Abstract. Lung adenocarcinoma (LACA) is the leading cause of cancer-associated death worldwide. The present study intended to identify DNA methylation patterns that may serve as diagnostic and prognostic biomarkers for LACA. Data on DNA methylation and the survival data of the patients of LACA were obtained from The Cancer Genome Atlas. Kaplan-Meier curves and receiver operating characteristic curve analysis were utilized to build diagnostic and prognostic models. A total of 13 CpG sites were identified and validated as the optimal diagnostic and prognostic signature for overall survival. It was concluded that the CpG-based signature is a reliable predictor for the diagnosis and prognosis of patients with LACA.

Introduction

Lung cancer, including small cell lung cancer (SCLC) and non-SCLC (NSCLC) (1,2), may be regarded as the most common tumor type and major contributor to the high tumor-associated mortality rate worldwide (3). Recent epidemiological data have revealed that it affected 1.8 million patients and resulted in 1.6 million deaths in 2012 (4). As one of the most frequent histological subtypes of lung cancer, lung adenocarcinoma (LACA) is a major cause of tumor-associated death (5,6). Despite recent advances in surgical techniques, radiotherapeutic interventions and combined chemotherapy strategies, the long-term survival rate of patients diagnosed with primary LACA has not significantly improved (7,8). Due to tumor heterogeneity factors and different molecular subtypes of LACA, its treatment faces large challenges. In this light, it is significant to identify specific details regarding characteristic molecules in LACA tissues to delineate the heterogeneity of LACA and develop strategies for personalized therapy.

In recent years, epigenetics, which has a critical role in carcinogenesis, has gained attention (9). Aberrant DNA methylation, as the core element of epigenetic modification, influences certain tumor suppressor genes and regulates gene functions (10,11). Increasing studies also demonstrated that DNA methylation is associated with genome stability, gene imprinting and cell differentiation (12,13). Thus, the methylation level is deemed a molecular biomarker for the diagnosis and prognostication of patients with tumors. However, the current expertise on the association between the epigenetic modifications and the clinically predicted outcomes of LACA is limited. Thus, in the present study, distinctive DNA methylation data for LACA vs. control tissues were acquired to evaluate the prognostic significance of distinctive DNA methylation patterns and provide insight regarding survival prediction for patients with LACA.

Materials and methods

Data processing. Original, publicly available and open-access genetic representation data of LACA samples and relevant clinical information of the patients obtained from The Cancer Genome Atlas (TCGA) database (http://cancergenome.nih.gov/) were included in the present study. The clinical information included the following attributes: Age, sex, ethnicity, stage and histological type of LACA. The exclusion criteria were as follows: i) Samples without clinical information, ii) samples without survival data, and iii) samples from patients that survived for <1 month. Ultimately, 447 LACA samples with DNA methylation data and clinical information were screened for further testing. The data were provided by TCGA and the
study was performed in compliance with the TCGA publication guidelines (14).

Selection of differential DNA methylation sites. In the present study, aggregation and collection of DNA methylation information was performed by using R. First, the data were normalized by log2 transformation. The Limma package was employed for analyzing the differential DNA methylation sites between LACA tumor tissues and normal tissues. The fold changes (FCs) of DNA methylation were also calculated and significant aberrations in gene methylation were defined as those having a log2FC>2.0, P-value <0.01 and beta value >0.1. The least absolute shrinkage and selection operator (LASSO) method, which is suitable for regression of high-dimensional data (15), was used to select the most useful predictive features from the primary data set. The potential association of the CpG-based signature with LACA status was first assessed in the primary cohort and then validated in the validation cohort using a Mann-Whitney U-test. With this CpG-based signature, patients in each dataset were classified into a high-risk group and a low-risk group by using the median risk score. Kaplan-Meier curves and log-rank analysis were then performed to calculate the association between the CpG-based signature and patient’s OS in the two groups with high-risk and low-risk CpG-based signatures. The CpG-based signature, which was significantly associated with OS (P<0.001), was then subjected to receiver operating characteristic (ROC) curve analysis to evaluate the predictive accuracy and sensitivity of the prognostic model. The area under the ROC curve (AUC) was also calculated. In the Kaplan-Meier curve, log-rank test and ROC analysis, the significance was defined as P<0.05.

Functional and pathway enrichment analysis. To further elucidate the biological functions of the mapped genes and the molecular mechanisms, a functional enrichment analysis was performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) (16), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and Gene Ontology (GO) enrichment analyses with P<0.05 were identified and biological process terms were further clustered using the Enrichment Map Plugin of Cytoscape (17).

Statistical analyses. Fundamental characteristics of the sample in the study were summarized by using descriptive statistics. Data for continuous variables, which were expressed as the mean ± standard deviation, were compared using the Student's t-test, Mann-Whitney U-test, Kruskal-Wallis H-test or one-way ANOVA with post hoc Student-Newman-Keuls tests, depending on the normality of data distribution as tested by Kolmogorov-Smirnov tests; data for categorical variables, which were presented as percentages, were compared using the Chi-square test. Statistical analyses were performed using SPSS 17.0 (SPSS Inc.) and R version 3.5.1 software (http://www.r-project.org/) (18). P<0.05 was considered to indicate a statistically significant difference.

Results

Patient characteristics. The data of all 447 samples with clinical information and methylation data available were obtained from the TCGA database. The samples included 440 LACA tissues and 7 normal tissues. Table I lists the detailed clinical characteristics of patients in the initial stage and specific groups (primary and validation cohort), including age at diagnosis, sex, ethnicity, disease stage and survival status. The results revealed that there were no major distinctions between the two groups in terms of these five clinical characteristics.

Differential methylation sites in LACA. A total of 209 differential methylation sites were recognized between the LACA and regular tissue samples in the present study, including 133 hypermethylation and 76 hypomethylation sites. The distribution of hypermethylated and hypomethylated sites were visualized in Fig. 1. Furthermore, the methylation data were analyzed using the Limma incremental model. The five most hypermethylated sites (cg16306898, cg006648301, cg01869632, cg18837178 and cg22449330) and hypomethylated sites (cg05100666, cg15998127, cg07764932, cg12581354 and cg27649653) between LACA and normal tissue samples are listed in Table II.

Methylated loci signature. Based on the LASSO regression, 13 potential predictors were showed in the primary cohort (Fig. S1). As indicated in Fig. 2, Kaplan-Meier curves were drawn and the log-rank test was performed to evaluate the association between the CpG-based signals and the survival rates of patients with LUAD. The 13 methylated CpG signature, including cg00002719, cg02769743, cg05239163, cg05507908, cg07918170, cg08213398, cg08516516, cg08623223, cg12748948, cg14904034, cg16007456, ch.6.2958553R and cg19868631, was a significant predictor of OS in the primary cohort (P<0.01; Fig. 2A). The risk-score formula is as follows: (0.93x cg00002719) + (1.15x cg02769743) + (0.24x cg05239163) + (-1.20x cg05507908) + (1.31x cg07918170) + (0.83x cg08213398) + (0.54x cg08516516) + (0.06x cg08623223) + (0.12x cg12748948) + (0.03x cg14904034) + (-0.36x cg16007456) + (-0.36x ch.6.2958553R) + (-0.21x cg19868631). The mapped genes of the methylation sites are also provided in Table III. In addition, for identifying LACA, the approach of ROC curve analysis was pursued (Fig. 3). The AUC was 0.79, and the optimal cut-off value was 0.11 (Fig. 3A).

Diagnostic and prognostic validation of the signature. To assess the utility of the 13 CpG-based signature in the diagnosis and prognosis of LUAD, the above-mentioned analyses were performed using the validation cohort. There was a marked distinction between the high-risk and low-risk groups in the primary cohort (P<0.01), which successfully provided confirmation in this process of validation (Fig. 2B). Subsequently, the signatures were tested using an ROC analysis of the validation cohort and the results revealed the AUC was 0.70, indicating that the signature was an effective predictor for LACA, although the AUC was lower than that in the primary cohort (Fig. 2B).

Functional enrichment analysis. The significant terms of the GO enrichment analysis performed by DAVID and the KEGG pathways are provided in Fig. 4 and Fig. S2. The genes were significantly enriched KEGG pathways including extracellular matrix (ECM)-receptor interaction, cell adhesion molecules,
PI3K-Akt signaling pathway, dilated cardiomyopathy and cysteine and methionine metabolism. In addition, the GO biological process terms were mainly enriched in ECM organization, embryonic digit morphogenesis, cell adhesion, cardiovascular system development, sequence-specific DNA binding, positive regulation of transcription from RNA polymerase II promoter and positive regulation of cell differentiation.

Discussion

Due to high invasion and poor prognosis, the outcome for patients with LACA remains unsatisfactory, with a 5-year OS rate of 4-17%, depending on the stage and regional differences (19,20). Most lung cancer patients are in the advanced stages at the time of diagnosis. It is achievable to enhance the efficiency of diagnosing and prognosticating LACA patients once the indication of the tumor’s presence is able to be detected and explored at an early stage. Thus, in-depth studies on the aetiological elements and progressive mechanisms, early detection of the prognostic markers and identification of specific methylation CpG sites are urgently required.

For the purpose of clear classification of different secondary types of cancer, making use of CpG methylation locations is probably more efficient than it would be to collect detailed information on genetic representation based on covalent chemical alterations and steady hysterogenic markers of conjugated duplication (21). Hence, in the present study, a signature of 13 CpG sites with differential methylation was established, which were as follows: cg05100666, cg15998127, cg07764932, cg12581354, cg27649653, cg16306898, cg00648301, cg01869632, cg18837178, cg22449330, cg00002719, cg02769743, cg05239163, cg05507908, cg07918170, cg08213398, cg08516516, cg08623223, cg12748948, cg14904034, cg16007456, ch.6.2958553R and cg19868631. The 13 CpG signature, which was significantly associated with the OS of patients with LACA, was also recognized as an independent element for diagnosing LACA and predicting the prognosis of the patients. Furthermore, compared with that of
the high-risk patients, low-risk patients had better OS, and the 5-year survival rate in low-risk patients was also higher than that in high-risk patients. The functional enrichment analysis for the mapped genes of the CpG methylation sites were all presented by means of approaches of the fields of biology and information technology. The outcomes indicate that the methylation sites included in the 13-CpG-based prognostic signature may have a role in the molecular pathogenetic mechanisms and clinical progression in LACA patients, and this provides novel information for survival prediction and personalized treatment of patients with LACA.

As an important epigenetic mechanism in tumors, DNA methylation has a critical role in carcinogenesis (22). It regulates the extent of gene expression to control the function of the biomolecules encoded by those genes (23,24). Numerous studies have revealed that DNA methylation has a significant role in the initiation, progression and metastasis of cancer by controlling different aspects, including DNA repair, cell cycle regulation, angiogenesis and apoptosis (25). In the present study, 13 mapped genes were identified to be aberrantly methylated in LACA, which were as follows: Coiled-coil domain containing 181 (CCDC181), transmembrane protein 201 (TMEM201), chromosome 1 open reading frame 43 (1orf43), ankyrin repeat domain 31 (ANKRD31), tubulin folding cofactor D (TBCD), ubiquitin specific peptidase 48 (USP48), G protein-coupled receptor 88 (GPR88), spectrin repeat containing nuclear envelope protein 1 (SYNE1), V-set and transmembrane domain containing 2A (VSTM2A).

### Table III. Methylation loci significantly associated with survival.

<table>
<thead>
<tr>
<th>Composite</th>
<th>Chromosome</th>
<th>Start</th>
<th>End</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg00002719</td>
<td>1</td>
<td>169427468</td>
<td>169427469</td>
<td>CCDC181</td>
</tr>
<tr>
<td>cg02769743</td>
<td>1</td>
<td>9608344</td>
<td>9608345</td>
<td>TMEM201</td>
</tr>
<tr>
<td>cg05239163</td>
<td>1</td>
<td>154218790</td>
<td>154218791</td>
<td>C1orf43</td>
</tr>
<tr>
<td>cg05507908</td>
<td>5</td>
<td>75237454</td>
<td>75237455</td>
<td>ANKR31</td>
</tr>
<tr>
<td>cg07918170</td>
<td>17</td>
<td>82932272</td>
<td>82932273</td>
<td>TBCD</td>
</tr>
<tr>
<td>cg08213398</td>
<td>11</td>
<td>9722579</td>
<td>9722580</td>
<td>SWAP70</td>
</tr>
<tr>
<td>cg08516516</td>
<td>5</td>
<td>115816795</td>
<td>115816796</td>
<td>CDO1</td>
</tr>
<tr>
<td>cg08623223</td>
<td>8</td>
<td>11283906</td>
<td>11283907</td>
<td>AF131216.1</td>
</tr>
<tr>
<td>cg12748948</td>
<td>1</td>
<td>21779802</td>
<td>21779803</td>
<td>USP48</td>
</tr>
<tr>
<td>cg14904034</td>
<td>2</td>
<td>10389150</td>
<td>10389151</td>
<td>HPCAL1</td>
</tr>
<tr>
<td>cg16007456</td>
<td>1</td>
<td>100539082</td>
<td>100539083</td>
<td>GPR88</td>
</tr>
<tr>
<td>ch.6.2958553R</td>
<td>6</td>
<td>152198831</td>
<td>152198831</td>
<td>SYNE1</td>
</tr>
<tr>
<td>cg19868631</td>
<td>7</td>
<td>54542083</td>
<td>54542084</td>
<td>VSTM2A</td>
</tr>
</tbody>
</table>

CCDC181, coiled-coil domain containing 181; TMEM201, transmembrane protein 201; C1orf43, chromosome 1 open reading frame 43; ANKRD31, ankyrin repeat domain 31; TBCD, tubulin folding cofactor D; SWAP70, switching B cell complex subunit SWAP70; CDO1, cysteine dioxygenase type 1; USP48, ubiquitin specific peptidase 48; HPCAL1, hippocalcin like 1; GPR88, G protein-coupled receptor 88; SYNE1, spectrin repeat containing nuclear envelope protein 1; VSTM2A, V-set and transmembrane domain containing 2A.

Figure 2. Kaplan-Meier survival curves of the methylated sites in (A) the primary cohort and (B) the validation cohort. The red data-points represent high risk group and the blue data-points represent low risk group. HR, hazard ratio.
(TMEM201), chromosome 1 open reading frame 43 (Clorf43), ankyrin repeat domain 31 (ANKRD31), tubulin folding cofactor D (TBCD), switching B cell complex subunit SWAP70 (SWAP70), cysteine dioxygenase type 1 (CDO1), AF131216.1, ubiquitin specific peptidase 48 (USP48), hippocalcin like 1 (HPCAL1), G protein-coupled receptor 88 (GPR88), spectrin repeat containing nuclear envelope protein 1 (SYNE1) and V-set and transmembrane domain containing 2A (VSTM2A). Previous studies have directly shown that CCDC181, CDO1 and SYNE1 are associated with LACA. Gao et al (26) stated that CCDC181 could be a provisional prognostic biomarker of LACA. Moreover, Diaz-Lagares et al (27) indicated that there is a possibility for the cancer-specific methylation of CDO1 to enhance the diagnosis approach at the early stage and also achievements for patients. The methylation status of SYNE1 has also been valuable in estimating the sporadic lung cancer prognosis (28). However, no previous studies have reported on the association of TMEM201, Clorf43, ANKRD31, TBCD, SWAP70, AF131216.1, USP48, HPCAL1, GPR88 or VSTM2A and LACA. Hence, it is necessary to perform further studies on the methylation of these 10 genes and LACA.

In addition, GO annotations and KEGG pathways were established to provide detailed information regarding the molecular functions of the genes. The differentially methylated genes were mainly enriched in ECM, embryonic...
morphogenesis, cell adhesion and vascular system development; these are why high-risk tumors are more biologically aggressive. Generated through direct diffusion, lymphatic and vascular metastases are the common metastases in LACA. Once metastasis occurs, the prognosis of patients with LACA is poor. Increasing evidence has indicated that the ECM has a significant role in tumor occurrence and progression (29). Lim et al (30) developed a 29-gene ECM-associated signature to predict the prognosis of the patients at the early stage of NSCLC. Furthermore, it has been well established that the ECM is involved in regulating metastasis and invasion of lung cancer (31,32). Thus, it is necessary to perform in-depth research on these molecules to confirm these predictions, and simultaneously develop novel therapeutic interventions for LACA.

The present study does have certain limitations. First, in view of the LACA cohort exhibiting a reasonably high censored rate, this probably had an effect on the credibility of the Kaplan-Meier evaluation. Furthermore, as all of the samples analyzed in the present study were acquired from TCGA only, in-depth verification should be performed using independent datasets. In addition, the mechanistic role of each component of the prognostic signature remains to be investigated. Therefore, experimental research on cancer cell lines may provide significant information to further the understanding of their functional roles.

In conclusion, a 13 CpG-based prognostic signature for OS prediction in patients with LACA was obtained through comprehensively analyzing DNA methylation. The present results suggest that further research is required to validate the diagnostic ability of the novel diagnostic model in LACA. Retrospective and prospective studies may be performed to verify the prognostic utility of the CpG-based signature model.

Acknowledgements

Not applicable.

Funding

No funding was received.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Authors' contributions

RZ, HX, WM, ZD and MH participated in the study design, analysis and interpretation of data and drafting of the manuscript. MW performed the research and revised the paper. XG contributed to the acquisition of data, collection of relevant literature and drafting of the article together. 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


This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) License.