Cancer Management and Research

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

A network-based signature to predict the survival of non-smoking lung adenocarcinoma

This article was published in the following Dove Press journal:
Cancer Management and Research

Qixing Mao,[1–4,*]
Louqian Zhang,[1–3,*]
Yi Zhang,[1,*]
Gaochao Dong,[1,3]
Yao Yang,[4]
Wenjie Xia,[1–4]
Bing Chen,[1–3]
Weidong Ma,[1–3]
Jianzhong Hu,[4]
Feng Jiang,[1,3]
Lin Xu[1,3]

[1]Department of Thoracic Surgery, Jiangsu Cancer Hospital, Jiangsu Institute of Cancer Research, Nanjing Medical University Affiliated Cancer Hospital, Nanjing, China; [2]The Fourth Clinical College of Nanjing Medical University, Nanjing, China; [3]Jiangsu Key Laboratory of Molecular and Translational Cancer Research, Nanjing Medical University Affiliated Cancer Hospital, Cancer Institute of Jiangsu Province, Nanjing, China; [4]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA

*These authors contributed equally to this work

Correspondence: Lin Xu
Department of Thoracic Surgery, Jiangsu Cancer Hospital, No. 42 Baiziting Road, Nanjing, 210009, China
Tel +86 258 328 4700
Fax +86 258 364 1062
Email xulin_83@hotmail.com

Feng Jiang
Department of Thoracic Surgery, Jiangsu Cancer Hospital, No. 42 Baiziting Road, Nanjing, 210009, China
Tel +86 258 328 3408
Email zengnljf@hotmail.com

Background: A substantial increase in the number of non-smoking lung adenocarcinoma (LAC) patients has been drawing extensive attention in the past decade. However, effective biomarkers, which could guide the precise treatment, are still limited for identifying high-risk patients. Here, we provide a network-based signature to predict the survival of non-smoking LAC.

Materials and methods: Gene expression profiles were downloaded from The Cancer Genome Atlas and Gene Expression Omnibus. Significant gene co-expression networks and hub genes were identified by Weighted Gene Co-expression Network Analysis. Potential mechanisms and pathways of co-expression networks were analyzed by Gene Ontology. The predictive signature was constructed by penalized Cox regression analysis and tested in two independent datasets.

Results: Two distinct co-expression modules were significantly correlated with the non-smoking status across 4 Gene Expression Omnibus datasets. Gene Ontology revealed that nuclear division and cell cycle pathways were main mechanisms of the blue module and that genes in the turquoise module were involved in lymphocyte activation and cell adhesion pathways. Seventeen genes were selected from hub genes at an optimal lambda value and built the prognostic signature. The prognostic signature distinguished the survival of non-smoking LAC (training: hazard ratio [HR]=3.696, 95% CI: 2.025–6.748, $P<0.001$; testing: HR=2.9, 95% CI: 1.322–6.789, $P=0.006$; HR=2.78, 95% CI: 1.658–6.654, $P=0.022$) and had moderate predictive abilities in the training and validation datasets.

Conclusion: The prognostic signature is a promising predictor of non-smoking LAC patients, which might benefit clinical practice and precision therapeutic management.

Keywords: weighted gene co-expression network analysis, WGCNA, lung adenocarcinoma, LAC, co-expressing, prognostic signature

Introduction

Lung adenocarcinoma (LAC) is the main histological type of non-small cell lung carcinoma (NSCLC), making up 40% of lung cancer patients. There is a growing concern about the increasing number of the non-smoking LAC in the past decade. Previous evidence shows that the non-smoking LAC patients are more likely to be young, women, and carrying epidermal growth factor receptor mutations, which are different from smoking LAC patients.[1,2] In addition, etiology and biological behaviors of non-smoking LAC are remarkably different from smoking LAC, which make them different in therapeutic responses and prognosis.[3,4] Reliable signatures can accurately estimate the prognosis of disease and have tremendous significance in therapeutic management. Increasing number of studies are proposing gene expression-based signatures for survival stratification of NSCLC patients.[5] However, predictive signatures

submit your manuscript | www.dovepress.com
Dovepress
http://dx.doi.org/10.2147/CMAR.S163918

for non-smoking LAC have not been well addressed. Therefore, promising prognostic signatures for the non-smoking LAC are needed to stratify patients and predict the outcomes.

Weighted gene co-expression network analysis (WGCNA) is a feasible approach, which handles multi-dimensional expression data to construct sub-network atlas related to clinical features. Clarke et al reported a large-scale co-expression analysis in breast cancer from 13 microarrays.[6] Sun et al identified several hub genes related to the stage and grade of ovarian cancer.[7] Another study uncovered biomarkers for the prognosis of stage II and III colon cancer by the network-based approach.[8] These studies reinforced the effect of WGCNA as a method for discovering useful and reliable cancer biomarkers.

In this study, we developed a prognostic signature to predict the survival of non-smoking LAC patients by WGCNA, which could help better understand the potential mechanisms and aid in optimizing treatment.

Materials and methods

Microarray analysis

The expression profiles of LAC were retrieved from the Gene Expression Omnibus (GEO) data repository (http://www.ncbi.nlm.nih.gov/geo/) and The Cancer Genome Atlas (TCGA). The datasets that contained non-smoking status and normal tissues were included into subsequent analysis. To avoid bias, we only included datasets with large sample sizes. Gene expression profile was assessed by the HG-U133A microarrays from Affymetrix Human Genome. The background correction was performed by robust multi-array (RMA) method to remove the noise signals. Quantile standard method was used to normalize the data. Gene annotation was conducted using the Bioconductor annotation package hgu133plus2.db. These steps were performed using the R package "affy".[9] The mean expression was calculated as the final expression of genes measured by multiple probes. Differential expression analysis of the microarray was conducted by the R package "limma" and "EdgeR".[10,11] The batch effects between different datasets were adjusted by R package "limma". The threshold of different expression genes (DEGs) were defined as fold change (FC) over 2 with adjusted *P*-value <0.05. The R packages of "pHeatmap" was used for data visualization.[12]

Gene Ontology (GO) biological process analysis and string network

The clusterProfiler package was used to perform the GO functional enrichment analysis among the DEGs.[13] The significance of each GO term was defined by the $P<0.05$. The top 10 GO terms with the least *P*-value were listed. The active interaction sources contained experiments, databases and co-expression. The minimum interaction scores were defined as 0.4. The cytoscape software (http://www.cytoscape.org/) was used to visualize the co-expression network.[14]

WGCNA analysis

The R package "WGCNA" was used to cluster the gene co-expression network.[15] In case of outlier effects, pre-processing step was performed before the WGCNA analysis. Based on the expression matrix, the clustering analysis was performed to identify abnormal samples, which might bring bias to subsequent analysis. A soft thresholding power of 7 with a scale-free model fitting index R2>0.9 was applied to maximize scale-free topology, maintaining a high mean number of connections and eliminating small correlations. In WGCNA, a neighborhood proximity measurement was defined as topological overlap matrix (TOM), which quantified the degree of shared network neighbors. One-step network was constructed with the following parameters: maxBlockSize=10,000, minModuleSize=30, deepSplit=4. Then, a hierarchical clustering dendrogram was plotted with identified modules. Modules were defined as clusters of interconnected genes. The module eigengene (ME) was defined as the first principal component of a given module, which was considered a representative of the gene expression profiling. MEs were calculated to evaluate the correlation between the modules and the clinical traits (non-smoking status). Associations can then be determined on the basis of MEs. Sum of correlation coefficients with other nodes in a "signed" TOM defined the connectivity of one node. Hub genes were loosely identified as those with high network connectivity. Unweighted node connectivity information was used to identify hub genes within the 2 modules. The top strongest connections within the significant modules were identified to show the distribution of hub genes. We defined the significant hub genes as the top 25 genes ranked by gene significance (GS) in each module.

Statistical analysis

The prognostic signature was estimated in the training cohort by using penalized lasso Cox proportional hazards regression (R package "Glmnet"). The optimal lambda value was defined by 10-fold cross-validation. The number of candidate genes and corresponding coefficients were calculated by the optimal lambda value. The predictive ability was evaluated by the time-dependent receiver operating characteristic (ROC)

curve (R package "survivalROC"). This signature was carried over the testing datasets to validate the predictive ability. Coefficients were not re-estimated in testing dataset.

Results

Selecting DEGs in discovering datasets

Comprehensive search was conducted in GEO for RNA sequencing data and microarray expression profiles with LAC tissue samples. Datasets without non-smoking clinical records were excluded. To avoid bias, we excluded datasets with small sample sizes. The final accession numbers of datasets were GSE10072,[16] GSE31210,[17] GSE40419,[18] GSE68465,[19] and GSE50081.[20] One of them was RNA sequencing data and the others were microarray profiles. In addition, 214 LAC patients from TCGA met the inclusion criteria and were included. The baseline information of these datasets are listed in Table S1. We assigned GSE10072, GSE31210, GSE40419, and GSE68465 datasets into the discovering group. The patients from TCGA were treated as training group to build prognostic signature. GSE50081 and GSE31210 were identified as 2 external testing groups for validation (Figure S1). Different expression analysis of these datasets in discovery group revealed that 180 genes were down-regulated and 318 genes were up-regulated in GSE10072, 348 genes were down-regulated and 248 genes were up-regulated in GSE31210, 1,620 genes were down-regulated and 1,238 genes were up-regulated in GSE40419, and 660 genes were down-regulated and 803 genes were up-regulated in GSE68465 (FC>2, P<0.05). Different expression genes plotted are shown in Figure S2.

Constructing gene co-expression networks

To construct gene co-expression modules, DEGs of each dataset were submitted to WGCNA. DEGs were assigned to different co-expression networks by cluster dendrogram trees (Figure 1A). Unassigned genes were categorized into gray module. Different numbers of the co-expression modules were obtained from the different datasets, ranging from 4 to 10. The relationships between the clinical records with the co-expression networks are presented in Figure 1C. We found that the blue and turquoise modules were 2 significant networks related to non-smoking status across 4 datasets (Table 1). In addition, 2 networks showed an opposite correlation with non-smoking status (Figure 1D). TOM was visualized by heat map, which could depict adjacencies or topological overlaps. The topological overlap of two nodes reflected their similarity in terms of commonality of the nodes they connected to (Figure 1B).

Gene Ontology (GO) analysis of significant modules

GO enrichment analysis was conducted to identify the potential mechanisms of the significant modules. Based on the GO biological process, we observed that different expression genes of the blue module were mainly enriched in the nuclear division and cell cycle pathways (Figure 2A). For the turquoise modules, genes were engaged in the lymphocyte activation and cell adhesion pathways (Figure 2B). There were 131 genes in the blue modules from 4 datasets. In addition, a total of 352 genes were assigned in the turquoise modules across 4 datasets. String analysis plotted the inter-connection of different expression genes in 2 modules by the co-expression networks (Figure 2C, D). It could be seen that BUB1B, CCNB2, and TPX2 exhibited high connectivity with neighboring genes in the blue module. VWF, END1, and THBS2 were highly connecting hub nodes in the turquoise network. The size of nodes represented the degree of the correlation with the non-smoking status. The width of the lines was based on the co-expression value of 2 nodes. The hub genes were selected from the top 25 genes of each dataset based on the GS. Several genes were overlapped in the 4 datasets (Table 2).

Building prognostic signature

After combining the genes from 2 networks related to the non-smoking status, 234 overlapped genes were identified, including 58 genes selected simultaneously from 4 datasets. We clustered the expression of 17 candidate genes in the training cohort (Figure 3A). To build an efficient prognostic model, the penalized Cox regression model was used to narrow down the candidate genes and calculate the coefficients (Figure 3B). Finally, the prognostic signature was built by 17 genes at the optimal λ value (Risk score=0.016*$ADAM12$+0.001*$ASPN$+0.068*$COL1A2$−0.016*$DNALI1$−0.09*$FCGBP$+0.007*$FNDC1$ +0.03*$FOSL1$ + 0.069*$FSCN1$−0.107 *$GDF15$−0.041*HLF +0.088*$IGF2$BP3+0.002*$LRCH2$ + 0.027*$S100$A8+0.087*SCD−0.027*$SFTPB$ +0.043*$ST6GALNAC5$−0.014*$UNC5CL$). The performance of the signature was tested by time-dependent ROC curve. The area under the curve (AUC) of the training dataset was 0.736 (Figure 4A). X-tile was used to find out the optimal cut-off value of risk score for the training dataset. Patients with a risk score more than 0.204 were assigned to the high-risk subgroup and the rest of the patients were included into low-risk subgroup (Figure 4C). Log-ranked survival analysis showed that the high-risk subgroup had a poorer prognosis than the low-risk subgroup (hazard ratio [HR]=3.696, 95% CI: 2.025–6.748,

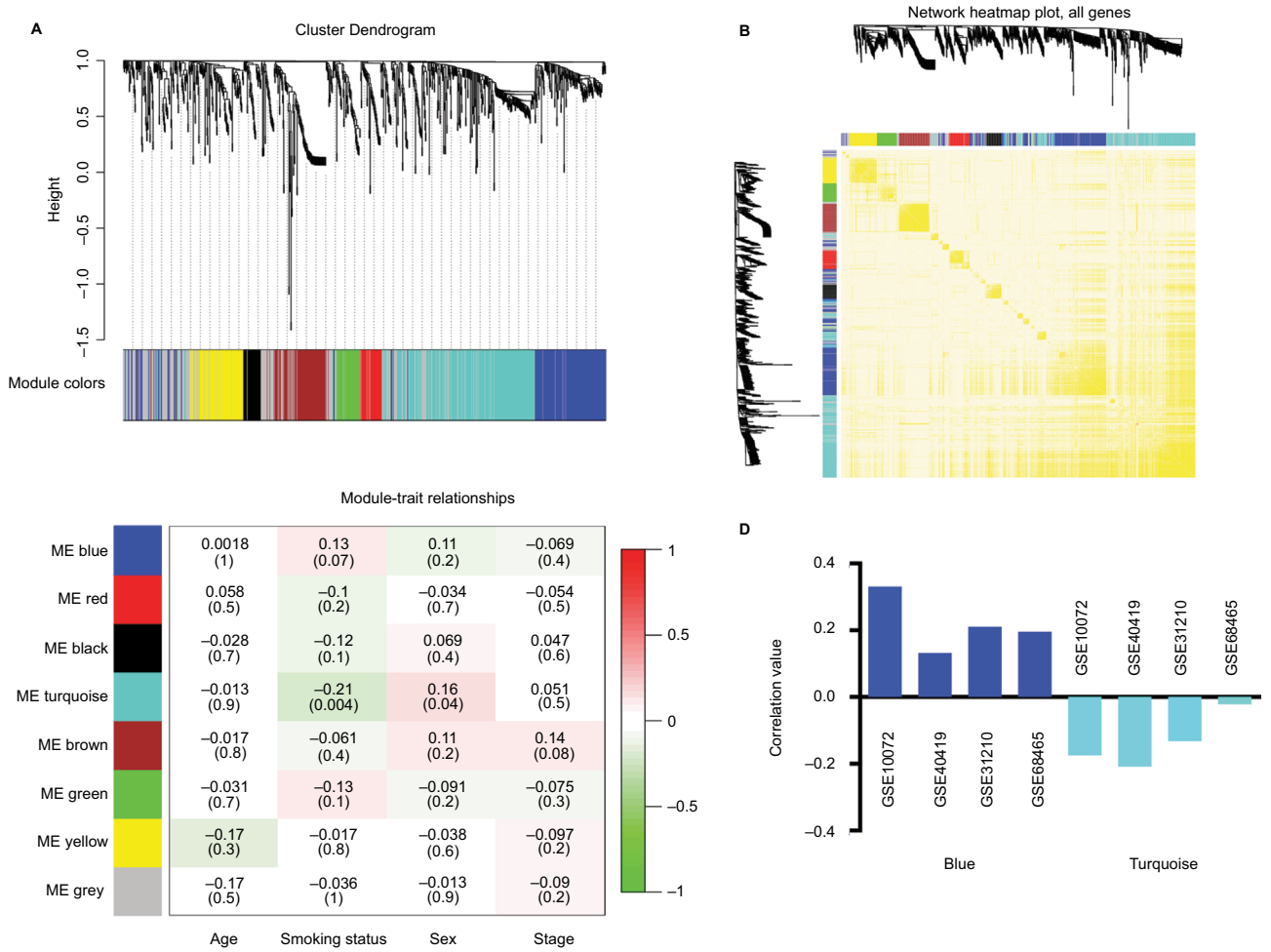

Figure 1 Weighted gene coexpression network analysis identified co-expression gene modules of LAC.
Notes: (**A**) Clustering dendrogram of different expression genes. Hierarchical cluster analysis dendrogram used to detect co-expression clusters. Each color is assigned to 1 module (gray represented unassigned genes). (**B**) Network heatmap plot. Genes were sorted in the rows and columns by the clustering tree. Light colors denoted low adjacency and darker colors denoted higher adjacency. (**C**) Correlation values of different module-trait relationships with different clinical records. (**D**) Correlation values of module-trait relationships of non-smoking related modules across 4 training datasets.
Abbreviation: ME, module eigengene.

Table 1 *P*-values of module-trait relationships of two non-smoking related modules across 4 training datasets

Datasets	Blue modules (*P*-value)	Turquoise modules (*P*-value)
GSE10072	0.041*	<0.001*
GSE40419	0.07	0.004*
GSE31210	0.038*	<0.001*
GSE68465	<0.001*	<0.001*

Note: *P-value is significant.

P<0.001). Multi-variable Cox analysis revealed that the risk score was an independent risk factor for survival of non-smoking LAC (Figure 4B).

Testing the prognostic signature

GSE50081 and GSE31210 were adopted to test performance of the prognostic signature to avoid over-fitting. Eighty cases were selected from GSE50081 dataset and 105 patients were included from GSE31210 with non-smoking LAC and survival records. ROC curve showed a good predictive ability of GSE50081 and a moderate predictive ability of GSE31210 (GSE50081: AUC =0.818, GSE31210: AUC =0.662). The prognostic signature could discriminate the high-risk subgroup from the low-risk subgroup in 2 testing datasets by survival analysis (GSE50081:HR =2.9, 95% CI: 1.322–6.789, *P*=0.006; GSE31210: HR =2.78, 95% CI: 1.658–6.654, *P*=0.022) (Figure 4D, E).

Discussion

The rapid increase in the number of non-smoking LAC makes it a novel hotspot of lung cancer prevention.[21] Current evidence indicates that non-smoking LAC carries more characterized driver genes and somatic mutations, which results in clinical disparities between the non-smoking LAC and the smoking LAC.[22,23] In addition, the differences of the

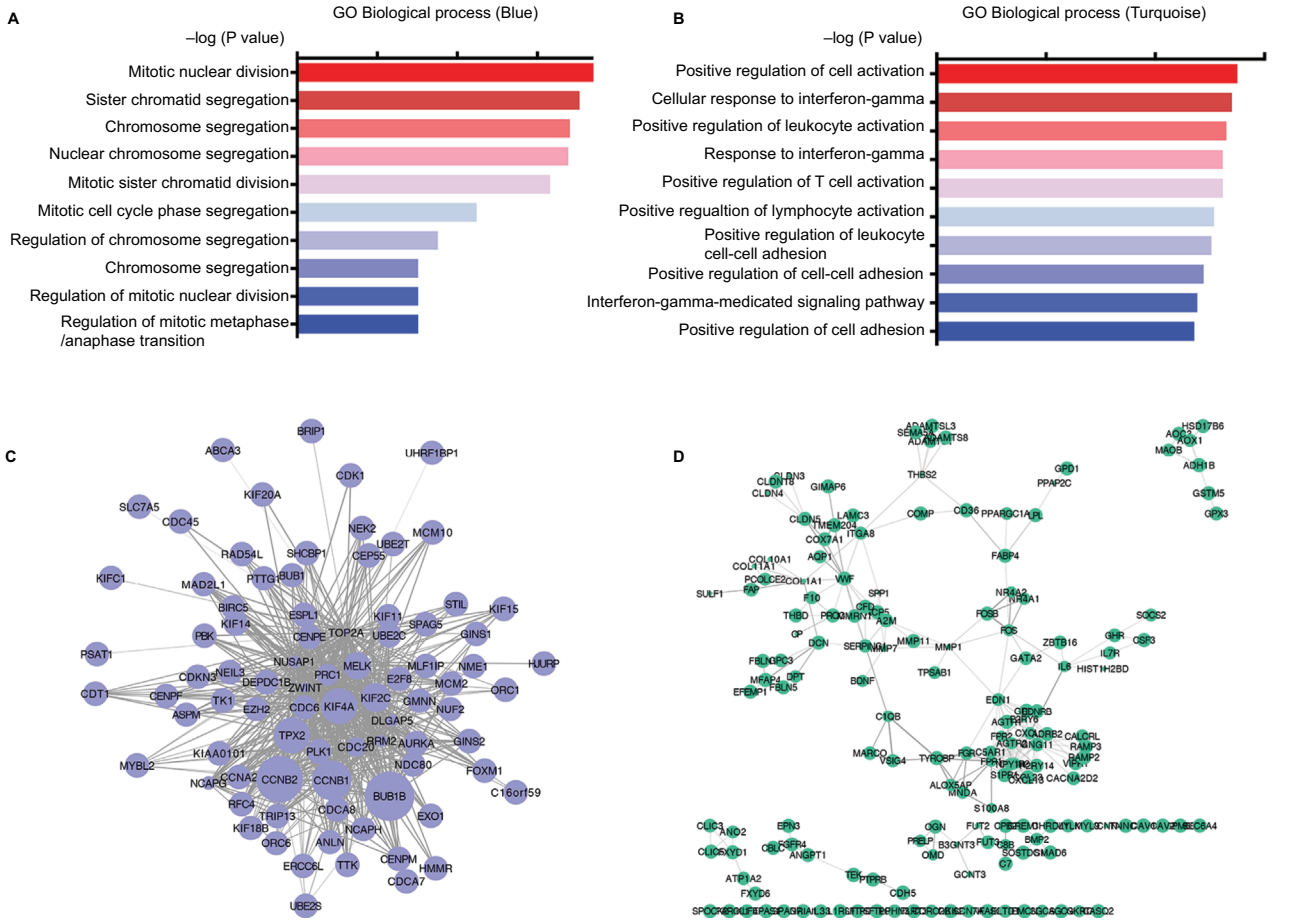

Figure 2 GO analysis and co-expression network of non-smoking related modules.
Notes: GO enrichment analysis of the blue and turquoise modules (**A** and **B**); visual representation of co-expression networks in the blue and turquoise modules. The width of the lines represented co-expression correlation value (**C** and **D**); different sizes of nodes indicated different Module Membership values.
Abbreviation: GO, Gene Ontology.

demographics and survival between the non-smoking group and the smoking group suggest that non-smoking LAC should be recognized as a separate group.[1,2,24] To explore reliable biomarkers of the non-smoking LAC, we identified 2 co-expression networks by the WGCNA and built a prognostic signature to predict the survival.

WGCNA is a promising approach to identify hub genes related to clinical features and mine significant gene co-expression networks. Zhang et al revealed a unique 22 carbon-metabolism gene expression signature in hepatocellular carcinoma (HCC), which might provide new therapeutic targets for HCC treatment.[25] Another study built a prognostic signature to predict the survival of gastric cancer by WGCNA.[26] In the discovery phase, we identified 2 significant co-expression networks related to the non-smoking status by the WGCNA across 4 datasets. The blue module positively correlated with the non-smoking status with the highest *P*-value. GO annotation analysis revealed that genes in the blue module were mainly enriched in the nuclear division and cell cycle pathways, indicating that aberrant nuclear division and dysregulated cell cycle might play critical roles in progression of non-smoking LAC. Wu et al's study found that a differentially regulated gene module was enriched for cell cycle related genes, which played a role in the molecular differences between the smoking and the non-smoking LAC.[27] Zhang et al reported that MID1-PP2A complex plays an important role in cell cycle arrest among non-smoking LAC patients.[28] In addition, Chen et al found that AhR modulated NFkB activity and up-regulated the interleukin-6 expression, which promoted lung carcinogenesis in non-smokers.[29] Our results were consistent with theirs. Another significant module that correlated with the non-smoking status was defined as turquoise. Compared with the blue module, the turquoise module negatively correlated with the non-smoking status. GO biological pathways demonstrated that genes of the turquoise module were enriched in the lymphocyte activation

Table 2 The top 25 hub genes of the blue and turquoise modules in 4 training datasets

Rank	GSE10072	GSE40419	GSE31210	GSE68465
Blue				
1	CDK1	SPAG5	ASPM	CCNA2
2	**CCNB1**	**KIF4A**	**MELK**	**CCNB2**
3	TOP2A	CDCA5	**DLGAP5**	PRC1
4	MAD2L1	**BUB1B**	KIF2C	**CCNB1**
5	CDC20	**CCNB2**	**RRM2**	**TPX2**
6	**BUB1B**	BIRC5	**TPX2**	**BUB1B**
7	PRC1	**TPX2**	**CCNB2**	KIF2C
8	ZWINT	KIF2C	NCAPG	**RRM2**
9	**NUSAP1**	KIF15	CENPA	**NUSAP1**
10	ECT2	KIF20A	CENPF	MAD2L1
11	CEP55	**CCNB1**	CEP55	**KIF4A**
12	KIF11	**NUSAP1**	TOP2A	CHEK1
13	ASPM	TTK	**KIF4A**	AURKA
14	**MELK**	NCAPG	**NUSAP1**	KIAA0101
15	**RRM2**	ESPL1	CDC20	**MELK**
16	KPNA2	KIF14	**BUB1B**	CDC6
17	**TPX2**	**MELK**	**CCNB1**	CENPA
18	CDKN3	UHRF1	BIRC5	HJURP
19	**KIF4A**	**RRM2**	TTK	CDK1
20	**CCNB2**	NCAPH	NEK2	FEN1
21	**DLGAP5**	CDCA8	HMMR	BIRC5
22	CENPU	CAV1	KIAA0101	**DLGAP5**
23	CKS1B	POLQ	FOXM1	BUB1
24	CENPA	KIFC1	ZWINT	KIF18B
25	RFC4	**DLGAP5**	ORC6	CDKN3
Turquoise				
1	**TCF21**	GRK5	TAL1	HLA-DRA
2	FHL1	FHL1	**LDB2**	**LDB2**
3	EDNRB	ARHGAP6	TEK	HLA-DQA1
4	GRK5	CAV1	MMRN2	C1QA
5	TEK	NXPH3	PTPRB	HLA-DPA1
6	FAM107A	ABCA8	FHL5	HLA-DQB1
7	FIGF	CDO1	S1PR1	HLA-DRB1
8	PECAM1	**LDB2**	KANK3	CD14
9	JAM2	ABI3BP	EDNRB	HLA-DRB1
10	AGER	CCBE1	ARHGAP6	HLA-DQB1
11	**CDH5**	GPM6A	ERG	**TCF21**
12	CLEC3B	ADH1B	SASH1	CSF1R
13	ABCA8	**TCF21**	EMCN	GIMAP4
14	**LDB2**	FGD5	TGFBR3	**CDH5**
15	CA4	AOC3	PKNOX2	SERPING1
16	HIGD1B	RADIL	FAM107A	CD163
17	FOXF1	SCUBE1	AOC3	HLA-DPB1
18	TACC1	**CDH5**	CLEC3B	PECAM1
19	STARD13	LTBP4	ADAMTSL3	ENTPD1
20	RAMP2	TGFBR3	**TCF21**	GIMAP6
21	AOC3	GRIA1	ASPA	FCER1G
22	TGFBR3	KANK3	RASIP1	CD4
23	ADH1B	CAV2	DACH1	HLA-DQB1
24	VWF	ACVRL1	CDO1	SPARCL1
25	S1PR1	ADAMTSL3	**CDH5**	SLC7A7

Notes: Bold font: Overlapped hub genes in 4 datasets.

and cell adhesion pathways. These results were consistent with the previous studies.[30] Peng et al's study indicated that several genes affect the prognosis of LAC patients through regulating cell cycle and cell adhesion.[31] Another study found that different active T cells promoted the progression of non-smoking LAC by creating an immunosuppressive microenvironment,[32] which supported our findings.

Due to heterogeneity of expression profiles, hub genes of different datasets were not absolutely accordant. However, several significant hub genes were shared by 4 different datasets, and some of them were reported by published studies. Based on previous studies, DLGAP5 was identified as a promising diagnostic and prognostic biomarker in lung cancer.[33] In addition, integrated genome-scale co-expression network revealed that DLGAP5 played a crucial role in cell cycle progression of LAC, which was consistent with our analysis.[34] And our analysis further pointed out that high expression of DLGAP5 showed poor survival for the non-smoking LAC. In addition, TPX2 was defined as a prognostic biomarker for lung cancer and engaged in cell division and cell cycle pathways according to published studies.[35] Our analysis further reinforced the important roles of TPX2 in the non-smoking LAC. Published studies indicated that epigenetic deregulation of TCF21 inhibited malignant behavior of lung cancer.[36] Existing evidence demonstrated that LDB2 engaged in the epithelial–mesenchymal transition (EMT) and the cell adhesion pathways.[37] Our analysis highlighted LDB2 as a prognostic biomarker and potential therapeutic target for non-smoking LAC. However, several novel targets, like *COL1A2* and *CDH5*, have not been well reported by previous studies. Further studies are needed to identify the mechanisms of these genes.

The prognostic signature was defined by combining several transcriptome profiles from non-smoking samples in GEO datasets and TCGA. The multiple datasets and analysis method avoided the biases from batch effect and platform.[38] The Cox penalized regression model was used to identify prognostic genes and corresponding coefficients. The predictive ability of prognostic signature was moderate in the training dataset, but it was good in 1 testing dataset, indicating excellent generalization of the prognostic signature. Survival analysis showed that significant distinction between the high-risk and low-risk groups in 2 testing datasets, which implied that the signature was a feasible tool to stratify high-risk non-smoking LAC patients.

Increasing studies have proposed the prognostic signatures for survival prediction of LAC. The first RNA-seq

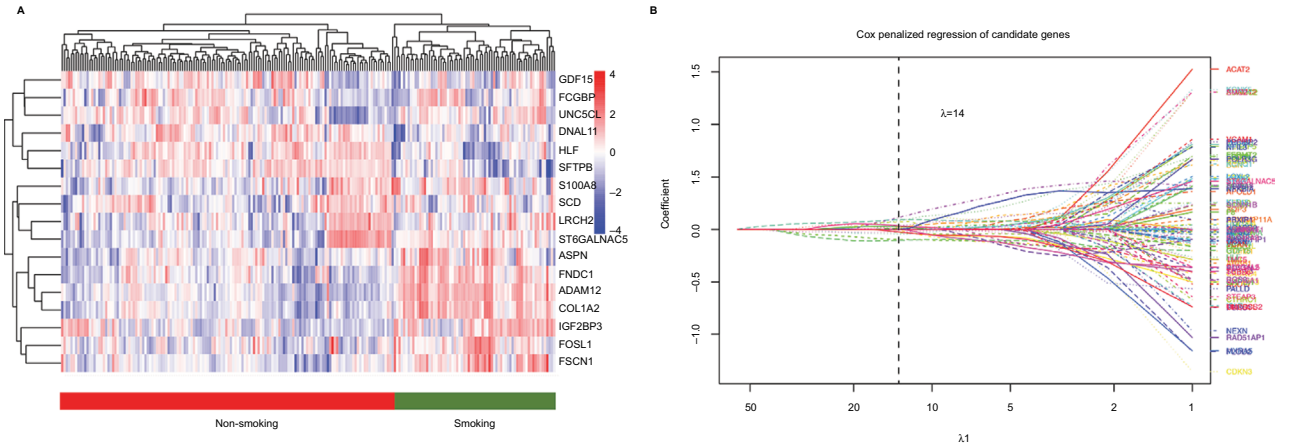

Figure 3 Cluster analysis of 17 candidate genes selected by penalized Cox regression in the training group (**A**). Penalized Cox regression analysis to select survival-associated genes in the training group. (**B**) The optimal λ value is 14.

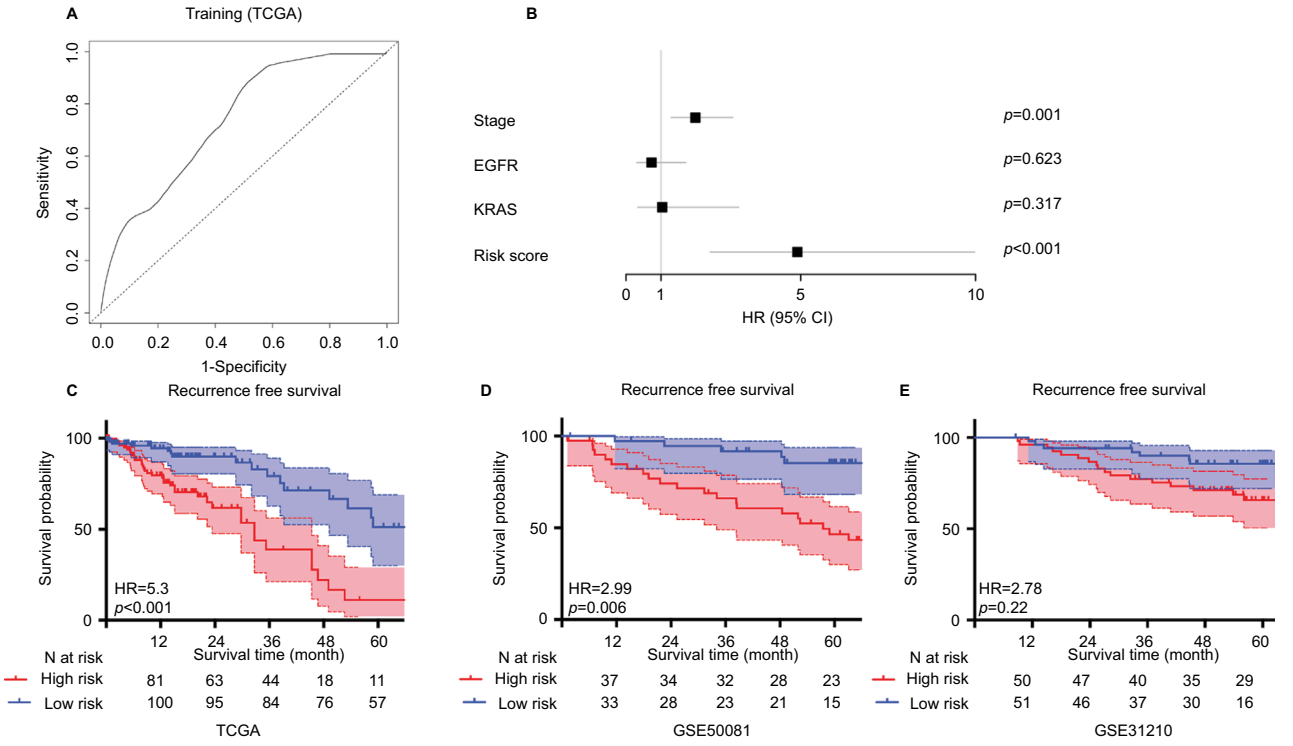

Figure 4 ROC curves for 17 genes to predict the survival of non-smoking LAC (**A**). Multi-variable Cox analysis indicated that risk score was an independent prognostic risk factor by adjusting other variables (**B**). The performance of the prognostic signature in stratifying the high-risk and low-risk groups. (**C**) training cohort (TCGA), (**D**) external testing cohort 1 (GSE50081), (**E**) external testing cohort 2 (GSE50081).

Abbreviations: EGFR, epidermal growth factor receptor; HR, hazard ratio; KRAS, Kirsten ras; LAC, lung adenocarcinoma; TGCA, The Cancer Genome Atlas.

prognostic signature for LAC was proposed by Shukla et al, which provided a powerful prognostic tool for precision oncology.[39] In addition, the prognostic predictor based on alternative splicing events uncovered prognostic effect of the splicing networks in LAC.[40] A recent study reported that a P53-deficiency gene signature could predict recurrence risk of patients with early-stage LAC.[41] However, few predicted the survival of non-smoking LAC patients. This was the first study to develop a prognostic signature based on 17 non-smoking related genes for survival of non-smoking LAC. The prognostic signature was tested in 2 independent datasets from different demographics to guarantee the generalization. In addition, our signature could stratify patients into the high-risk group and the low-risk group with different

survival outcomes. Compared with previous biomarkers, our model first leveraged the molecular biomarkers from co-expression networks by the WGCNA to accurately estimate the survival of the non-smoking LAC, which might aid to guide the therapeutic management.

The current study had several limitations. First, we did not test the expression of hub genes and performance of prognostic signature by our own samples. Second, we only used expression profiles in our signature. However, combining meta-omics biomarkers into signature would further improve the predictive ability.[42] Furthermore, the role of hub genes should be explored by further experimental procedures, which might reinforce the significance and robustness of this analysis.

In this study, we highlighted 2 gene modules related to non-smoking LAC and built a prognostic signature, which provide the novel compendium of biomarkers and guide the therapy in the non-smoking LAC.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Nos. 81472702, 81501977 and 81672294), Natural Science Foundation of Jiangsu Province (No. SBK016030028), and the Innovation Capability Development Project of Jiangsu Province (No. BM2015004). Thanks to Jing Han from Department of Epidemiology and Biostatistics, School of Public Health, Nanjing Medical University for helping with statistical analysis. The abstract of this paper was presented at the European Lung Cancer Congress as a poster presentation with interim findings. The poster's abstract was published in "Poster Abstracts" in the *Journal of Thoracic Oncology*.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Stiles BM, Rahouma M, Hussein MK, et al. Never smokers with resected lung cancer: different demographics, similar survival. *Eur J Cardiothorac Surg*. 2018;53(4):842–848.
2. Cho J, Choi SM, Lee J, et al. Proportion and clinical features of never-smokers with non-small cell lung cancer. *Chin J Cancer*. 2017;36(1):20.
3. Thu KL, Vucic EA, Chari R, et al. Lung adenocarcinoma of never smokers and smokers harbor differential regions of genetic alteration and exhibit different levels of genomic instability. *PLoS One*. 2012;7(3):e33003.
4. Song MA, Benowitz NL, Berman M, et al. Cigarette Filter Ventilation and its Relationship to Increasing Rates of Lung Adenocarcinoma. *J Natl Cancer Inst*. 2017;109(12).
5. Dong Y, Li Y, Jin B, et al. Pathologic subtype-defined prognosis is dependent on both tumor stage and status of oncogenic driver mutations in lung adenocarcinoma. *Oncotarget*. 2017;8(47):82244–82255.
6. Clarke C, Madden SF, Doolan P, et al. Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. *Carcinogenesis*. 2013;34(10):2300–2308.
7. Sun Q, Zhao H, Zhang C, et al. Gene co-expression network reveals shared modules predictive of stage and grade in serous ovarian cancers. *Oncotarget*. 2017;8(26):42983–42996.
8. Liu R, Zhang W, Liu ZQ, Zhou HH. Associating transcriptional modules with colon cancer survival through weighted gene co-expression network analysis. *BMC Genomics*. 2017;18(1):361.
9. Gautier L, Cope L, Bolstad BM, Irizarry RA, Fau CL, Fau BB, Irizarry RA. affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*. 2004;20(3):307–315.
10. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47–e47.
11. Robinson MD, Mccarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139–140.
12. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*; Springer-Verlag, New York, 2009.
13. Yu G, Wang LG, Han Y, He QY,. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012;16(5):284–287.
14. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498–2504.
15. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9(1):559.
16. Landi MT, Dracheva T, Rotunno M, et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS One*. 2008;3(2):e1651.
17. Yamauchi M, Yamaguchi R, Nakata A, et al. Epidermal growth factor receptor tyrosine kinase defines critical prognostic genes of stage I lung adenocarcinoma. *PLoS One*. 2012;7(9):e43923.
18. Seo JS, Ju YS, Lee WC, et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome Res*. 2012;22(11):2109–2119.
19. Shedden K, Taylor JMG, Enkemann SA, et al. Gene Expression-Based Survival Prediction in Lung Adenocarcinoma: A Multi-Site, Blinded Validation Study: Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma. *Nat Med*. 2008;14(8):822–827.
20. der SD, Sykes J, Pintilie M, et al. Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including stage IA patients. *J Thorac Oncol*. 2014;9(1):59–64.
21. Lu TP, Tsai MH, Lee JM, et al. Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. *Cancer Epidemiol Biomarkers Prev*. 2010;19(10):2590–2597.
22. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers--a different disease. *Nat Rev Cancer*. 2007;7(10):778–790.
23. Gou L-Y, Niu F-Y, Y-L W, Zhong W-Z. Differences in driver genes between smoking-related and non–smoking-related lung cancer in the Chinese population. *Cancer*. 2015;121(S17):3069–3079.
24. Yano T, Miura N, Takenaka T, et al. Never-smoking nonsmall cell lung cancer as a separate entity. *Cancer*. 2008;113(5):1012–1018.
25. Zhang J, Baddoo M, Han C, et al. Gene network analysis reveals a novel 22-gene signature of carbon metabolism in hepatocellular carcinoma. *Oncotarget*. 2016;7(31):49232–49245.
26. Zhao X, Cai H, Wang X, Ma L. Discovery of signature genes in gastric cancer associated with prognosis. *Neoplasma*. 2016;63(2):239–245.
27. Wu C, Zhu J, Zhang X. Network-based differential gene expression analysis suggests cell cycle related genes regulated by E2F1 underlie the molecular difference between smoker and non-smoker lung adenocarcinoma. *BMC Bioinformatics*. 2013;14(1):365.
28. Zhang L, Li J, Lv X, Guo T, Li W, Zhang J. MID1-PP2A complex functions as new insights in human lung adenocarcinoma. *J Cancer Res Clin Oncol*. 2018;144(5):855–864.

29. Chen PH, Chang H, Chang JT, Lin P. Aryl hydrocarbon receptor in association with RelA modulates IL-6 expression in non-smoking lung cancer. *Oncogene*. 2011;31:2555.
30. Zhou My W, Cui H, Wang N, et al. Identification of potential therapeutic target genes and mechanisms in non-small-cell lung carcinoma in non-smoking women based on bioinformatics analysis. *Eur Rev Med Pharmacol Sci*. 2015;19(18):3375–3384.
31. Peng F, Wang R, Zhang Y, et al. Differential expression analysis at the individual level reveals a lncRNA prognostic signature for lung adenocarcinoma. *Mol Cancer*. 2017;16(1):98.
32. Kinoshita T, Kudo-Saito C, Muramatsu R, et al. Determination of poor prognostic immune features of tumour microenvironment in non-smoking patients with lung adenocarcinoma. *Eur J Cancer*. 2017;86:15–27.
33. Shi YX, Yin JY, Shen Y, Zhang W, Zhou HH, Liu ZQ. Genome-scale analysis identifies NEK2, DLGAP5 and ECT2 as promising diagnostic and prognostic biomarkers in human lung cancer. *Sci Rep*. 2017;7(1):8072.
34. Bidkhori G, Narimani Z, Hosseini Ashtiani S, Moeini A, Nowzari-Dalini A, Masoudi-Nejad A. Reconstruction of an Integrated Genome-Scale Co-Expression Network Reveals Key Modules Involved in Lung Adenocarcinoma. *PLoS ONE*. 2013;8(7):e67552.
35. Orth M, Unger K, Schoetz U, Belka C, Lauber K. Taxane-mediated radiosensitization derives from chromosomal missegregation on tripolar mitotic spindles orchestrated by AURKA and TPX2. *Oncogene*. 2018;37(1):52–62.
36. Richards KL, Zhang B, Sun M, et al. Methylation of the Candidate Biomarker TCF21 Is Very Frequent Across A Spectrum of Early Stage Non-Small Cell Lung Cancers. *Cancer*. 2011;117(3):606–617.
37. Chen HN, Yuan K, Xie N, et al. PDLIM1 Stabilizes the E-Cadherin/β-Catenin Complex to Prevent Epithelial-Mesenchymal Transition and Metastatic Potential of Colorectal Cancer Cells. *Cancer Res*. 2016;76(5):1122–1134.
38. Chu SH, Huang YT. Integrated genomic analysis of biological gene sets with applications in lung cancer prognosis. *BMC Bioinformatics*. 2017;18(1):336.
39. Shukla S, Evans JR, Malik R, et al. Development of a RNA-Seq Based Prognostic Signature in Lung Adenocarcinoma. *J Natl Cancer Inst*. 2017;109(1):djw200.
40. Li Y, Sun N, Lu Z, et al. Prognostic alternative mRNA splicing signature in non-small cell lung cancer. *Cancer Letters*. 2017;393(Supplement C):40–51.
41. Zhao Y, Varn FS, Cai G, Xiao F, Amos CI, Cheng C. A P53-Deficiency Gene Signature Predicts Recurrence Risk of Patients with Early-Stage Lung Adenocarcinoma. *Cancer Epidemiol Biomarkers Prev*. 2018;27(1):86–95.
42. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clin Cancer Res*. 2018;24(6):1248–1259.

Supplementary material

Table S1 Information of training and validation GEO datasets

Datasets	Platform	Sample size	Smoking status (never/smoker)	Stage (I/II/III/IV)	Gender (female/male)
Discovery					
GSE10072	Affymetrix Human Genome U133A Array	107	30/77	45/35/21/6	38/69
GSE40419	Illumina HiSeq 2000	164	70/94	109/24/23/8	67/97
GSE31210	Affymetrix Human Genome U133 Plus 2.0 Array	246	123/123	168/58	130/116
GSE68465	Affymetrix Human Genome U133A Array	440	49/391	276/102/50/12	220/220
Training					
TCGA	Illumina Hiseq	524	214/310	283/125/84/27	277/243
Validation					
GSE50081	Affymetrix Human Genome U133 Plus 2.0 Array	181	103/58	127/54	84/97
GSE31210	Affymetrix Human Genome U133 Plus 2.0 Array	246	123/123	168/58	130/116

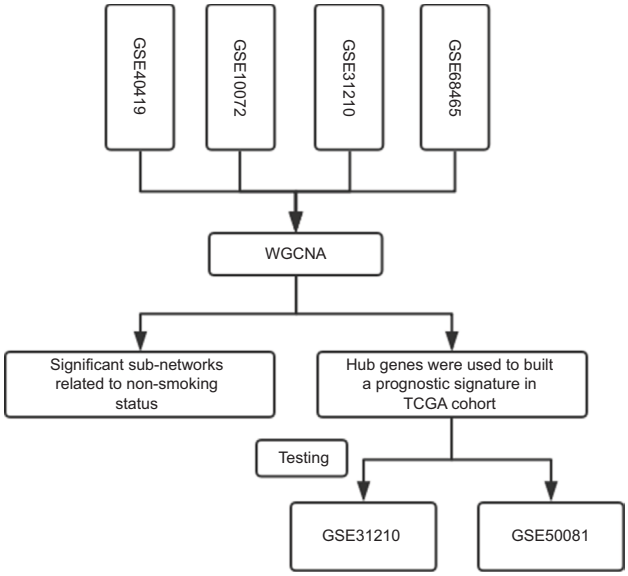

Figure S1 The flow chat of the study.
Abbreviation: WGCNA, Weighted correlation network analysis.

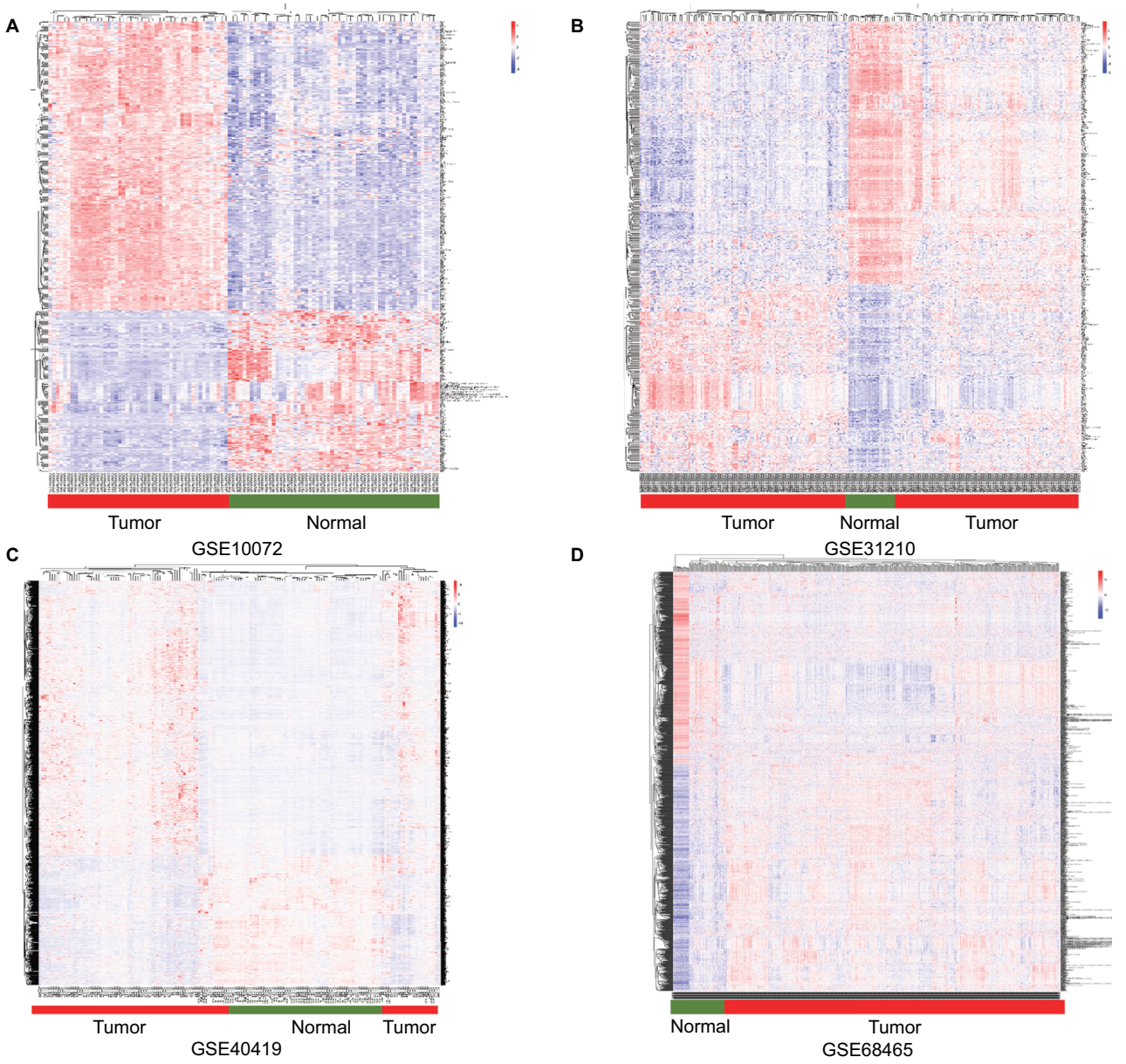

Figure S2 The expression profile in lung adenocarcinoma tissues and normal tissues. (**A–D**) Heatmap of the different expression genes in GSE10072, GSE31210, GSE40419 and GSE68465 datasets.