Open Access Full Text Article

ORIGINAL RESEARCH

Twenty-gene-based prognostic model predicts lung adenocarcinoma survival

Kai Zhao[1,2]
Zulei Li[2]
Hui Tian[1]

[1]Department of Thoracic Surgery, Qilu Hospital of Shandong University, Jinan, Shandong Province, China; [2]Department of Thoracic Surgery, Zibo Central Hospital, Zibo, Shandong Province, China

Introduction: Lung adenocarcinoma (LAC) accounts for more than a half of non-small cell lung cancer with high morbidity and mortality. Progression of treatment has not accelerated the improvement of its prognosis. Hence, it is an urgent need to develop novel biomarkers for its early diagnosis and treatment.

Materials and methods: In this study, we proposed to identify LAC survival-related genes through comprehensive analysis of large-scale gene expression profiles. LAC gene expression data sets were obtained from The Cancer Genome Atlas (TCGA). Identification of differentially expressed genes (DEGs) in LAC compared with adjacent normal lung tissues was first performed followed by univariate Cox regression analysis to obtain genes that are significantly associated with LAC survival (SurGenes). Then, we conducted sure independence screening (SIS) for SurGenes to identify more reliable genes and the prognostic signature for LAC survival prediction. Another two lung cancer data sets from TCGA and Gene Expression Omnibus (GEO) were used for the validation of prognostic signature.

Results: A total of 20 genes were obtained, which were significantly associated with the overall survival (OS) of LAC patients. The prognostic signature, a weighted linear combination of the 20 genes, could successfully separate LAC samples with high OS from those with low OS and had robust predictive performance for survival (training set: p-value $<2.2\times10^{-16}$; testing set: p-value $=2.04\times10^{-5}$, area under the curve (AUC) $=0.615$). Combined with GEO data set, we obtained four genes, that is, *FUT4*, *SLC25A42*, *IGFBP1*, and *KLHDC8B* that are found in both the prognostic signature and DEGs of LAC in GEO data set.

Discussion: The prognostic signature combined with multi-gene expression profiles provides a moderate OS prediction for LAC and should be helpful for appropriate treatment method selection.

Keywords: GEO, lung adenocarcinoma, SIS, survival, TCGA

Introduction

Lung cancer is both the most common cancer and the leading cause of cancer-related mortality worldwide.[1,2] The 5-year survival rate after diagnosis of lung cancer is 15.6%, which is lower than the survival rate for colon, breast, or prostate cancer.[3] Non-small cell lung cancer (NSCLC) accounts for 85% of all lung cancers, and lung adenocarcinoma (LAC) is the most diagnosed histological subtype of NSCLC.[4,5] The lung cancer is most commonly diagnosed at a late stage, which results in a poor patient survival rate. Although advanced progression had been made in treatment, including surgical, medical, and radiotherapeutic interventions, the long-term survival rate of patients diagnosed with primary LAC has not improved.[6,7] Previous studies have reported that LAC tends to grow and spread faster than other types of lung cancer and accounted for almost half of all lung cancer.[8] To date, the survival of LAC patients

Correspondence: Hui Tian
Department of Thoracic Surgery, Qilu Hospital of Shandong University, No. 107, Wenhuaxi Road, Lixia District, Jinan City, Shandong Province 250012, China
Tel +86 136 2215 7177
Email huitianqilu@163.com

remains dismal and 5-year survival rate is only ~10%.[9] The molecular mechanisms of LAC carcinogenesis and progression remain unclear.

The prognosis of lung cancer mainly depends on the probability of recurrence and metastasis. Recently, many efforts were made to identify the potential molecules that are the prognosis markers of lung cancer patients. Cui et al[10] reported a significant association between the prognosis of LAC patients and the expression of *fibulin-1*, a member of fibulin gene family, and showed that overexpression of *fibulin-1* in patients was positively associated with the overall survival (OS) of LAC patients. This result suggested that *fibulin-1* may possess a novel role in LAC prognosis and act as a potential molecular target for the detection and treatment of LAC.[10] *HoxB9* is a member of Hox family and plays an important role in cancer progression. Zhan et al found that enhanced expression of *HoxB9* correlates with T classification and more lymph node metastasis, leading to a poor prognosis of LAC patients.[11] A strata analysis of *CCND1* gene was performed in the LAC patients, the results of which showed that overexpression of *CCND1* was significantly associated with poor prognosis and low OS of LAC, suggesting that *CCND1* may act as a potential predictive marker or molecular target for the treatment and prognosis of LAC patients.[12] Although some molecular targets were identified and regarded as potential markers for the prognosis of LAC patients at present, the predicting accuracy of those molecules was insufficient. Identification of novel biomarkers would be one of the promising approaches for developing new diagnostic, therapeutic, and prognostic strategies of LAC.

Biomarkers possess some special features, such as measurable, dependable, inexpensive, and high sensitivity and specificity. These features confer the biomarkers as a potential tool in screening and recurrence detection of cancer.[13] Diverse techniques including biological information indexing and database provision have been developed for identifying novel biomarkers in many diseases, which are helpful for us to better understand the biological reactions including invasion, metastasis, proliferation, and prognosis. Bioinformatics is recently growing in the field of cancer biology, and several public databases, such as The Cancer Genome Atlas (TCGA), Gene Expression Omnibus (GEO), Embase, Surveillance, and Epidemiology and End Results, are open access for researchers.[14] In this study, the transcriptomes of more than 500 LAC samples obtained from TCGA were analyzed. Through the combination of univariate survival analysis and sure independence screening (SIS)-based dimensionality reduction, 20 genes were identified as the potential prognosis markers. Besides, we validated the predictive accuracy of the 20 genes in another data set from TCGA which contains more than 200 LAC samples. Our study should provide important clues for the LAC survival prediction and therapeutic strategy selection for different patients.

Materials and methods
Study population

The training set was downloaded from TCGA including 521 LAC samples diagnosed at 33–88 years of age and 488 (~93.67%) of them were classified as stage I ~ stage III. Seventeen samples were removed because of the lack of survival information. The clinicopathological features of patients are provided in Table S1. One hundred LAC patients had both tumor and adjacent normal tissues, which were used for the differential expression analysis. Another LAC data set consisting of 230 LAC samples from TCGA was used as validation set with 28 samples removed due to missing survival information. We downloaded an LAC gene expression data set from GEO with the accession number GSE85841 consisting of eight LAC and eight adjacent normal tissues for the confirmation of differential expression analysis in the training set.

Gene expression data analysis

Raw read counts of transcriptomic data from TCGA were normalized by quartile normalization method and logarithmic transformed to a normal distribution. DESeq2[15] bioconductor package was used for the identification of differentially expressed genes (DEGs) in LAC samples compared with adjacent normal tissues with the criteria of adjusted p-value <0.01 and absolute log2-based fold change >1. For gene expression data from GEO, we used GEO2R online tool to conduct differential expression analysis with the thresholds of adjusted p-value <0.01 and absolute log2-based fold change >1 as significantly differential expression.

Screening of LAC survival-related genes

We conducted a two-step method for the screening of LAC survival-related genes. First, univariate Cox regression analysis between expression values of DEGs and survival of samples in training set was performed, and the significant genes were abbreviated as SurvGenes. The relatively high number of SurvGenes in contrast to the small sample number might lead to overfitting. Second, we used SIS based on LASSO Cox penalized regression model to identify more important variables and construct a multi-gene-based prognostic signature for OS prediction. The two-step screening method should be more robust and was performed via SIS R package.

Statistical analysis

We used the multi-gene-based prognostic signature to calculate risk score (RS) for every sample and divided the samples into two groups according to the median RS. Kaplan–Meier survival curves were drawn and compared between the subgroups via log-rank tests. Receiver operating characteristic (ROC) curves were drawn for the predicted 3-year OS based on the nearest neighbor method, and area under the curve (AUC) was calculated. AUC p-values were obtained through bootstrap resampling.

All statistical analyses were performed through R 3.4.1. Two-sided p-value <0.05 was considered as statistically significant.

Results
Candidate genes

After preprocessing, we first conducted clustering analysis for samples in the training set based on their stages via the top two principal components. As a result, LAC samples of stage I ~ stage IV could not be separated from each other (Figure 1A). Hence, it is important to identify other diagnosis biomarkers for the assistance of personalized treatment.

Differential expression analysis was conducted based on the 100 LAC patients with both tumor and adjacent normal tissues in the training set. A total of 2,011 genes with adjusted p-value <0.01 and absolute log2-based fold change >1 were identified. We then used univariate Cox regression
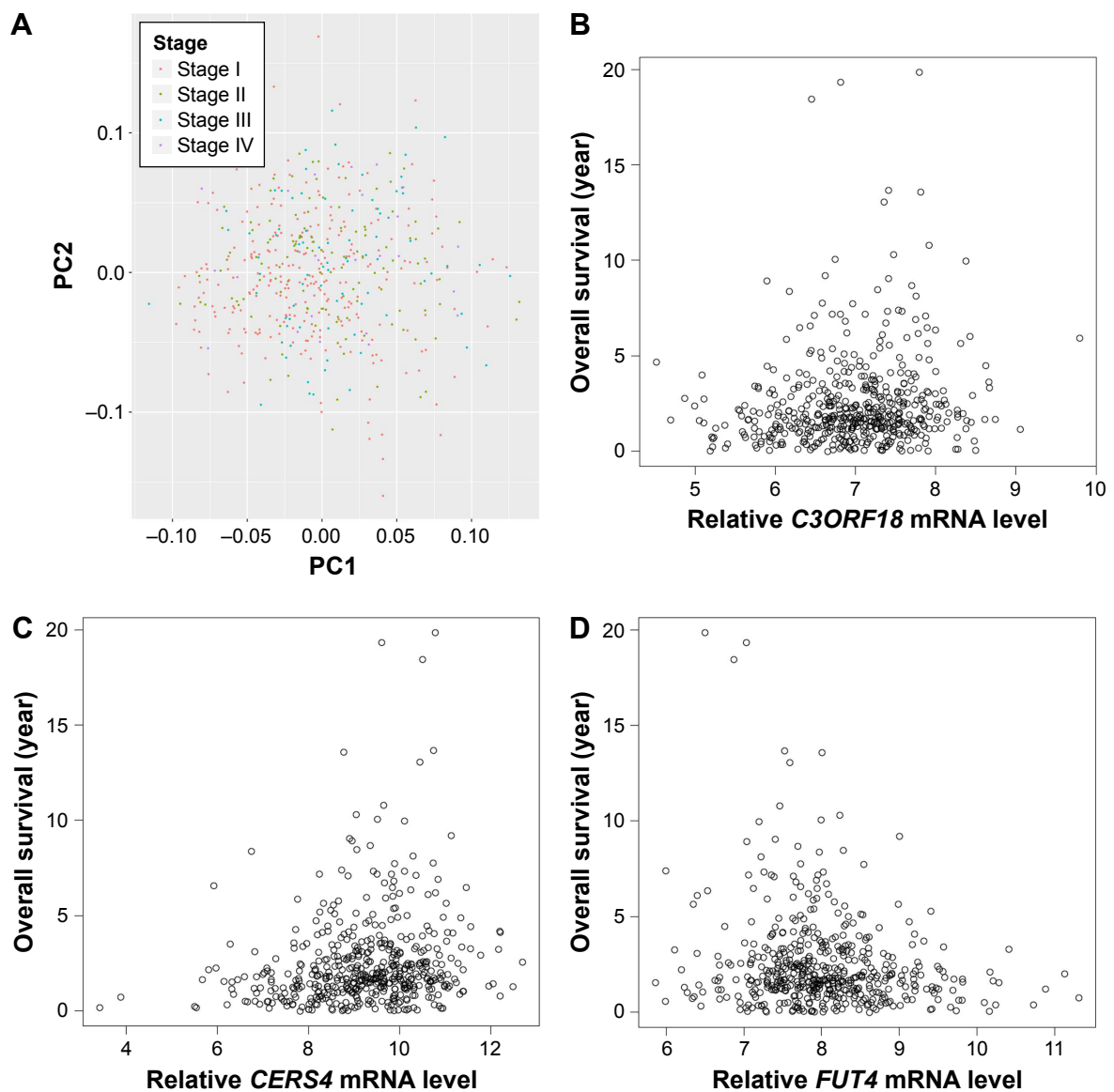

Figure 1 Principal component analysis of raw read count and correlation analysis between selected genes expression values and sample survival in training set. (**A**) Principal component analysis of raw read count of samples in different clinicopathological stages in training set. Red, blue, green, and purple dots represent stage I, stage II, stage III, and stage IV, respectively. (**B–D**) Spearman correlation between expression value of *C3ORF18*, *CERS4*, and *FUT4* and survival times of samples in training set. X-axis and Y-axis represent gene expression value and sample survival time, respectively.

analysis to evaluate associations between DEGs and OS in the training set and identified 1,425 SurvGenes with *p*-value <0.05. Figure 1B–D illustrates the Spearman correlation between the expression values of three of the 1425 genes, that is, *C3ORF18* (cor =0.177, *p*-value =6.30×10^{-5}), *CERS4* (cor =0.125, *p*-value =4.85×10^{-3}), and *FUT4* (cor =–0.178, *p*-value =6.063×10^{-5}), and OS in years. Through functional enrichment analysis in the Database for Annotation, Visualization and Integrated Discovery,[16] 1,425 SurvGenes were found to be significantly enriched in biological processes/pathways related to cell cycle, immune response (Figure 2).

SIS analysis was used to further identify robust gene combination. Twenty of the 1,425 candidate SurvGenes were obtained, including *CERS4, FUT4, C3ORF18, CYP17A1, ASPM, HJURP, LOC645166, DENND1C, SLC25A42, CCNA2, LDHA, IGFBP1, SLC2A1, DAAM2, RGS20, MFI2, LDLRAD3, KLHDC8B, CREG2,* and *SPATA6*. Table 1 shows their regression coefficients and forward selection Akaike information criterion.

Differential expression analysis of LAC data sets from GEO obtained a total of 1,282 DEGs in LAC compared with adjacent normal tissues. Intersecting those DEGs and the 20 candidate genes, we identified four overlaps, including

Figure 2 Functional enrichment analysis of SurvGenes via DAVID. (**A**) Clustering analysis of significantly enriched biological process (BP) terms. Nodes and edges are BP terms and their interactions, respectively. Larger node size indicates more contained genes and darker color indicates more significant, that is, smaller *p*-value. Thicker line indicates more overlaps between the two BP terms. (**B**) Significantly enriched pathways. X-axis is gene number contained in pathways. Darker color indicates more significant, that is, smaller *p*-value.

Table 1 The 20-gene signatures screened by forward selection and their regression coefficients

Gene	AIC	Coefficient
CERS4	494.2123	0.1375
FUT4	470.2554	−0.5066/−0.4321[a]
C3ORF18	469.7183	0.08425
CYP17A1	430.3143	0.2910
ASPM	407.2271	−0.4360
HJURP	403.3921	0.3888
LOC645166	383.6688	−0.1540
DENND1C	380.4144	0.3972
SLC25A42	362.051	−0.2404/0.07663[a]
CCNA2	361.8942	0.1753
LDHA	338.4287	−0.06559
IGFBP1	295.4821	−0.09503/−0.05506[a]
SLC2A1	294.8089	0.2518
DAAM2	282.4452	0.02896
RGS20	196.9195	−0.03843
MFI2	184.1124	−0.08505
LDLRAD3	182.1311	−0.2181
KLHDC8B	144.8857	0.1977/0.2943[a]
CREG2	128.026	0.05795
SPATA6	118.6571	0.10865

Note: [a]Regression coefficient in the four-gene prognostic signature.
Abbreviation: AIC, Akaike information criterion.

FUT4, *SLC25A42*, *IGFBP1*, and *KLHDC8B*. In the following analysis, we used both the prognostic signature of 20 candidate gene combination and the prognostic signature of four overlap combination for OS prediction. The prognostic score for each sample based on individualized values of the 20 genes: prognostic $score_{20}$ =$0.1375 \times CERS4 - 0.5066 \times FUT4 + 0.08425 \times C3ORF18 + 0.291 \times CYP17A1 - 0.436 \times ASPM + 0.3888 \times HJURP - 0.154 \times LOC645166 + 0.3972 \times DENND1C - 0.2404 \times SLC25A42 + 0.1753 \times CCNA2 - 0.06559 \times LDHA - 0.09503 \times IGFBP1 + 0.2518 \times SLC2A1 + 0.02896 \times DAAM2 - 0.03843 \times RGS20 - 0.08505 \times MFI2 - 0.2181 \times LDLRAD3 + 0.1977 \times KLHDC8B + 0.05795 \times CREG2 + 0.10865 \times SPATA6$; prognostic score based on the four genes: prognostic $score_4$ =$-0.4321 \times FUT4 + 0.07663 \times SLC25A42 - 0.05506 \times IGFBP1 + 0.2943 \times KLHDC8B$.

Prognostic signature for LAC patients

We divided the patients in testing sets into two subgroups based on the expression values of the 20 candidate genes (Figure 3A) and four overlaps (Figure 3B) through hierarchical clustering. Log-rank test was used for the testing of differences of OS between subgroups. As a result, both the 20 candidate genes and the four overlaps could significantly separate LAC patients with high OS from those with low OS as shown in Figure 3C (*p*-value =5.35×10^{-10}) and Figure 3D (*p*-value =4.44×10^{-4}), respectively. Furthermore,

we also divided the patients in testing set into two subgroups according to their prognostic $score_{20}$ and prognostic $score_4$ based on their median values and tested the significance of differences of OS between subgroups through log-rank test. As shown in Figure 4A and B, both prognostic $score_{20}$ (*p*-value =3.66×10^{-9}) and prognostic $score_4$ (*p*-value =2.04×10^{-5}) could robustly predict LAC survival. To evaluate the sensitivity of the two prognostic scores, we predicted OS through the nearest neighbor method for ROC curves of censored survival value. As a result, they could effectively predict 3-year OS (AUC of prognostic $score_{20}$ =0.615, AUC of prognostic $score_4$ =0.5731; *p*-value <0.05). The ROC curves for the expression of 20 genes were shown in Figure 4C.

Discussion

Lung cancer is the leading cause of cancer-related death in men and women worldwide.[17] LAC was a main type of lung cancer and accounts for more than half the incidence of morbidity and mortality.[8] A recent study by TCGA revealed potential novel drivers such as NF1, MET, ERBB2, and RIT1 which occur in 13% of cases and are enriched in samples lacking an activated oncogene.[18] Accumulated evidence suggested that the abnormally expressed genes in lung cancer, including LAC, were strongly associated with the prognosis and could be regarded as the potential prognostic factors, such as *Fibulin-1*, *HoxB9*, *TFEB*, and *BCL2*.[10,11,19,20] The predictive accuracy of those molecules was deemed insufficient because of no consideration on the condition of simultaneous changes in multiple genes. Several previous studies have focused on the differential expression or prognosis of gene signatures for their potential in LAC diagnosis and treatment.[21–23] Combination of those two strategies was applied for the screening of LAC biomarkers, which should be more reliable. Here, we used high-throughput method of searching for genetic differences of multiple genes associated with LAC prognosis and conducted the comprehensive analysis to obtain more reliable genes and the prognostic signature for LAC survival prediction.

In this study, 521 LAC samples and 504 transcriptomes of them were analyzed as training set to identify survival-related biomarkers. Another data set consisting of 230 LAC samples was used as testing set to validate biomarker reliability. A microarray expression data set which included eight LAC samples and eight adjacent tissue samples, the DEGs in the LAC samples compared to adjacent tissues were identified. After univariate survival analysis followed by SIS-based dimensional reduction process,
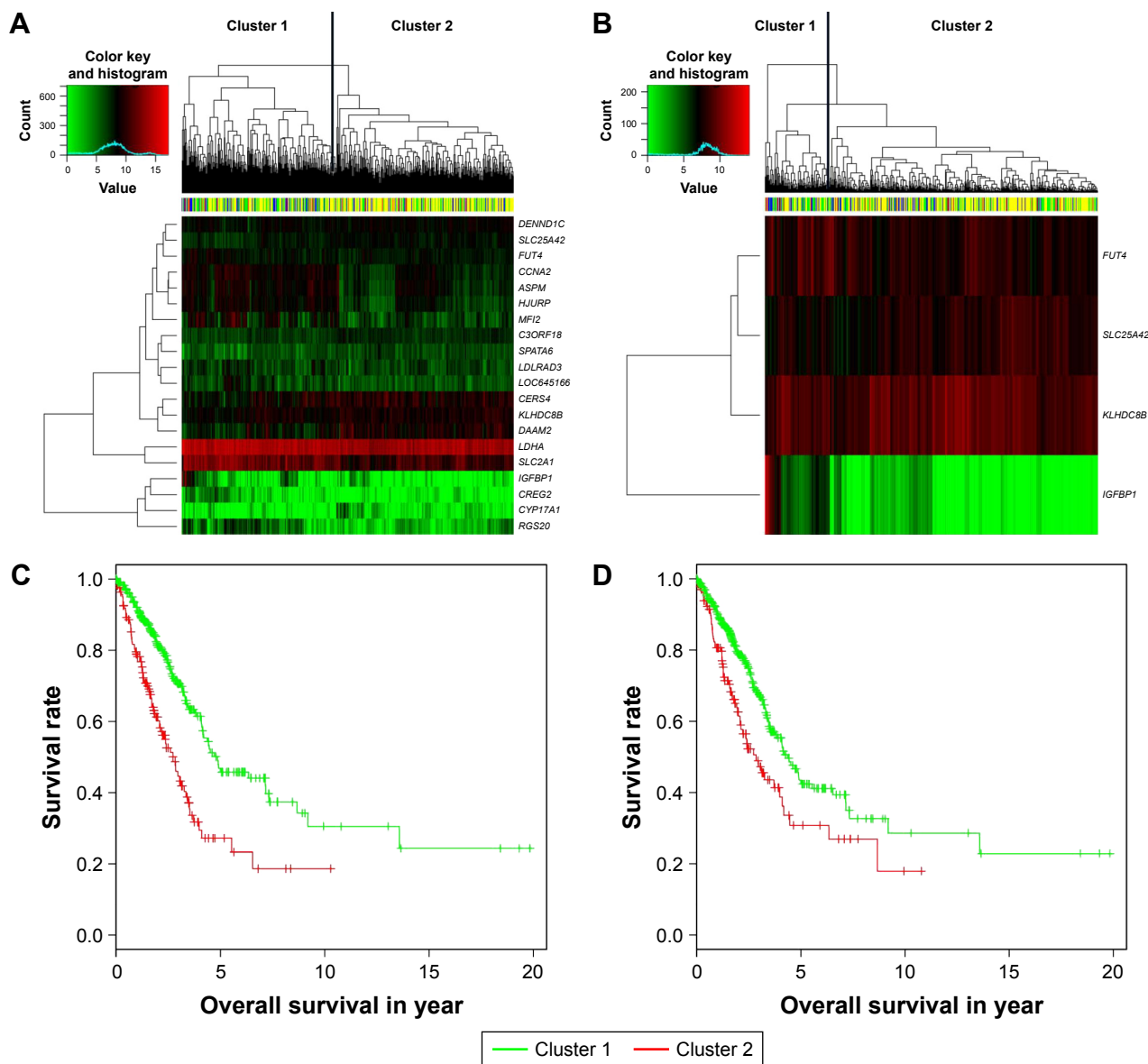
Figure 3 Clustering and Kaplan–Meier analysis in testing set. (**A** and **B**) The hierarchical clustering of samples in testing set based on the 20 sure independent screening (SIS)-resulted genes and four overlaps. Color bar at the top of the heatmap represents stages. (**C** and **D**) The Kaplan–Meier plot of different clusters obtained based on prognostic signature of 20 and four genes, respectively.

20 survival-related genes were retained. We identified four genes that exist both in the 20 genes and DEGs of GEO data sets, which might be more reliable biomarkers for LAC diagnosis and treatment.

Based on the tumor grading, we performed sample clustering by the principal component analysis in the training data set. Figure 1A showed that all the samples with different tumor stage were clustered together, suggesting an inaccurate grading standard of tumor, which might induce inaccurate prediction of cancer prognosis. Therefore, the construction-of-molecular model using the survival-related genes is needed.

In our research, univariate survival analysis of the DEGs in LAC samples of the training data set identified *C3ORF18*, *FUT4*, and *CERS4* as the top three most significant genes that are associated with LAC survival. *FUT4* is one of key enzymes to catalyze the α1,3-fucosylation of tumor sugar antigen Lewis Y (LeY), which is a specific tumor-associated sugar antigen (TASA).[24] Overexpressed *FUT4* was observed in the breast cancer patients, and it could be served as a novel biomarker in the diagnosis and prognosis of breast cancer.[25] *FUT4* has been reported to be closely correlated to tumor proliferation, apoptosis, metastasis, and epithelial–mesenchymal transition.[26] Ceramide synthases
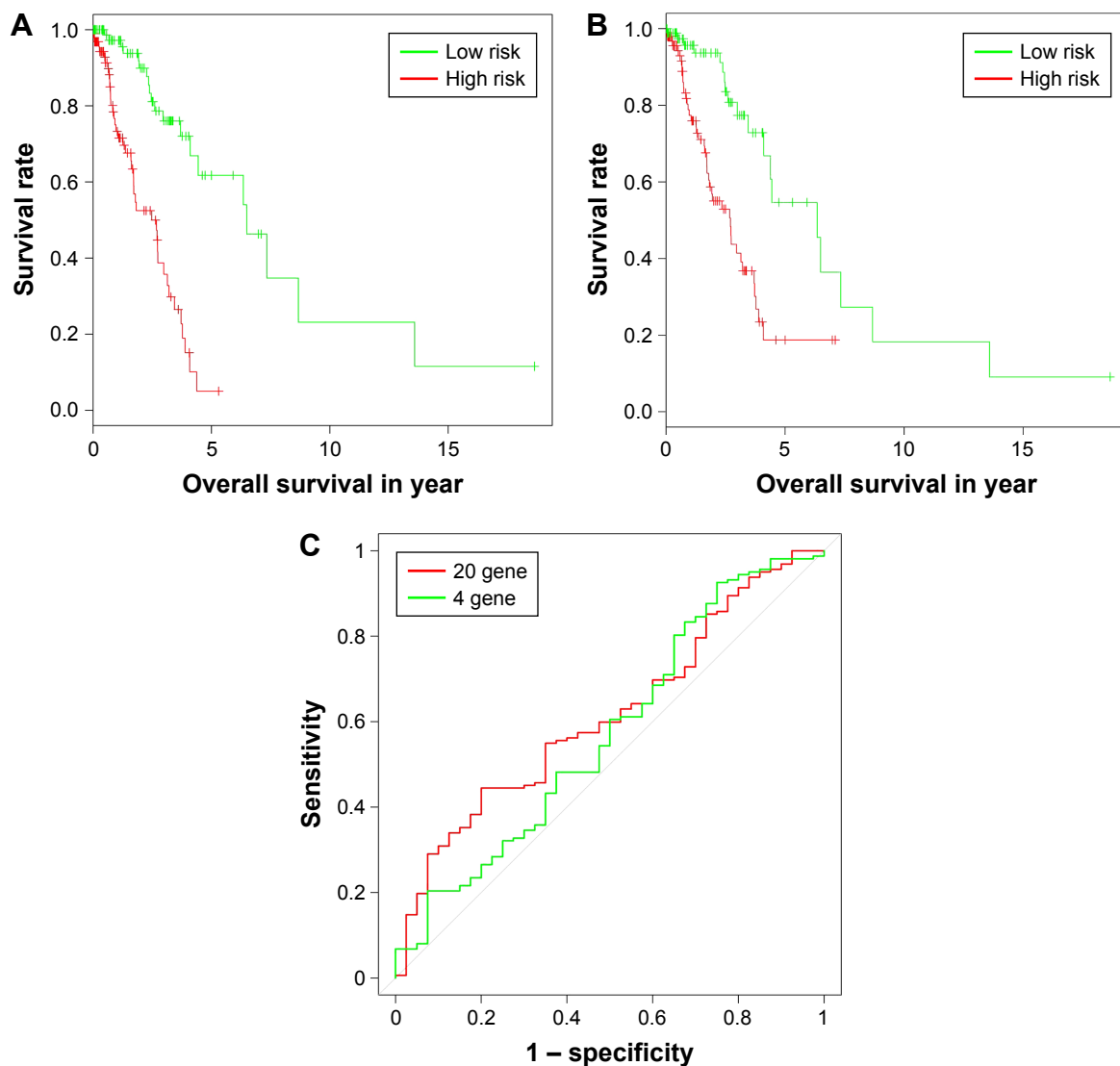
Figure 4 Sensitivity analysis for prognostic score$_{20}$ and prognostic score$_4$ in testing set. (**A** and **B**) The Kaplan–Meier plot of samples in testing set with higher and lower prognostic score$_{20}$ and prognostic score$_4$, respectively. (**C**) Receiver operating characteristic curves for the expression of 20 genes and four genes.

(CerSs) are important enzymes that play a central role in the sphingolipid pathway.[27] CerSs have been implicated in cancer biology, especially in apoptosis, through the action of ceramide.[28] Wegner et al reported that upregulation of *CERS4* in breast cancer cells is important for cell proliferation and tumor development.[29] In this study, expression values of *C3ORF18* and *CERS4* were positively correlated with the LAC OS, and expression value of *FUT4* shows a negative correlation with the OS in training data set (Figure 1B–D), which were consistent with the result of regression analysis (Table 1). These results suggested that the expression levels of *C3ORF18*, *FUT4*, and *CERS4* were strongly associated with the OS of LAC patients.

Except for *C3ORF18*, *FUT4*, and *CERS4*, many invasion-, metastasis-, and prognosis-related genes were included in

the 20 genes, such as *CYP17A1*, *ASPM*, *HJURP*, and so on. *CYP17A1* is one of the key enzymes involved in estrogen biosynthesis and was upregulated in various cancers.[30,31] Abnormal expression of *CYP17A1* might be used as a post-diagnostic marker and for choosing the right treatment method.[32] *ASPM* is an essential factor for normal mitotic spindle function in embryonic neuroblasts.[33] *ASPM* was overexpressed in many tumor cells and played an important role in malignant transformation and tumor progression.[34] Histone chaperone holliday junction recognition protein (*HJURP*) is responsible for depositing the histone variant centromere protein A at the centromere and is upregulated in lung and breast cancer.[35,36] Montes de Oca et al reported that the abnormal expression of *HJURP* could discriminate the favorable and unfavorable outcomes within the luminal A

subtype of breast carcinoma and could be an independent prognostic marker for luminal A patients.[37]

In recent years, some prognostic models based on cancer-related genes were constructed, and the predictive performances of these models were validated in different cancer types. Choi et al investigated the genetic variation in immune-related genes in breast cancer of Korean women, and the results showed that the genetic polymorphisms in immune-related genes, including *IL-6R*, *IL-8*, *IL-10RB*, *IL-12A*, and *IL-12B*, had significant relevance to breast cancer prognosis among Korean women.[38] A multivariable predictive model for relapse-free survival of neoadjuvant-resistant rectal cancer patients was constructed, which consists of gender, pT stage, N status, *CXCR4*, and *CXCR7*. Higher expression of *CXCR4* was significantly correlated with a shorter relapse-free survival and cancer-specific survival, and *CXCR4* could be regarded as a new therapeutic target in rectal cancer.[39] In this study, the prognostic model constructed with the 20 genes for the LAC was evaluated and compared with the model constructed with the four overlapping genes in the testing data set. Both of the two models could robustly predict LAC OS and the 20-gene-based model with relatively higher prediction accuracy than the four-gene-based model.

In this study, we integrated differential expression and regression analysis for LAC data sets from TCGA and GEO. Two OS prognostic models, that is, the 20-gene-based and the four-gene-based model, were obtained, which should be helpful for LAC early diagnosis and personalized treatment.

Disclosure
The authors report no conflicts of interest in this work.

References
1. Wu X, Dong L, Zhang R, Ying K, Shen H. Transgelin overexpression in lung adenocarcinoma is associated with tumor progression. *Int J Mol Med*. 2014;34:585–591.
2. Woods LM, Coleman MP, Lawrence G, Rashbass J, Berrino F, Rachet B. Evidence against the proposition that "UK cancer survival statistics are misleading": simulation study with National Cancer Registry data. *BMJ*. 2011;342:d3399.
3. Nanavaty P, Alvarez MS, Alberts WM. Lung cancer screening: advantages, controversies, and applications. *Cancer Control*. 2014;21:9–14.
4. Shao L, Zhang B, He C, et al. Efficacy and safety of icotinib in Chinese patients with advanced non-small cell lung cancer after failure of chemotherapy. *Chin Med J (Engl)*. 2014;127:266–271.
5. Travis WD. Pathology of lung cancer. *Clin Chest Med*. 2011;32:669–692.
6. Field JK, Raji OY. The potential for using risk models in future lung cancer screening trials. *F1000 Med Rep*. 2010;2.pii:38.
7. Ganti AK, Mulshine JL. Lung cancer screening: panacea or pipe dream? *Ann Oncol*. 2005;16(Suppl 2):ii215–ii219.
8. Travis WD, Brambilla E, Noguchi M, et al. International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society: international multidisciplinary classification of lung adenocarcinoma: executive summary. *Proc Am Thorac Soc*. 2011;8:381–385.
9. Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2008. *CA Cancer J Clin*. 2008;58:71–96.
10. Cui Y, Liu J, Yin HB, Liu YF, Liu JH. Fibulin-1 functions as a prognostic factor in lung adenocarcinoma. *Jpn J Clin Oncol*. 2015;45:854–859.
11. Zhan J, Wang P, Niu M, et al. High expression of transcriptional factor HoxB9 predicts poor prognosis in patients with lung adenocarcinoma. *Histopathology*. 2015;66:955–965.
12. Xu P, Zhao M, Liu Z, et al. Elevated nuclear CCND1 expression confers an unfavorable prognosis for early stage lung adenocarcinoma patients. *Int J Clin Exp Pathol*. 2015;8:15887–15894.
13. Malati T. Tumour markers: an overview. *Indian J Clin Biochem*. 2007;22:17–31.
14. Lee E, Moon A. Identification of biomarkers for breast cancer using databases. *J Cancer Prev*. 2016;21:235–242.
15. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15:550.
16. Sherman BT, Huang da W, Tan Q, et al. DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC Bioinformatics*. 2007;8:426.
17. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. *CA Cancer J Clin*. 2011;61:69–90.
18. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014;511:543–550.
19. Giatromanolaki A, Kalamida D, Sivridis E, et al. Increased expression of transcription factor EB (TFEB) is associated with autophagy, migratory phenotype and poor prognosis in non-small cell lung cancer. *Lung Cancer*. 2015;90:98–105.
20. Xu P, Liu L, Wang J, et al. Genetic variation in BCL2 3′-UTR was associated with lung cancer risk and prognosis in male Chinese population. *PLoS One*. 2013;8:e72197.
21. Lim SM, Kim EY, Kim HR, et al. Genomic profiling of lung adenocarcinoma patients reveals therapeutic targets and confers clinical benefit when standard molecular testing is negative. *Oncotarget*. 2016;7:24172.
22. Finn RS, Dering J, Conklin D, et al. PD 0332991, a selective cyclin D kinase 4/6 inhibitor, preferentially inhibits proliferation of luminal estrogen receptor-positive human breast cancer cell lines in vitro. *Breast Cancer Res*. 2009;11:R77.
23. Logan JE, Mostofizadeh N, Desai AJ, et al. PD-0332991, a potent and selective inhibitor of cyclin-dependent kinase 4/6, demonstrates inhibition of proliferation in renal cell carcinoma at nanomolar concentrations and molecular markers predict for sensitivity. *Anticancer Res*. 2013;33:2997–3004.
24. Taniguchi A, Suga R, Matsumoto K. Expression and transcriptional regulation of the human alpha1, 3-fucosyltransferase 4 (FUT4) gene in myeloid and colon adenocarcinoma cell lines. *Biochem Biophys Res Commun*. 2000;273:370–376.
25. Yan X, Lin Y, Liu S, Aziz F, Yan Q. Fucosyltransferase IV (FUT4) as an effective biomarker for the diagnosis of breast cancer. *Biomed Pharmacother*. 2015;70:299–304.
26. Tian L, Shen D, Li X, et al. Ginsenoside Rg3 inhibits epithelial-mesenchymal transition (EMT) and invasion of lung cancer by down-regulating FUT4. *Oncotarget*. 2016;7:1619.
27. Kageyama-Yahara N, Riezman H. Transmembrane topology of ceramide synthase in yeast. *Biochem J*. 2006;398:585–593.
28. Mojakgomo R, Mbita Z, Dlamini Z. Linking the ceramide synthases (CerSs) 4 and 5 with apoptosis, endometrial and colon cancers. *Exp Mol Pathol*. 2015;98:585–592.
29. Wegner MS, Wanger RA, Oertel S, et al. Ceramide synthases CerS4 and CerS5 are upregulated by 17beta-estradiol and GPER1 via AP-1 in human breast cancer cells. *Biochem Pharmacol*. 2014;92:577–589.
30. Bulun SE, Simpson ER. Aromatase expression in women's cancers. *Adv Exp Med Biol*. 2008;630:112–132.
31. Carruba G. Aromatase in nontumoral and malignant human liver tissues and cells. *Ann N Y Acad Sci*. 2009;1155:187–193.

32. Tuzuner MB, Ozturk T, Eronat AP, et al. Evaluation of local CYP17A1 and CYP19A1 expression levels as prognostic factors in postmenopausal invasive ductal breast cancer cases. *Biochem Genet*. 2016;54: 784–802.

33. Bond J, Roberts E, Mochida GH, et al. ASPM is a major determinant of cerebral cortical size. *Nat Genet*. 2002;32:316–320.

34. Lin SY, Pan HW, Liu SH, et al. ASPM is a novel marker for vascular invasion, early recurrence, and poor prognosis of hepatocellular carcinoma. *Clin Cancer Res*. 2008;14:4814–4820.

35. Hu Z, Huang G, Sadanandam A, et al. The expression level of HJURP has an independent prognostic impact and predicts the sensitivity to radiotherapy in breast cancer. *Breast Cancer Res*. 2010;12:R18.

36. Kato T, Sato N, Hayama S, et al. Activation of Holliday junction recognizing protein involved in the chromosomal stability and immortality of cancer cells. *Cancer Res*. 2007;67:8544–8553.

37. Montes de Oca R, Gurard-Levin ZA, Berger F, et al. The histone chaperone HJURP is a new independent prognostic marker for luminal A breast carcinoma. *Mol Oncol*. 2015;9:657–674.

38. Choi J, Song N, Han S, et al. The associations between immunity-related genes and breast cancer prognosis in Korean women. *PLoS One*. 2014;9:e103593.

39. D'Alterio C, Avallone A, Tatangelo F, et al. A prognostic model comprising pT stage, N status, and the chemokine receptors CXCR4 and CXCR7 powerfully predicts outcome in neoadjuvant resistant rectal cancer patients. *Int J Cancer*. 2014;135:379–390.

Supplementary material

Table S1 The clinicopathological features of patients

Clinicopathological features	Number
Gender	
Male	241
Female	280
Ethnicity category	
Hispanic or Latino	7
Not Hispanic or Latino	389
Not available	125
Neoplasm Disease Stage American Joint Committee on Cancer Code	
Stage I	279
Stage II	124
Stage III	85
Stage IV	25
Discrepancy	7
Not available	1
Patient's vital status	
Alive	333
Dead	188
Overall survival (months)	
>36 months	136
≤36 months	376
Not available	9