Genetics and Molecular Biology, 43, 3, e20190164 (2020)

DOI: https://doi.org/10.1590/1678-4685-GMB-2019-0164

Original Article
Genomics and Bioinformatics

Comprehensive analysis of gene expression and DNA methylation data identifies potential biomarkers and functional epigenetic modules for lung adenocarcinoma

XiaoCong Wang[1*], YanMei Li[1*], HuiHua Hu[2], FangZheng Zhou[1], Jie Chen[1] and DongSheng Zhang[1]

[1]*Hubei University of Medicine, Department of Oncology, Suizhou Hospital, Suizhou, Hubei, China.*
[2]*Hubei University of Medicine, Department of ICU, Suizhou Hospital, Suizhou, Hubei, China.*

Abstract

Lung cancer has one of the highest mortality rates of malignant neoplasms. Lung adenocarcinoma (LUAD) is one of the most common types of lung cancer. DNA methylation is more stable than gene expression and could be used as a biomarker for early tumor diagnosis. This study is aimed to screen potential DNA methylation signatures to facilitate the diagnosis and prognosis of LUAD and integrate gene expression and DNA methylation data of LUAD to identify functional epigenetic modules. We systematically integrated gene expression and DNA methylation data from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO), bioinformatic models and algorithms were implemented to identify signatures and functional modules for LUAD. Three promising diagnostic and five potential prognostic signatures for LUAD were screened by rigorous filtration, and our tumor-normal classifier and prognostic model were validated in two separate data sets. Additionally, we identified functional epigenetic modules in the TCGA LUAD dataset and GEO independent validation data set. Interestingly, the MUC1 module was identified in both datasets. The potential biomarkers for the diagnosis and prognosis of LUAD are expected to be further verified in clinical practice to aid in the diagnosis and treatment of LUAD.

Keywords: DNA methylation, biomarkers, lung adenocarcinoma, cancer diagnosis, prognosis.

Received: May 25, 2019; Accepted: January 30, 2020.

Introduction

Lung cancer has one of the highest incidence and mortality rates of neoplasms and can be classified into small cell lung cancer and non-small cell lung cancer (NSCLC), NSCLC consists of adenocarcinoma, squamous cell carcinoma, large cell carcinoma and other types (Travis, 2011). Among them, adenocarcinoma is the most common subtype (Liu *et al.*, 2000). Despite treatment with surgery followed by radiotherapy or chemotherapy, many patients still have poor clinical outcomes (Perez *et al.*, 2014; Ramnath *et al.*, 2013; Verdecchia *et al.*, 2007). Hence, early diagnosis and treatment are the key to reducing the mortality of lung cancer. To date, no widely used DNA methylation markers have been identified for the early diagnosis and prognosis of lung adenocarcinoma (LUAD).

According to a previous study, alterations in DNA methylation (DNAm) appear to mark preneoplastic normal cells that later transform and become enriched in tumors (Teschendorff *et al.*, 2016), which indicates that DNAm could act as biomarkers for the diagnosis of early cancer. Studies on DNAm in lung cancer strongly suggest that the analyses of DNA methylation profiles will be of great utility both for understanding the molecular basis of lung cancer development (Toyooka *et al.*, 2001, 2003, 2004; Virmani *et al.*, 2002), and for developing epigenetic signatures for lung cancer (Shi *et al.*, 2017; Walter *et al.*, 2018). Recently, some studies have developed new biomarker screening algorithms for cancer diagnosis and prognosis (Wei *et al.*, 2015; Hao *et al.*, 2017), and algorithms to integrate DNA methylation and gene expression data to better understand tumor biology (Jiao *et al.*, 2014).

In this study, the gene expression and DNA methylation data of LUAD patients were systematically integrated and analyzed. We screened three potential DNA methylation signatures for early diagnosis and five prognostic gene expression signatures for LUAD. The tumor-normal classification model and prognostic model were validated in two separate data-sets. In addition, we also identified functional epigenetic modules (FEMs) in The Cancer Genome Atlas (TCGA) LUAD data set and Gene Expression Omnibus (GEO) independent validation data set. The MUC1 module was identified in both data-sets. The potential biomarkers identified in this study are expected to be further validated and may aid decision for diagnosis and treatment of LUAD.

Send correspondence to DongSheng Zhang. Hubei University of Medicine, Department of Oncology, Suizhou Hospital, Suizhou, Hubei, China. E-mail: zhangdons@163.com. *These authors contributed equally to this work.

Material and Methods

Dataset

Gene expression data analysis

We downloaded the RNA-seqV2 sequencing data (level 3, normalized count) and the corresponding clinical data of LUAD patients from the UCSC Xena database (http://xena.ucsc.edu), including 553 gene expression samples (tumor: 495, normal: 58). The screening criteria for significantly differentially expressed genes were as follows: 1.5-fold change and corrected P-value < 0.05 (independent *t*-test and p-value was adjusted by Benjamini/Hochberg correction method.

DNA methylation data analysis

We downloaded the methylation data (level 3, Methylation 450K) and the corresponding clinical data of LUAD patients from Xena (http://xena.ucsc.edu), including 492 samples (tumor: 460, normal 32) for methylation data analysis. The methylation level of each gene was calculated by defining the average methylation level of probes within gene promoter area (TSS1500, 1stExon, TSS200, 5` UTR), as the methylation level of the gene (TSS 1500 and TSS 200 represents 1500 bp and 200 bp downstream from the transcription start site, 1stExon represents the first exon). The screening criteria for significantly differentially methylated genes were as follows: delta Beta > 0.2 and corrected P-value < 0.05(independent *t*-test plus Benjamini/Hochberg method).

Validation dataset

DNA methylation 450K chip data (series_matrix.txt) and gene expression data were downloaded from the NCBI-GEO database (http: // www. ncbi. nlm. Nih. Gov/geo/), including GSE39279 (Sandoval *et al.*, 2013), GSE52401 (Shi *et al.*, 2014), GSE66836 (Bjaanaes *et al.*, 2016), GSE75037 (Girard *et al.*, 2016), GSE56044 (Karlsson *et al.*, 2014), GSE50081 (Der *et al.*, 2014) and GSE42127 (Tang *et al.*, 2013).

Construction of diagnostic classifier

First, the recursive features elimination method was used to screen diagnostic probes from those 24,116 differentially methylated CpG sites. Then, probes were used to build the logistic regression function in the Python Sklearn package (version 0.19, http://scikit-learn.org/stable/index.html). The parameters were all default parameters, and the model was trained with the TCGA data. Cox proportional hazards model was built based on the screened genes and survival analysis was performed for all patients (Python lifeline 0.11.1 (http:/ /lifelines.readthedocs.io/en/latest/index.html)), and validated in a separate data-set.

Construction of prognostic model

We first preprocessed the GSE50081 and GSE42127 datasets downloaded from GEO. The expression of each gene was the average expression level of the corresponding probes. Then genes from the 469 differentially expressed genes were selected with cox regression analysis and log-rank test (the criterion was as follows: false discovery rate (FDR) of cox regression ≤ 0.05 and FDR of log-rank test ≤ 0.05). Then, the cox proportional hazards model was constructed with five selected genes, and survival analysis was performed for all patients.

Gene enrichment analysis

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed using a web-based gene annotation tool, DAVID (Huang *et al.*, 2009ab).

FEMs analysis

The FEM algorithm (Jiao *et al.*, 2014) is a functional supervised algorithm that uses a network of relations between genes (in our case a protein-protein interaction (PPI) network) to identify subnetworks where a significant number of genes are differentially methylated and differentially expressed. The association is measured at both the DNA methylation and gene expression levels. The algorithm thus consists of two main parts: (i) construction of an integrated network in which the associations with the phenotype are encapsulated as weights on the network edges, and (ii) inference of the FEMs as heavy subgraphs on this weighted network.

We first constructed a PPI network by integrating the InBio (Li *et al.*, 2017) and BioPlex (Huttlin *et al.*, 2015) databases. Then, we conducted FEM with the FEM algorithm by integrating the gene expression and DNA methylated data from both the TCGA and GEO datasets.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Results

Analysis of differentially expressed genes and differentially methylated genes

A total of 2469 differentially expressed genes were obtained, including 1457 upregulated and 1012 downregulated genes (Figure 1a). Principal component analysis (PCA) (Figure 1b) indicated that 1324 differentially expressed genes could effectively distinguish tumor samples from normal samples. A total of 24,116 differentially methylated CpGs retained, mapping to 981 genes, including 472 hypermethylated genes and 509 hypomethylated genes (Figure 1c). PCA indicated that the 981 differentially methylated genes could significantly separate the normal samples from the tumor samples (Figure 1d).

Diagnostic classifier effectively distinguishes tumor samples

After feature selection with recursive feature elimination (see Material and Methods), three probes (cg20568402,

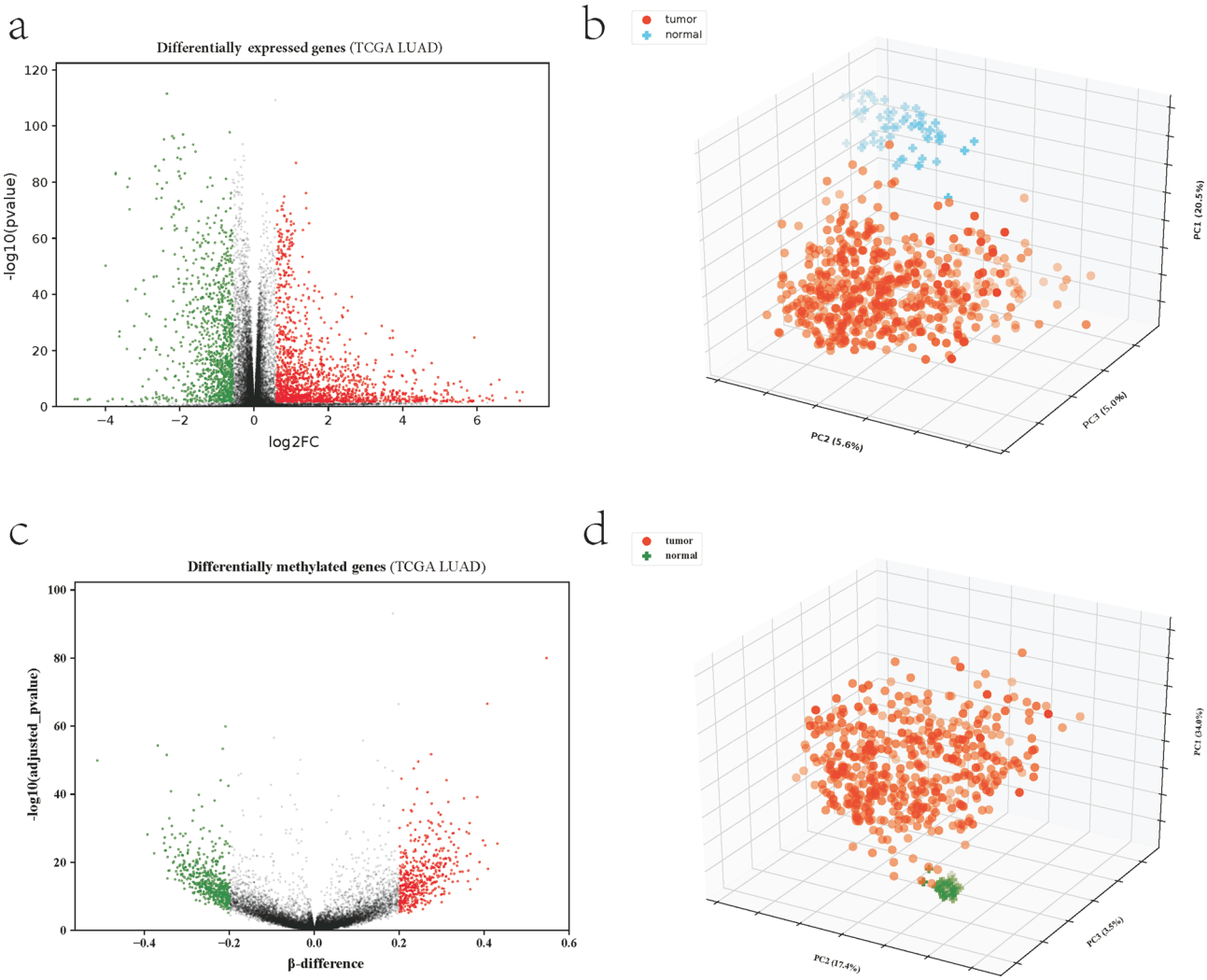

Figure 1 - Differential expression and differential methylation analyses. (a) Volcano plot of differentially expressed genes. (b) PCA of differentially expressed genes. (c) Volcano plot of differentially methylated genes. (d) PCA of differentially methylated genes.

cg11302791 and cg01302240, see Table 1) remained. The logistic regression model constructed with these three probes performed well in the TCGA training dataset (Figure 2a, area under the curve (AUC) > 0.99). The unsupervised cluster map of the DNA methylation level of these 3 probes could clearly distinguish tumor samples from normal samples (Figure 2b), indicating that the selected three probes can be used as potential biomarkers for the diagnosis of LUAD.

To further verify the repeatability of our feature selection method and classifier, we verified our model with GEO datasets (GSE39279, GSE52401 and GSE66836). As shown in Figure 2c and d, the results from the independent validation datasets are also very good (AUC > 0.92), which further indicated the reliability and accuracy of our method and model. In conclusion, our identified potential signatures may be helpful in distinguishing LUAD samples from normal samples, although further verification is needed.

Table 1 - Detailed information of three methylation markers (probes) for LUAD diagnosis.

Probe	GeneID	Gene Symbol	Relation To Island	Group
cg20568402	55208	DCUN1D2	OpenSea	Body
cg11302791	54984	PINX1	OpenSea	Body
cg01302240	5998	RGS3	OpenSea	TSS200; Body

Prognostic model significantly predicts the outcome of LUAD

We first screened five genes from all of the differentially expressed genes (see Material and Methods). A cox proportional hazards model was constructed with the five selected genes (COL6A6, WFIKKN2, PLA2G1B, UMODL1 and CNGA3, see Table S1) from the TCGA LUAD dataset. Of these genes, PLA2G1B was reported to be associated with smoking-related lung adenocarcinoma (Liu *et al.*, 2016), and UMODL1 may drive lung adenocarcinoma metastasis by involving the G-protein coupled receptor protein signaling pathway (Tan *et al.*, 2016). Then, survival analysis was performed for all patients. Finally, the tumor patients were divided into high-risk and low-risk groups, which were

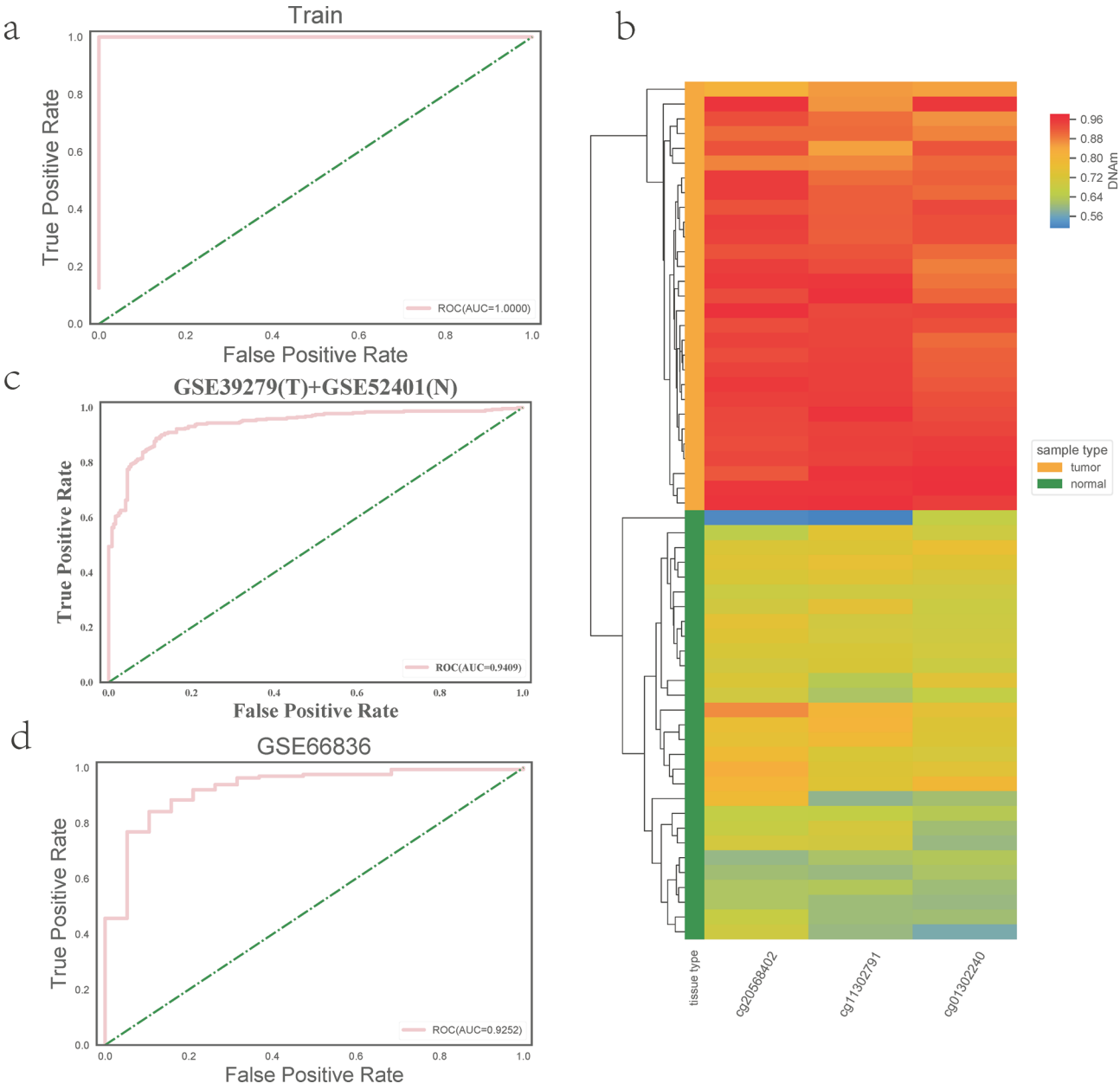

Figure 2 - Screening of methylation markers for lung adenocarcinoma and the construction and validation of the diagnostic classifier. (a) The ROC curve of the logistic regression model. (b) Unsupervised clustering map of the methylation profile for the three DNA methylation markers. (c, d) ROC curves in the independent validation datasets.

verified in the independent data-sets GSE50081 and GSE42127. A summary of the patients in the training and validation datasets for the five-gene-based classifier is listed in Table 2.

As shown in Figure 3, whether in the TCGA training dataset (Figure 3a) or independent validation datasets (Figure 3b-d), the five genes can significantly divide patients into high-risk and low-risk groups (P-value < 0.05), and the prognosis of patients in the high-risk group is significantly worse than that of patients in the low-risk group. In addition, the five potential prognostic markers screened from the LUAD can also significantly divide all NSCLC samples (including lung squamous carcinoma) into high- and low-risk groups (P-value < 0.01, Figure 3d), which further illustrates the repeatability of our classifier. We further performed survival analysis with regard to the five-gene-based classifier in subsets of patients with different clinical variables in the TCGA LUAD dataset. When stratified by clinical variables (sex, age, and pathologic stage), the five-gene-based classifier was still a statistically significant prognostic model (Figure S1).

The FEM MUC1 was identified in LUAD

We implemented the FEM algorithm (Jiao *et al.*, 2014) to integrate gene expression and DNA methylation data to perform FEM analysis, and four different functional mod-

Table 2 - Characteristics of patients by the five-gene-based classifier assessment set.

	TCGA LUAD (N=574)	GSE50081 (Adenocarcinoma N=127)	GSE42127 (Adenocarcinoma N=133)	GSE42127 (Squamous N=43)
Age (Years, mean ± std)	65.52 ± 9.91	68.73 ± 9.71	65.76 ± 10.29	68.11 ± 7.76
Gender				
MALE	238	62	65	18
FEMALE	272	65	68	25
Stage				
I	5	0	0	0
IA	132	36	32	10
IB	134	56	57	13
II	1	0	0	0
IIA	50	7	6	3
IIB	70	28	16	7
IIIA	73	0	7	6
IIIB	11	0	13	4
IV	26	0	1	0
Survival status				
Alive	317	76	90	22
Dead	181	51	43	21
Survival time (Months, mean ± std)	30.41 ± 30.04	42.39 ± 27.66	49.67 ± 31.70	53.53 ± 34.45

N indicates the number of tumor samples.

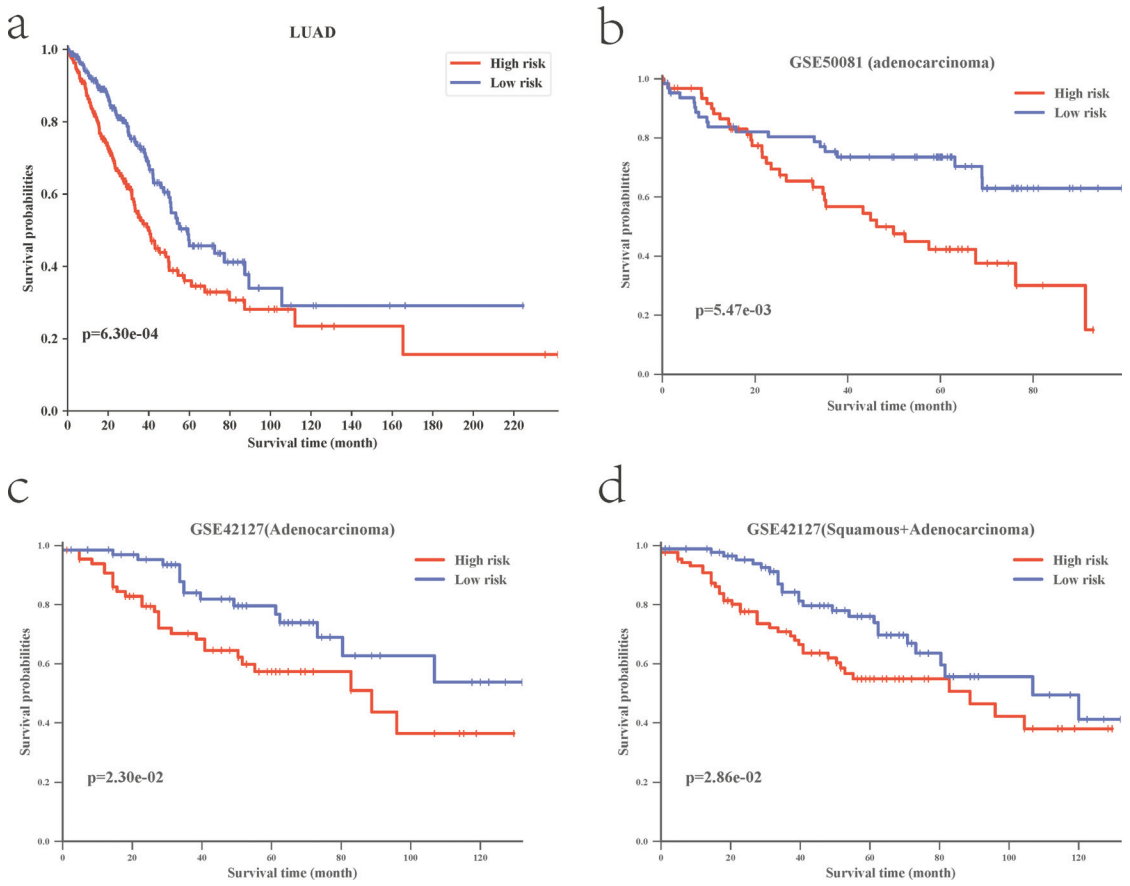

Figure 3 - Screening of the prognostic markers for lung adenocarcinoma and the construction of the prognostic classifier. (a) K-M curve in the TCGA training dataset. (b,c,d) K-M curve in the independent validation dataset.

ules were identified in the TCGA dataset: (MUC1) (Figure 4a), ADCY8, CAOLEC10 and WNT3A (Figure S2). Then, in the validation data set (GSE75037 and GSE56044), three modules were identified: MUC1 (Figure 4b), GSTMS and OTX1 (Figure S2), of which, the MUC1 module was identified in both of the datasets. The MUC1 modules identified in the two datasets were significantly overlapping (45 overlapping genes, P-value < 0.01, hypergeometric test) (Figure 4 and Table S2).

The genes in the MUC1 module were enriched in the ricin lectin domain structure and participated in biological processes such as O polysaccharide processing, protein glycosylation, keratin sulfate biosynthesis, mucin O polysaccharide biosynthesis, and sheath sugar lipid biosynthesis as well as metabolic signaling pathways (Figure 4). Pro-oncogenic mucin MUC1 was reported to contribute to smoking-induced lung cancers that are driven by inflammatory signals from macrophages (Xu *et al.*, 2014). A previous study showed that overexpression of MUC1 induces epithelial-mesenchymal transition and promotes the metastasis of lung cancer cells (Xue *et al.*, 2017). In addition, MUC1 was reported to be useful in predicting prognosis in NSCLC patients (Zhu *et al.*, 2014) and may contribute to the treatment of patients with NSCLC resistant to EGFR kinase inhibitors

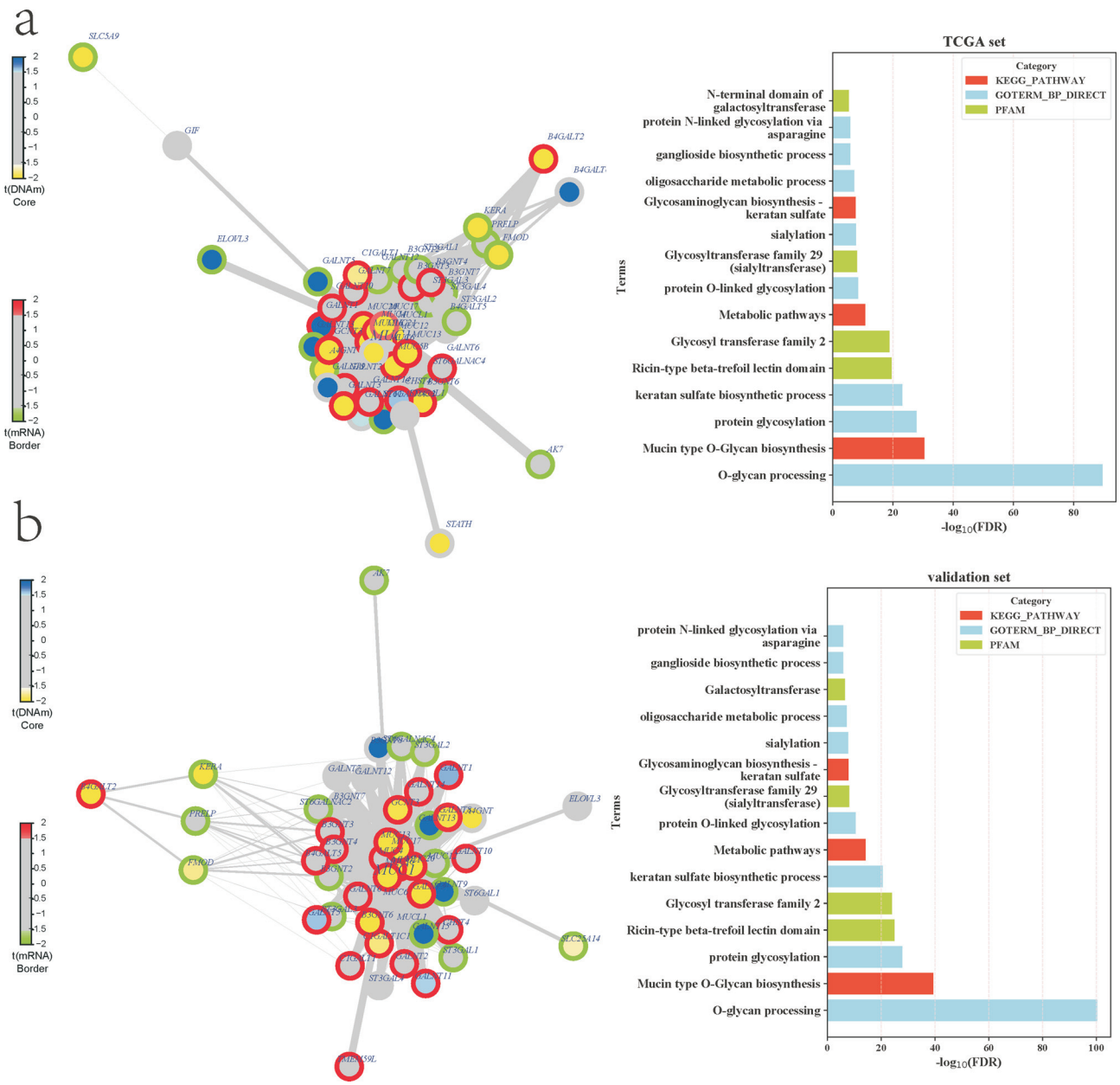

Figure 4 - Functional epigenetic modules of lung adenocarcinoma. (a) MUC1 module identified in TCGA (left) and enrichment analysis results (right). (b) MUC1 module identified in the GEO validation set (left) and enrichment analysis (right). The node color indicates the DNA methylation difference (blue indicates high methylation, and yellow indicates low methylation), and the edge color indicates differentially expressed genes (red represents genes with high expression level in tumors and green represents genes with low expression in tumors).

(Kharbanda *et al.*, 2014), indicating that our identified functional epigenetic module MUC1 may play an important role in the development and prognosis of LUAD.

Discussion

LUAD is one of the most common neoplasms, and the early diagnosis of LUAD has always been challenging. DNA methylation changes have been reported to occur early in carcinogenesis (Teschendorff *et al.*, 2016), and DNA methylation analysis seems to be a promising strategy in cancer diagnosis. This study used the data of the TCGA LUAD and GEO public datasets, performed an integrative analysis of gene expression and DNA methylation data, and selected three potential DNA methylation biomarkers for the diagnosis of LUAD. All three CpGs are differentially variable and differentially methylated CpGs (DVMCs) in the TCGA LUAD dataset (Figure S3). DVMC was defined by Teschendorff *et al.* (2016) and could be useful to identify field defects. In addition, five potential prognostic gene expression signatures selected from the differentially expressed genes could be used to predict the outcome of LUAD patients. We also performed stratification analysis, in which, different clinical variables (sex, age and stage) were evaluated separately and our prognostic model could also divide tumor patients into high-risk and low-risk groups with significantly different outcomes. Finally, we identified the functional epigenetic module MUC1, which plays a certain role in LUAD, in both the TCGA and GEO datasets. The potential DNA methylation biomarkers identified in this study may be used to design appropriate methylation targeted therapies.

Conflict of Interest

The authors report no conflicts of interest and have no relevant disclosures.

Author Contributions

DZ, X.W and YL designed the study and wrote the manuscript. XW and HH conducted the data analysis. FZ, JC and DZ revised and finalized the manuscript. All authors read and approved the final manuscript.

References

Bjaanaes MM, Fleischer T, Halvorsen AR, Daunay A, Busato F, Solberg S, Jorgensen L, Kure E, Edvardsen H, Borresen-Dale AL *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 10:330-43.

Der SD, Sykes J, Pintilie M, Zhu CQ, Strumpf D, Liu N, Jurisica I, Shepherd FA and Tsao MS (2014) Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including stage IA patients. J Thorac Oncol 9:59-64.

Girard L, Rodriguez-Canales J, Behrens C, Thompson DM, Botros IW, Tang H, Xie Y, Rekhtman N, Travis WD, Wistuba II *et al.* (2016) An expression signature as an aid to the histologic classification of non-small cell lung cancer. Clin Cancer Res 22:4880-4889.

Hao X, Luo H, Krawczyk M, Wei W, Wang W, Wang J, Flagg K, Hou J, Zhang H, Yi S *et al.* (2017) DNA methylation markers for diagnosis and prognosis of common cancers. Proc Natl Acad Sci U S A 114:7414-7419.

Huang da W, Sherman BT and Lempicki RA (2009a) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37:1-13.

Huang da W, Sherman BT and Lempicki RA (2009b) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44-57.

Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, Tam S, Zarraga G, Colby G, Baltier K *et al.* (2015) The BioPlex Network: A systematic exploration of the human interactome. Cell 162:425-40.

Jiao Y, Widschwendter M and Teschendorff AE (2014) A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. Bioinformatics 30:2360-6.

Karlsson A, Jonsson M, Lauss M, Brunnstrom H, Jonsson P, Borg A, Jonsson G, Ringner M, Planck M and Staaf J (2014) Genome-wide DNA methylation analysis of lung carcinoma reveals one neuroendocrine and four adenocarcinoma epitypes associated with patient outcome. Clin Cancer Res 20:6127-40.

Kharbanda A, Rajabi H, Jin C, Tchaicha J, Kikuchi E, Wong KK and Kufe D (2014) Targeting the oncogenic MUC1-C protein inhibits mutant EGFR-mediated signaling and survival in non-small cell lung cancer cells. Clin Cancer Res 20:5423-34.

Li T, Wernersson R, Hansen RB, Horn H, Mercer J, Slodkowicz G, Workman CT, Rigina O, Rapacki K, Staerfeldt HH *et al.* (2017) A scored human protein-protein interaction network to catalyze genomic interpretation. Nat Methods 14:61-64.

Liu NS, Spitz MR, Kemp BL, Cooksley C, Fossella FV, Lee JS, Hong WK and Khuri FR (2000) Adenocarcinoma of the lung in young patients: the M. D. Anderson experience. Cancer 88:1837-41.

Liu Y, Ni R, Zhang H, Miao L, Wang J, Jia W and Wang Y (2016) Identification of feature genes for smoking-related lung adenocarcinoma based on gene expression profile data. Onco Targets Ther 9:7397-7407.

Perez EA (2014) Highlights in breast cancer from the 2014 american society of clinical oncology annual meeting. Clin Adv Hematol Oncol 12:7-16.

Ramnath N, Dilling TJ, Harris LJ, Kim AW, Michaud GC, Balekian AA, Diekemper R, Detterbeck FC and Arenberg DA (2013) Treatment of stage III non-small cell lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest 143:e314S-e340S.

Sandoval J, Mendez-Gonzalez J, Nadal E, Chen G, Carmona FJ, Sayols S, Moran S, Heyn H, Vizoso M, Gomez A *et al.* (2013) A prognostic DNA methylation signature for stage I non-small-cell lung cancer. J Clin Oncol 31:4140-7.

Shi J, Marconett CN, Duan J, Hyland PL, Li P, Wang Z, Wheeler W, Zhou B, Campan M, Lee DS *et al.* (2014) Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. Nat Commun 5:3365.

Shi YX, Wang Y, Li X, Zhang W, Zhou HH, Yin JY and Liu ZQ (2017) Genome-wide DNA methylation profiling reveals no-

vel epigenetic signatures in squamous cell lung cancer. BMC Genomics 18:901.

Tan Q, Cui J, Huang J, Ding Z, Lin H, Niu X, Li Z, Wang G, Luo Q and Lu S (2016) Genomic alteration during metastasis of lung adenocarcinoma. Cell Physiol Biochem 38:469-86.

Tang H, Xiao G, Behrens C, Schiller J, Allen J, Chow CW, Suraokar M, Corvalan A, Mao J, White MA *et al.* (2013) A 12-gene set predicts survival benefits from adjuvant chemotherapy in non-small cell lung cancer patients. Clin Cancer Res 19:1577-86.

Teschendorff AE, Gao Y, Jones A, Ruebner M, Beckmann MW, Wachter DL, Fasching PA and Widschwendter M (2016) DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. Nat Commun 7:10478.

Toyooka S, Toyooka KO, Maruyama R, Virmani AK, Girard L, Miyajima K, Harada K, Ariyoshi Y, Takahashi T, Sugio K *et al.* (2001) DNA methylation profiles of lung tumors. Mol Cancer Ther 1:61-7.

Toyooka S, Maruyama R, Toyooka KO, McLerran D, Feng Z, Fukuyama Y, Virmani AK, Zochbauer-Muller S, Tsukuda K, Sugio K *et al.* (2003) Smoke exposure, histologic type and geography-related differences in the methylation profiles of non-small cell lung cancer. Int J Cancer 103:153-60.

Toyooka S, Suzuki M, Tsuda T, Toyooka KO, Maruyama R, Tsukuda K, Fukuyama Y, Iizasa T, Fujisawa T, Shimizu N *et al.* (2004) Dose effect of smoking on aberrant methylation in non-small cell lung cancers. Int J Cancer 110:462-4.

Travis WD (2011) Pathology of lung cancer. Clin Chest Med 32:669-92.

Verdecchia A, Francisci S, Brenner H, Gatta G, Micheli A, Mangone L, Kunkler I and EUROCARE-4 Working Group (2007) Recent cancer survival in Europe: a 2000-02 period analysis of EUROCARE-4 data. Lancet Oncol 8:784-96.

Virmani AK, Tsou JA, Siegmund KD, Shen LY, Long TI, Laird PW, Gazdar AF and Laird-Offringa IA (2002) Hierarchical clustering of lung cancer cell lines using DNA methylation markers. Cancer Epidemiol Biomarkers Prev 11:291-7.

Walter RFH, Rozynek P, Casjens S, Werner R, Mairinger FD, Speel EJM, Zur Hausen A, Meier S, Wohlschlaeger J, Theegarten D *et al.* (2018) Methylation of L1RE1, RARB, and RASSF1 function as possible biomarkers for the differential diagnosis of lung cancer. PLoS One 13:e0195716.

Wei JH, Haddad A, Wu KJ, Zhao HW, Kapur P, Zhang ZL, Zhao LY, Chen ZH, Zhou YY, Zhou JC *et al.* (2015) A CpG-methylation-based assay to predict survival in clear cell renal cell carcinoma. Nat Commun 6:8699.

Xu X, Padilla MT, Li B, Wells A, Kato K, Tellez C, Belinsky SA, Kim KCn and Lin Y (2014) MUC1 in macrophage: contributions to cigarette smoke-induced lung cancer. Cancer Res 74:460-70.

Xue M and Tao W (2017) Upregulation of MUC1 by its novel activator 14-3-3zeta promotes tumor invasion and indicates poor prognosis in lung adenocarcinoma. Oncol Rep 38:2637-2646.

Zhu WF, Li J, Yu LC, Wu Y, Tang XP, Hu YM and Chen YC (2014) Prognostic value of EpCAM/MUC1 mRNA-positive cells in non-small cell lung cancer patients. Tumour Biol 35:1211-9.

Supplementary Material

The following online material is available for this article:

Figure S1 - Kaplan Meier curves of the five-genes-based classifier between high-risk and low-risk groups stratified by clinical variables (sex, age, and pathologic stage).

Figure S2 - Other functional epigenetic modules identified in the TCGA dataset (ADCY8, CAOLEC10 and WNT3A) and the validation data set (GSTMS and OTX1).

Figure S3 - Illustration of differentially variable and differentially methylated CpGs (DVMCs) for the three diagnostic markers in the TCGA LUAD dataset.

Table S1 - Table of cox proportional hazards model result for the selected five genes used to construct the prognostic model.

Table S2 - Information for the members of MUC1 modules identified in the TCGA and independent validation datasets.

Associate Editor: Emmanuel Dias Neto