Furthermore, we also validated our regression model on the external dataset from one study (Bjaanæs et al., 2016). Results showed an AUC value of 98.4% to differentiate the tumor and normal tissues by the ROC analysis. The external dataset further confirmed the accuracy of the regularized logistic model which we applied to build the both classifiers above. From the two classifiers, we obtained an overlap cluster of CpG sites: cg03502002 and cg07915921. The information for cg07915921 is not clear. cg03502002 is on the CpG island of the promotor region of the GALR1 gene. The methylation status of the GALR1 promoter and the level of GALR1 gene expression have been correlated in a large number of head and neck squamous tumor specimens (Misawa et al., 2008). Ectopic expression of GALR1 suppresses tumor cell proliferation through Erk1/2-mediated regulation of cyclin-dependent kinase inhibitors and cyclin D1 (Kanazawa et al., 2009). One study revealed that hypermethylated GALR1 plays important roles in smoking-associated LUAD (Tan et al., 2013).

We also built a model to predict the OS of LUAD patients by means of methylation CpG sites. The LASSO Cox regression model generated risk score for each patient. When we assessed the survival status and distribution of risk scores, patients with low risk scores generally had a better OS than those with high risk scores. The model will help guide individualized follow-up schedules for LUAD patients. The high-risk patients have poor OS prediction. This could be the basis of a future clinical trial. The LASSO Cox regression results were further confirmed by the time-dependent ROC analysis. When we compared the time-dependent ROC from the OS-prediction model and metastasis-prediction classifier, the OS-prediction model turned out to be more precise in the long-term survival prediction while the metastasis-prediction classifier worked better in the short-term survival prediction. One explanation could be that when the LUAD patients were accompanied by lymph node metastasis, the tumor progressed and the patients had a poorer prognosis. The OS expectation of patients with lymph node metastasis was shorter than those without lymph node metastasis. Hence, the metastasis-prediction classifier would work better for the short-term prediction.

To further utilize the risk scores from the Cox regression model, we classified patients into several subgroups according to the clinicopathological risk factors (EGFR mutation, KRAS mutation, LNM status and AJCC stages). The 16-CpG-based classifier still showed clinical and statistical significance regardless of the clinicopathological status of LUAD patients.

The independent prognostic values of the 16-CpG-based model were validated by multivariable survival analysis, which integrated other clinicopathological risk factors for the OS of LUAD patients. The Cox regression risk scores were applied together with age, gender, EGFR status, AJCC stages as indicators to generate a nomogram to predict the 3- and 5-year survival probability. We verified the performance of the nomogram by calibration plots. The predicted OS of LUAD patients by the nomogram was highly consistent with the observed 3- and 5-year OS of LUAD patients. Log-rank test and time-dependent ROC curves at vary follow-up times further confirmed the nomogram. Thus, the nomogram could provide an accurate and simple prognostic prediction for LUAD patients.

In previous studies, mRNA expression files (Beer et al., 2002), the mutation of key genes (Takano et al., 2008; Kosaka et al., 2009), long no-coding RNA expressions (Kosaka et al., 2009; Huarte, 2015; Zhou et al., 2016), and histone modifications (Seligson et al., 2009; Zhou et al., 2016) showed the prognostic potential for different types of cancer. Here, we emphasized that the methylation patterns could also be a meaningful tool for the prognosis of LUAD patients. Some studies have identified that multiple CpG sites are differentially methylated in lung cancer compared to normal tissues (Genome Atlas Research Network., 2014; Poirier et al., 2015; Hao et al., 2017). The key for methylation pattern-based early diagnosis is the identification of crucial CpG sites in LUAD. The use of supervised machine learning methods allowed us to integrate all methylation CpG sites identified by the methylation chip into one model, which improved the prognostic accuracy over that of a single CpG site alone. Our findings show that three DNA CpG signature-based models can effectively identify lymph node metastasis by the CpG sites from primary tumor tissues, differentiate between tumor and normal tissues, and predict the OS of LUAD patients. The tissues would be collected by preoperative biopsy or at surgery. The classifiers for identifying lymph node metastasis and differentiation between tumor and normal tissues would help the preoperative diagnosis. The Lasso Cox model would be helpful for adjuvant treatment and prognostic planning. Therefore, the 3 methylation signatures could be of great value in assessing the status, predicting prognosis and achieving individualized treatments of LUAD patients.

The limitations of our study should be mentioned. The methylation 450 k chip did not identify as many CpG sites as the methylation 850 k chip or methylation sequencing. The methylation CpG site candidates identified here did not represent the complete CpG sites in the genome of LUAD patients.

In conclusion, we built three DNA CpG signature-based models to identify LUAD lymph node metastasis by the CpG sites from primary tumor tissues, differentiate between tumor tissue and normal tissue, and predict the OS of LUAD patients, which highlight the relationship between clinical results (metastasis, survival) and methylation biomarkers in LUAD patients. The nomogram comprising LASSO Cox risk scores and clinicopathological factors may help predict the OS of LUAD patients and help individualized treatment of LUAD patients.

AUTHOR CONTRIBUTIONS

XB and RS conceived and designed the experiments. XB and HD wrote the code (The Python code is available due to request). XB, YW, and SX wrote the paper. RS and KZ reviewed the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We greatly thank Dr. Michael Rosemann and Prof. Dr. Michael J. Atkinson for helpful discussions and suggestions. We sincerely thank Sino-German (CSC-DAAD) Postdoc Scholarship Program for supporting the research of YW, and greatly thank the China Scholarship Council (CSC) for supporting the research and work of SX, RS, and XB.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00349/full#supplementary-material

Supplementary Figure 1 | Kaplan-Meier survival according to risk scores from the 3616-CpG-based classifier in the training, validation sets for 5-fold cross-validation. **(A, C, E, G, I)** The training sets 1-5. **(B, D, F, H, J)** The validation sets 1-5.

Supplementary Figure 2 | Time-dependent ROC analysis at varying follow-up times (500, 1,000, 1,500, 2,000, 2,500, 3,000 days) according to risk scores from the 3616-CpG-based classifier.

REFERENCES

Beer, D. G., Kardia, S. L., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., et al. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.* 8, 816–824. doi: 10.1038/nm733

Bjaanæs, M. M., Fleischer, T., Halvorsen, A. R., Daunay, A., Busato, F., Solberg, S., et al. (2016). Genome-wide DNA methylation analyses in lung adenocarcinomas: association with EGFR, KRAS and TP53 mutation status, gene expression and prognosis. *Mol. Oncol.* 10, 330–343. doi: 10.1016/j.molonc.2015.10.021

Cline, M. S., Craft, B., Swatloski, T., Goldman, M., Ma, S., Haussler, D., et al. (2013). Exploring TCGA pan-cancer data at the UCSC cancer genomics browser. *Sci. Rep.* 3:2652. doi: 10.1038/srep02652

Diaz-Lagares, A., Mendez-Gonzalez, J., Hervas, D., Saigi, M., Pajares, M. J., Garcia, D., et al. (2016). A novel epigenetic signature for early diagnosis in lung cancer. *Clin. Cancer Res.* 22, 3361–3371. doi: 10.1158/1078-0432.CCR-15-2346

Du, J., Johnson, L. M., Jacobsen, S. E., and Patel, D. J. (2015). DNA methylation pathways and their crosstalk with histone methylation. *Nat. Rev. Mol. Cell Biol.* 16, 519–532. doi: 10.1038/nrm4043

Ettinger, D. S., Akerley, W., Bepler, G., Blum, M. G., Chang, A., Cheney, R. T., et al. (2010). Non–small cell lung cancer. *J. Natl. Compreh. Cancer Net.* 8, 740–801. doi: 10.6004/jnccn.2010.0056

Franco, R., Schoneveld, O., Georgakilas, A. G., and Panayiotidis, M. I. (2008). Oxidative stress, DNA methylation and carcinogenesis. *Cancer Lett.* 266, 6–11. doi: 10.1016/j.canlet.2008.02.026

Genome Atlas Research Network. (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 511, 543–550. doi: 10.1038/nature13385

Hankey, B. F., Ries, L. A., and Edwards, B. K. (1999). The surveillance, epidemiology, and end results program: a national resource. *Cancer Epidemiol. Prev. Biomarkers* 8, 1117–1121.

Hao, X., Luo, H., Krawczyk, M., Wei, W., Wang, W., Wang, J., et al. (2017). DNA methylation markers for diagnosis and prognosis of common cancers. *Proc. Natl. Acad. Sci. U.S.A.* 114, 7414–7419. doi: 10.1073/pnas.1703577114

Hatano, Y., Semi, K., Hashimoto, K., Lee, M. S., Hirata, A., Tomita, H., et al. (2015). Reducing DNA methylation suppresses colon carcinogenesis by inducing tumor cell differentiation. *Carcinogenesis* 36, 719–729. doi: 10.1093/carcin/bgv060

Herman, J. G., Merlo, A., Mao, L., Lapidus, R. G., Issa, J.-P. J., Davidson, N. E., et al. (1995). Inactivation of the CDKN2/p16/MTS1 gene is frequently associated with aberrant DNA methylation in all common human cancers. *Cancer Res.* 55, 4525–4530.

Huarte, M. (2015). The emerging role of lncRNAs in cancer. *Nat. Med.* 21, 1253–1261. doi: 10.1038/nm.3981

Jemal, A., Miller, K. D., Ma, J., Siegel, R. L., Fedewa, S. A., Islami, F., et al. (2018). Higher lung cancer incidence in young women than young men in the United States. *N. Engl. J. Med.* 378, 1999–2009. doi: 10.1056/NEJMoa1715907

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* 13, 484–492. doi: 10.1038/nrg3230

Kanazawa, T., Kommareddi, P. K., Iwashita, T., Kumar, B., Misawa, K., Misawa, Y., et al. (2009). Galanin receptor subtype 2 suppresses cell proliferation and induces apoptosis in p53 mutant head and neck cancer cells. *Clin. Cancer Res.* 15, 2222–2230. doi: 10.1158/1078-0432.CCR-08-2443

Klutstein, M., Nejman, D., Greenfield, R., and Cedar, H. (2016). DNA methylation in cancer and aging. *Cancer Res*. 76, 3446–3450. doi: 10.1158/0008-5472.CAN-15-3278

Kosaka, T., Yatabe, Y., Onozato, R., Kuwano, H., and Mitsudomi, T. (2009). Prognostic implication of EGFR, KRAS, and TP53 gene mutations in a large cohort of Japanese patients with surgically treated lung adenocarcinoma. *J. Thorac. Oncol.* 4, 22–29. doi: 10.1097/JTO.0b013e3181914111

Li, X., Shi, Y., Yin, Z., Xue, X., and Zhou, B. (2014). An eight-miRNA signature as a potential biomarker for predicting survival in lung adenocarcinoma. *J. Transl. Med.* 12:159. doi: 10.1186/1479-5876-12-159

Marusyk, A., Almendro, V., and Polyak, K. (2012). Intra-tumour heterogeneity: a looking glass for cancer? *Nat. Rev. Cancer* 12, 323–334. doi: 10.1038/nrc3261

Misawa, K., Ueda, Y., Kanazawa, T., Misawa, Y., Jang, I., Brenner, J. C., et al. (2008). Epigenetic inactivation of galanin receptor 1 in head and neck cancer. *Clin. Cancer Res.* 14, 7604–7613. doi: 10.1158/1078-0432.CCR-07-4673

Ogino, S., Nishihara, R., VanderWeele, T. J., Wang, M., Nishi, A., Lochhead, P., et al. (2016). The role of molecular pathological epidemiology in the study of neoplastic and non-neoplastic diseases in the era of precision medicine. *Epidemiology* 27, 602–611. doi: 10.1097/EDE.0000000000000471

Poirier, J. T., Gardner, E. E., Connis, N., Moreira, A. L., De Stanchina, E., Hann, C. L., et al. (2015). DNA methylation in small cell lung cancer defines distinct disease subtypes and correlates with high expression of EZH2. *Oncogene* 34, 5869–5878. doi: 10.1038/onc.2015.38

Ramchandani, S., Bhattacharya, S. K., Cervoni, N., and Szyf, M. (1999). DNA methylation is a reversible biological signal. *Proc. Natl. Acad. Sci. U.S.A.* 96, 6107–6112. doi: 10.1073/pnas.96.11.6107

Reya, T., Morrison, S. J., Clarke, M. F., and Weissman, I. L. (2001). Stem cells, cancer, and cancer stem cells. *Nature* 414, 105–111. doi: 10.1038/35102167

Sandanger, T. M., Nøst, T. H., Guida, F., Rylander, C., Campanella, G., Muller, D. C., et al. (2018). DNA methylation and associated gene expression in blood prior to lung cancer diagnosis in the Norwegian Women and Cancer cohort. *Sci. Rep.* 8:16714. doi: 10.1038/s41598-018-34334-6

Sandoval, J., Heyn, H., Méndez-González, J., Gomez, A., Moran, S., Baiget, M., et al. (2013a). Genome-wide DNA methylation profiling predicts relapse in childhood B-cell acute lymphoblastic leukaemia. *Br. J. Haematol.* 160, 406–409. doi: 10.1111/bjh.12113

Sandoval, J., Mendez Gonzalez, J., Nadal, E., Chen, G., Carmona, F. J., Sayols, S., et al. (2013b). A prognostic DNA methylation signature for stage I non-small-cell lung cancer. *J. Clin. Oncol.* 31, 4140–4147. doi: 10.1200/JCO.2012.48.5516

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

Selamat, S. A., Chung, B. S., Girard, L., Zhang, W., Zhang, Y., Campan, M., et al. (2012). Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. *Genome Res*. 22, 1197–1211. doi: 10.1101/gr.132662.111

Seligson, D. B., Horvath, S., McBrian, M. A., Mah, V., Yu, H., Tze, S., et al. (2009). Global levels of histone modifications predict prognosis in different cancers. *Am. J. Pathol.* 174, 1619–1628. doi: 10.2353/ajpath.2009.080874

Takano, T., Fukui, T., Ohe, Y., Tsuta, K., Yamamoto, S., Nokihara, H., et al. (2008). EGFR mutations predict survival benefit from gefitinib in patients with advanced lung adenocarcinoma: a historical comparison of patients treated before and after gefitinib approval in Japan. *J. Clin. Oncol.* 26, 5589–5595. doi: 10.1200/JCO.2008.16.7254

Tan, Q., Wang, G., Huang, J., Ding, Z., Luo, Q., Mok, T., et al. (2013). Epigenomic analysis of lung adenocarcinoma reveals novel DNA methylation patterns associated with smoking. *Onco. Targets. Ther.* 6:1471–1479. doi: 10.2147/OTT.S51041

Toh, C.-K., Gao, F., Lim, W.-T., Leong, S.-S., Fong, K.-W., Yap, S.-P., et al. (2006). Never-smokers with lung cancer: epidemiologic evidence of a distinct disease entity. *J. Clin. Oncol.* 24, 2245–2251. doi: 10.1200/JCO.2005.04.8033

Travis, W. D., Brambilla, E., Noguchi, M., Nicholson, A. G., Geisinger, K. R., Yatabe, Y., et al. (2011). International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. *J. Thorac. Oncol.* 6, 244–285. doi: 10.1097/JTO.0b013e318206a221

Tsou, J. A., Galler, J. S., Siegmund, K. D., Laird, P. W., Turla, S., Cozen, W., et al. (2007). Identification of a panel of sensitive and specific DNA methylation markers for lung adenocarcinoma. *Mol. Cancer* 6:70. doi: 10.1186/1476-4598-6-70

Vizoso, M., Puig, M., Carmona, F. J., Maqueda, M., Velásquez, A., Gómez, A., et al. (2015). Aberrant DNA methylation in non-small cell lung cancer-associated fibroblasts. *Carcinogenesis* 36, 1453–1463. doi: 10.1093/carcin/bgv146

Wu, Q., and Ni, X. (2015). ROS-mediated DNA methylation pattern alterations in carcinogenesis. *Curr. Drug Targets* 16, 13–19. doi: 10.2174/1389450116666150113121054

Zhang, J.-X., Song, W., Chen, Z.-H., Wei, J.-H., Liao, Y.-J., Lei, J., et al. (2013). Prognostic and predictive value of a microRNA signature in stage II colon cancer: a microRNA expression analysis. *Lancet Oncol.* 14, 1295–1306. doi: 10.1016/S1470-2045(13)70491-1

Zhang, Y., Wilson, R., Heiss, J., Breitling, L. P., Saum, K.-U., Schöttker, B., et al. (2017). DNA methylation signatures in peripheral blood strongly predict all-cause mortality. *Nat. Commun.* 8:14617. doi: 10.1038/ncomms14617

Zhou, M., Xu, W., Yue, X., Zhao, H., Wang, Z., Shi, H., et al. (2016). Relapse-related long non-coding RNA signature to improve prognosis prediction of lung adenocarcinoma. *Oncotarget* 7, 29720–29738. doi: 10.18632/oncotarget.8825

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.