MEDICAL SCIENCE MONITOR

DATABASE ANALYSIS

e-ISSN 1643-3750
© Med Sci Monit, 2020; 26: e925833
DOI: 10.12659/MSM.925833

Received: 2020.05.09
Accepted: 2020.07.23
Available online: 2020.08.19
Published: 2020.10.06

Creation of a Prognostic Risk Prediction Model for Lung Adenocarcinoma Based on Gene Expression, Methylation, and Clinical Characteristics

Authors' Contribution:
Study Design A
Data Collection B
Statistical Analysis C
Data Interpretation D
Manuscript Preparation E
Literature Search F
Funds Collection G

CD 1 **Honggang Ke***
DEF 2 **Yunyu Wu***
CDF 3 **Runjie Wang**
AB 4 **Xiaohong Wu**

1 Department of Cardiovascular and Thoracic Surgery, Affiliated Hospital of Nantong University, Nantong, Jiangsu, P.R. China
2 Qixiu Campus, Nantong University, Nantong, Jiangsu, P.R. China
3 Department of Oncology, Wuxi People's Hospital, Wuxi, Jiangsu, P.R. China
4 Department of Medical Oncology, Affiliated Hospital of Jiangnan University and Wuxi 4th People's Hospital, Wuxi, Jiangsu, P.R. China

* Honggang Ke and Yunyu Wu are Co-first authors

Corresponding Author: Xiaohong Wu, e-mail: wxh_112112@163.com

Source of support: Departmental sources

Background: This study aimed to identify important marker genes in lung adenocarcinoma (LACC) and establish a prognostic risk model to predict the risk of LACC in patients.

Material/Methods: Gene expression and methylation profiles for LACC and clinical information about cases were downloaded from the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) databases, respectively. Differentially expressed genes (DEGs) and differentially methylated genes (DMGs) between cancer and control groups were selected through meta-analysis. Pearson coefficient correlation analysis was performed to identify intersections between DEGs and DMGs and a functional analysis was performed on the genes that were correlated. Marker genes and clinical factors significantly related to prognosis were identified using univariate and multivariate Cox regression analyses. Risk prediction models were then created based on the marker genes and clinical factors.

Results: In total, 1975 DEGs and 2095 DMGs were identified. After comparison, 16 prognosis-related genes (*EFNB2*, *TSPAN7*, *INPP5A*, *VAMP2*, *CALML5*, *SNAI2*, *RHOBTB1*, *CKB*, *ATF7IP2*, *RIMS2*, *RCBTB2*, *YBX1*, *RAB27B*, *NFATC1*, *TCEAL4*, and *SLC16A3*) were selected from 265 overlapping genes. Four clinical factors (pathologic N [node], pathologic T [tumor], pathologic stage, and new tumor) were associated with prognosis. The prognostic risk prediction models were constructed and validated with other independent datasets.

Conclusions: An integrated model that combines clinical factors and gene markers is useful for predicting risk of LACC in patients. The 16 genes that were identified, including *EFNB2*, *TSPAN7*, *INPP5A*, *VAMP2*, and *CALML5*, may serve as novel biomarkers for diagnosis of LACC and prediction of disease prognosis.

MeSH Keywords: **Genes, Intracisternal A-Particle • Medical Oncology • Molecular Conformation**

Full-text PDF: https://www.medscimonit.com/abstract/index/idArt/925833

2885 8 7 56

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Background

Lung cancer is one of the most common cancers and a severe threat to human health. The number of lung cancer-related deaths is growing, with an estimated one-quarter of cancer-related deaths due to the disease [1]. There are 2 main types of lung cancer: small cell lung cancer and non-small cell lung cancer (NSCLC). Lung adenocarcinoma (LACC) is the most frequent histological subtype of NSCLC, accounting for approximately 75% of all cases of lung cancer. Over the past few decades, incidence of LACC in China has rapidly increased [2]. Despite recent advances in multimodality therapy, the overall 5-year survival rate for patients with LACC is only 15% [3], because two-thirds of lung cancers are discovered at advanced stages. Furthermore, 30% to 55% or more of patients who undergo resection for lung cancer experience relapse of disease within 5 years and die of metastatic recurrence [4]. Currently, it is impossible to accurately identify specific patients at high risk of recurrence to provide individualized therapy.

In recent years, molecular characterization of NSCLC has reached an unprecedented level of detail [5,6]. Vascular invasion, poor differentiation, tumor size, and high tumor proliferation index have been found to have prognostic significance. In addition, advances in human genomics have revolutionized methods of identifying new prognostic factors for human cancer [7,8]. For instance, Jiang et al. [9] identified 16 survival marker genes on the basis of datasets from previous studies. Beer et al. [10] evaluated a group of survival marker genes for use in identification of high-risk patients with LACC. Moreover, global gene expression profiling based on microarray technology has identified novel gene signatures and potential biomarkers to better predict patient prognosis in lung cancer [11–15], such as *KRAS* [16], *p53* [17], *SLC1A6*, *MGB1*, *REG1A*, and *AKAP12* [18]. Despite this progress, however, it remains challenging to accurately predict prognosis in patients with LACC.

In this study, we integrated gene expression profiling, methylation profiling and clinical characteristics to identify important marker genes that could predict survival and prognosis in a cohort of patients with LACC. A comprehensive prognostic risk model was constructed based on tumor marker genes and clinical factors. Reasonable use of reliable tumor markers may be helpful in early diagnosis of LACC and prediction of prognosis in patients with the disease.

Material and Methods

Data collection for meta-analysis

The datasets for LACC, including gene expression and methylation profiles obtained from the same patient population, were downloaded from the National Center of Biotechnology Information Gene Expression Omnibus (GEO) database (*http://www.ncbi.nlm.nih.gov/geo/*) and the European Bioinformatics Institute database on September 5, 2017. The datasets were further screened according to the following inclusion criteria: (1) Presence of LACC and normal control samples; (2) Availability of more than 50 samples; and (3) More than 20 000 total probes detected in the dataset. Finally, a total of 7 gene expression profile (GSE75037, GSE33532, GSE43458, GSE30219, GSE32863, GSE10072 and GSE62949) and 4 methylation profile datasets (GSE32861, GSE49996, GSE63384, and GSE62948) were selected. Detailed information about them is shown in Table 1. Furthermore, GSE62949 and GSE62948 were both included in dataset GSE62950 (*https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62950*).

Data collection for construction of the prognostic risk prediction model

Data on gene expression and methylation profiles for LACC used to construct the prognostic risk prediction model were downloaded from The Cancer Genome Atlas (TCGA) database (*https://gdc-portal.nci.nih.gov/*). After matching the methylation and gene expression profiles, 473 matched tumor samples were obtained. A total of 335 tumor samples were obtained by removing the samples that did not have survival prognosis information. These data were used as the training dataset. At the same time, the expression profile for LACC tissue, GSE37745 [19] (platform: GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array), was downloaded from the GEO database. This dataset, which contains 106 LACC tissue samples, was used as an independent validation dataset for the prognostic risk prediction model. Clinical information about the 2 datasets is shown in Table 2.

Preprocessing, quality control, and differential expression analysis of data used in the meta-analysis

We used the oligo package [20] (*http://www.bioconductor.org/packages/release/bioc/html/oligo.html*) in R3.4.1 language for CEL data conversion, missing values supplementation (median method), background correction (MAS method), and data normalization (quantile method) of the GSE333532, GSE43458, GSE30219, and GSE1072 datasets, which were downloaded from the GEO database based on the Affy platform. Using the limma package [21] (*https://bioconductor.org/packages/release/bioc/html/limma.html*) in the R3.4.1 language with the Illumima platform (quantile method), gene annotation, log2 conversion, and data normalization were performed on the GSE75037 and GSE32863 datasets (TXT format). For methylation profiling of GSE32861, GSE49996, GSE63384, and GSE62948, we identified the chromosomal sites and methylation beta values using the GenomeStudio Methylation Module [22].

Table 1. The gene expression profiling and methylation profiling datasets in this study.

GEO accession		Platform	Total probe number	Total sample	Normal sample	Cancer sample
Gene expression	GSE75037	GPL6884 Illumina	48803	166	83	83
	GSE33532	GPL570 Affymetrix	54675	60	40	20
	GSE43458	GPL6244 Affymetrix	33297	110	80	30
	GSE30219	GPL570 Affymetrix	54675	98	84	14
	GSE32863	GPL6884 Illumina	48803	116	58	58
	GSE10072	GPL96 Affymetrix	22283	107	58	49
Gene methylation	GSE32861	GPL8490 Illumina	27578	118	59	59
	GSE49996	GPL8490 Illumina	27578	88	44	44
	GSE63384	GPL8490 Illumina	27578	70	35	35
	GSE62948	GPL8490 Illumina	27578	56	28	28

GEO – Gene Expression Omnibus.

Table 2. Clinical information from The Cancer Genome Atlas (TCGA) and GSE62254 datasets.

Clinical characteristics	TCGA (N=335)	GSE37745 (N=106)
Age (years, mean±SD)	65.19±10.25	62.94±9.22
Sex (Male/Female)	155/180	46/60
Pathologic M (M0/M1/–)	226/13/96	–
Pathologic N (N0/N1/N2/–)	214/60/55/6	–
Pathologic T (T1/T2/T3/T4/–)	111/180/29/14/1	–
Pathologic stage (I/II/III/IV)	180/81/61/13	70/19/13/4
Radiation therapy (yes/no/–)	41/254/40	–
Targeted molecular therapy (yes/no/–)	99/194/42	–
Tobacco smoking history (current/reformed/never/–)	70/206/45/14	–
Recurrence (yes/no/–)	104/176/55	26/27
Death (dead/alive)	120/215	77/29
Recurrence-free survival time (months, mean±SD)	22.27±27.77	54.11±53.48
Overall survival time (months, mean±SD)	27.54±29.74	61.74±49.96

'–' – Represents information unavailable.

The main purpose of the meta-analysis was to comprehensively generate multiple research results using multiple experimental datasets, improve the ability to generate statistics, and screen for more reliable genes. Because these datasets were collected from different samples and experiments, they may be subject to bias. Therefore, the MetaQC [23] package (*https://cran.r-project.org/web/packages/MetaQC/index.html*) in R3.4.1 was used to perform quality control on the datasets. Next, the differentially expressed genes (DEGs) and differentially methylated genes (DMGs) were screened out using MetaDE.ES in MetaDE [24] package (*https://cran.r-project.org/web/packages/MetaDE*). The tau^2=0, and Qpval >0.05 were used as the homogeneity test parameters; a false discovery rate (FDR) <0.05 was set as the threshold.

Analysis of correlation between gene expression and methylation levels

For the above obtained DEGs and DMGs, we selected the intersection genes, which then served as candidate tumor marker genes. The Pearson coefficient correlation between gene expression level and methylation level was calculated using the cor function (*https://www.rdocumentation.org/packages/stats/versions/3.4.1/topics/cor*) in R3.4.1. Then the DAVID

Table 3. MetaQC quality control of 6 expression profiling datasets and 4 methylation profiling datasets.

	IQC	EQC	CQCg	CQCp	AQCg	AQCp	SMR
Gene expression profiling							
GSE75037	5.27	3.23	106.65	158.86	32.71	90.88	1.62
GSE32863	4.38	3.16	64.14	146.51	26.46	96.74	2.42
GSE33532	4.81	3.23	59.25	171.49	25.50	84.37	2.86
GSE43458	6.09	1.10	101.10	114.30	19.53	29.46	3.92
GSE30219	6.64	3.71	83.97	107.69	47.87	63.89	4.33
GSE10072	8.06	9.19	12.24	8.92	9.78	14.52	7.76
Methylation profiling							
GSE32861	9.80	5.00	19.24	41.01	6.17	24.77	3.28
GSE49996	6.22	4.96	46.70	42.02	8.67	33.56	3.14
GSE63384	7.56	3.05	24.57	33.79	3.45	17.84	5.67
GSE62948	5.11	3.63	59.25	60.27	29.49	84.37	3.25

IQC – internal quality control; EQC – external quality control; CQC – consistency quality control; AQC – accuracy quality control; SMR – standardized mean rank score.

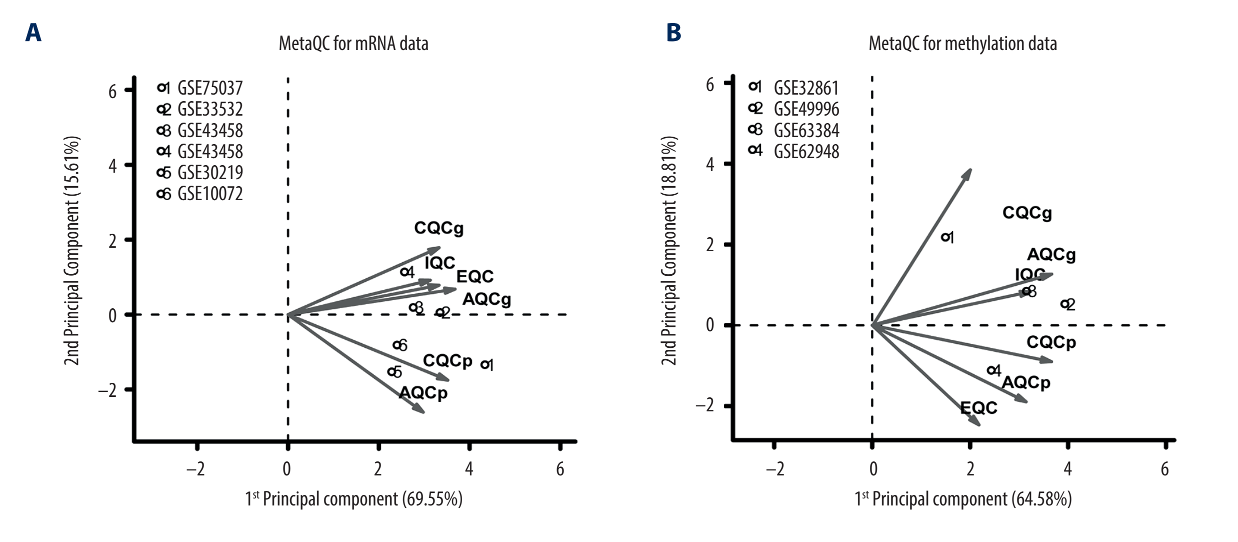


Figure 1. MetaQC quality control charts of (**A**) 5 gene expression profiles and (**B**) 2 gene methylation profiles. The horizontal and vertical axes represent the first and second principal components in principal component analysis. The numbers represent the corresponding datasets.

Bioinformatics Database, v6.8 [25,26] (*https://david.ncifcrf.gov/*), was used to perform enrichment analyses of the candidate tumor marker genes from the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases.

Screening of tumor marker genes and clinical factors related to prognosis

From among the total tumor marker gene set and the corresponding clinical factors for the tumor samples, we then identified the tumor marker genes and clinical factors significantly related to prognosis, using the univariate and multivariate Cox regression analyses in the Bioconductor R3.4.1 survival package [27] (*http://bioconductor.org/packages/survival/*). A log-rank test P<0.05 was used as the threshold for significance.

Construction and validation of the risk prediction model

Using the prognosis-associated tumor marker genes identified by the Cox regression analysis, we constructed a risk prediction model and calculated a prognostic index (PI) for each sample. The samples in the training set were divided into high- and low-risk

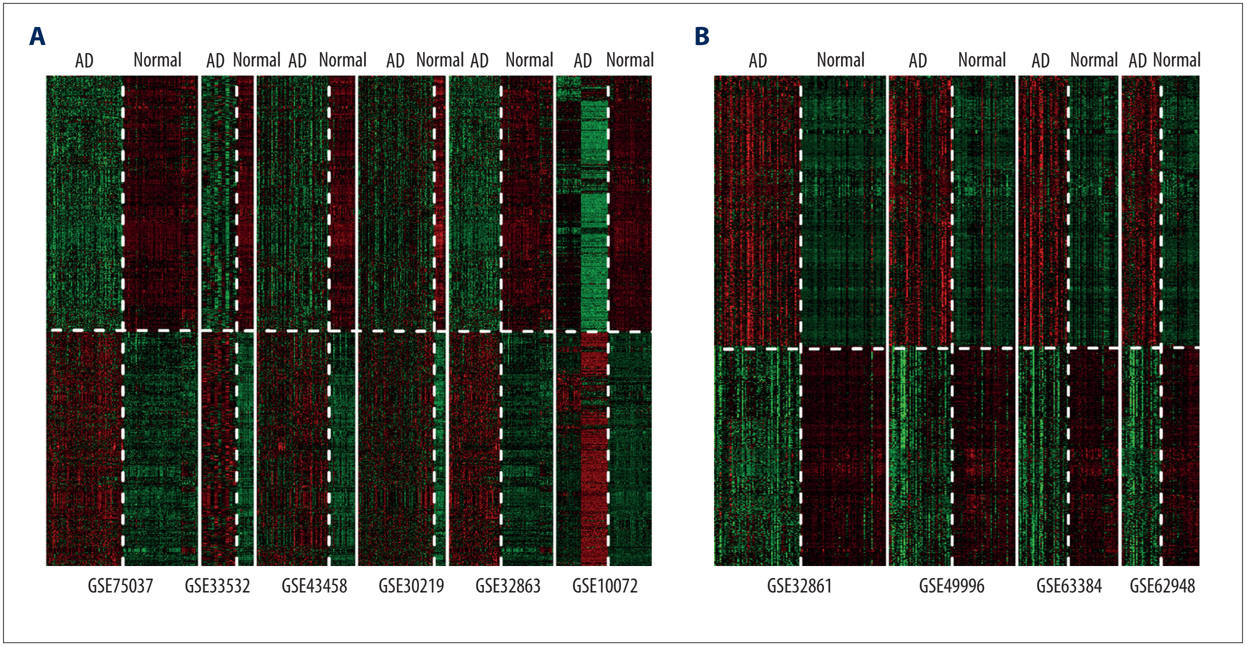


Figure 2. Heatmaps of (**A**) significant differentially expressed genes and (**B**) differentially methylated genes obtained based on MetaDE screening.

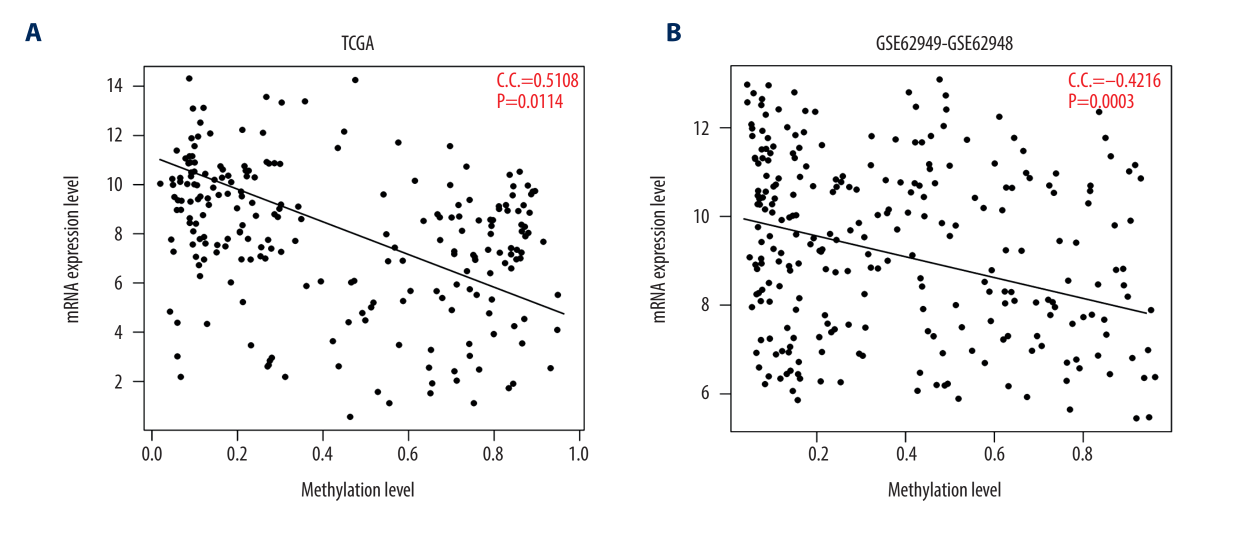


Figure 3. Correlation analysis of expression levels and methylation levels of 265 genes in (**A**) TCGA and (**B**) the GSE62950 dataset. The horizontal axis represents the gene expression level, the vertical axis represents the gene methylation level, the oblique line represents the trend line synthesized by points, and the red font represents the correlation coefficient (CC) and the significant P value.

groups, according to median PI. Then the correlation between the risk prediction model and prognosis was assessed with construction of a Kaplan-Meier survival curve [28] in the survival package of R3.4.1 and validated using the validation dataset.

In addition, following the same method, a risk prediction model was constructed using the clinical factors and the Cox regression analysis-generated prognosis associated with those factors. Similarly, the samples in the training set were divided into high- and low-risk groups, and the correlation between the risk prediction model and the prognosis was assessed through a Kaplan-Meier survival curve.

Finally, a risk prediction model that synthesized clinical factors and tumor marker genes was constructed based on the prognosis correlation coefficients obtained from the 2 models

Table 4. Functional enrichment analysis results for 265 candidate genes.

Category	Term	Count	P value	Genes
Biologic process	GO: 0032409 ~ regulation of transporter activity	6	0.0002	PLCG2, NDFIP1, PKD2, FKBP1B, NKX2-5, SYNGR3
	GO: 0009611 ~ response to wounding	21	0.0005	PPARA, A2M, ACHE, BMP2, UCN, FOXA2, EFEMP2, ATRN, CHST2, HOXB13, SERPING1, CD40, TNFRSF1B, THBD, PLSCR4, CTGF, PLA2G7, LTA4H, CFD, PLAU, ACVR1
	GO: 0050777 ~ negative regulation of immune response	5	0.0007	A2M, IL27RA, NDFIP1, CTLA4, SERPING1
	GO: 0048585 ~ negative regulation of response to stimulus	8	0.0013	PPARA, A2M, TNFRSF1B, IL27RA, NDFIP1, CTLA4, SERPING1, NT5E
	GO: 0015718 ~ monocarboxylic acid transport	6	0.0013	SLC16A3, SLC25A20, PPARA, SLC16A1, PLA2G1B, SLCO2A1
	GO: 0055082 ~ cellular chemical homeostasis	16	0.0016	FXYD1, TRPM8, IL6ST, NDFIP1, TP53, FZD2, FKBP1B, CKB, GCKR, PLCG2, CLDN1, PKD2, RGN, SV2A, KCNH2, KCNQ1
	GO: 0050878 ~ regulation of body fluid levels	9	0.0023	SCT, UCN, THBD, PLSCR4, FOXA2, EFEMP2, SERPING1, CD40, PLAU
	GO: 0006869 ~ lipid transport	9	0.0028	SLC25A20, PPARA, OSBPL3, SORL1, LIPG, PLA2G1B, VPS4B, VLDLR, SLCO2A1
	GO: 0031348 ~ negative regulation of defense response	5	0.0028	A2M, TNFRSF1B, NDFIP1, SERPING1, NT5E
	GO: 0050801 ~ ion homeostasis	16	0.0033	FXYD1, TRPM8, IL6ST, NDFIP1, TP53, FZD2, CPS1, FKBP1B, CKB, PLCG2, CLDN1, PKD2, RGN, SV2A, KCNH2, KCNQ1
	GO: 0035295 ~ tube development	11	0.0036	BMP2, FOXA2, CTGF, CRISPLD2, TGFBR1, HOXB13, PCSK5, NKX2-5, HECA, MYCN, ACVR1
	GO: 0006873 ~ cellular ion homeostasis	15	0.0037	FXYD1, TRPM8, IL6ST, NDFIP1, TP53, FZD2, FKBP1B, CKB, PLCG2, CLDN1, PKD2, RGN, SV2A, KCNH2, KCNQ1
	GO: 0010876 ~ lipid localization	9	0.0045	SLC25A20, PPARA, OSBPL3, SORL1, LIPG, PLA2G1B, VPS4B, VLDLR, SLCO2A1
	GO: 0019725 ~ cellular homeostasis	17	0.0046	FXYD1, TRPM8, PDIA2, IL6ST, NDFIP1, TP53, FZD2, FKBP1B, CKB, GCKR, PLCG2, CLDN1, PKD2, RGN, SV2A, KCNH2, KCNQ1
	GO: 0048878 ~ chemical homeostasis	18	0.0049	FXYD1, TRPM8, IL6ST, NDFIP1, TP53, FZD2, CPS1, FKBP1B, CKB, GCKR, PLCG2, LIPG, CLDN1, PKD2, RGN, SV2A, KCNH2,
KEGG pathway	hsa00562: Inositol phosphate metabolism	5	0.0017	ISYNA1, PLCG2, SYNJ2, ITPKB, INPP5A
	hsa04610: Complement and coagulation cascades	5	0.0037	A2M, THBD, SERPING1, CFD, PLAU
	hsa04070: Phosphatidylinositol signaling system	5	0.0046	PLCG2, SYNJ2, ITPKB, CALML5, INPP5A
	hsa00532: Chondroitin sulfate biosynthesis	3	0.0060	B3GAT1, XYLT1, CHSY1

Table 4 comtinued. Functional enrichment analysis results for 265 candidate genes.

Category	Term	Count	P value	Genes
KEGG pathway (continued)	hsa05217: Basal cell carcinoma	4	0.0393	BMP2, TP53, WNT11, FZD2
	hsa00534: Heparan sulfate biosynthesis	3	0.0081	B3GAT1, XYLT1, HS3ST1
	hsa00590: Arachidonic acid metabolism	4	0.0082	AKR1C3, CYP2C18, PLA2G1B, LTA4H
	hsa04514: Cell adhesion molecules (CAMs)	6	0.0093	NRCAM, CDH15, CLDN1, CTLA4, CD40, SDC2
	hsa00340: Histidine metabolism	3	0.0098	HDC, LCMT2, MAOB

KEGG – Kyoto Encyclopedia of Genes and Genomes.

Table 5. Tumor marker genes significantly associated with prognosis.

Gene	Coefficient	Hazard ratio	Lower.95	Upper.95	P value
EFNB2	0.7121	2.0384	1.5210	2.7317	<0.0001
TSPAN7	–0.5824	0.5586	0.4380	0.7123	<0.0001
INPP5A	–1.4730	0.2292	0.1103	0.4762	<0.0001
VAMP2	1.4277	4.1690	2.0004	8.6885	0.0001
CALML5	0.2006	1.2221	1.0996	1.3582	0.0002
SNAI2	0.5449	1.7245	1.2434	2.3916	0.0011
RHOBTB1	0.6348	1.8867	1.2467	2.8552	0.0027
CKB	–0.3511	0.7039	0.5578	0.8884	0.0031
ATF7IP2	–0.4666	0.6272	0.4299	0.9149	0.0155
RIMS2	0.1523	1.1645	1.0227	1.3259	0.0215
RCBTB2	–0.6106	0.5430	0.3189	0.9247	0.0246
YBX1	0.7766	2.1740	1.0909	4.3325	0.0273
RAB27B	0.2554	1.2909	1.0276	1.6218	0.0283
NFATC1	–0.5289	0.5892	0.3660	0.9487	0.0295
TCEAL4	–0.6401	0.5272	0.2933	0.9476	0.0324
SLC16A3	–0.4125	0.6620	0.4520	0.9696	0.0341

previously described. The PI of each sample was recalculated, the median of which was used to divide the samples in training set into high- and low-risk groups. The correlation between the risk prediction model and prognosis was evaluated via a Kaplan-Meier survival curve, and validated with the validation dataset.

Results

Quality control and differential expression analysis of data used in the meta-analysis

After normalization, quality control was performed on the datasets with MetaQC. Five parameter scores were calculated, including internal quality control (IQC), external quality control (EQC), accuracy quality control (AQCg), consistency quality control (CQCg), and standardized mean rank score (SMR), as shown in Table 3. In addition, results of principal component analysis

Table 6. Univariate and multivariate Cox regression analyses of clinical factors.

Clinical characteristics	Univariate Cox regression		Multivariate Cox regression	
	P value	HR (95%CI)	P value	HR (95%CI)
Age (above/below median, 65 years)	0.4370	1.155 (0.804~1.659)	–	–
Sex (Male/Female)	0.7450	1.062 (0.741~1.52)	–	–
Pathologic M (M0/M1)	0.1310	1.692 (0.848~3.378)	–	–
Targeted molecular therapy (yes/no)	0.1601	1.366 (0.883~ 2.114)	–	v
Tobacco smoking history (current/reformed/never)	0.9900	1.002 (0.737~1.362)	–	–
Radiation therapy (yes/no)	0.0035	2.033 (1.25~3.307)	0.5924	1.163 (0.669~2.019)
Pathologic N (N0/N1/N2)	<0.0001	1.85 (1.494~2.29)	0.0471	1.439 (1.005~2.060)
Pathologic T (T1/T2/T3/T4)	0.0002	1.537 (1.223~1.932)	0.0169	1.236 (0.914~1.672)
Pathologic stage (I/II/III/IV)	<0.0001	1.671 (1.413~1.976)	0.0103	1.279 (0.952~1.718)
New tumor (yes/no)	<0.0001	2.362 (1.535~3.634)	0.0001	2.395 (1.533~3.742)

HR – hazard ratio.

of these datasets are shown in Figure 1A and 1B. After combining Table 3 and Figure 1, we concluded that the distribution of 6 expression profiling datasets and 4 methylation profiling datasets was balanced, and all indexes fit the standard of data quality, so the 10 datasets were included in the subsequent analysis. Finally, 1975 significant DEGs and 2095 DMGs were identified using MetaDE. A heatmap of these DEGs and DMGs showed that that the DEGs and DMGs screened from different datasets were consistent in their differential degree and direction (Figure 2).

Correlation analysis between gene expression level and methylation level

By comparing the 1975 DEGs and 2095 DMGs, 265 intersecting genes (candidate genes) were identified. An analysis of the correlation between the expression and methylation levels of the 265 candidate genes then was performed, based on the methylation and expression profiles that matched the samples in TCGA and GSE62950 datasets. As shown in Figure 3, the expression values and methylation levels of 265 gene were negatively correlated in the TCGA and GSE62950 datasets, and the correlation coefficients were –0.5108 (P=0.0114) and –0.4216 (P=0.0003), respectively. Functional enrichment analysis of 265 candidate genes identified 15 significant GO biological processes and 9 KEGG pathways, as shown in Table 4.

Screening of prognosis-related tumor marker genes and clinical factors

From an initial pool of 256 candidate genes and based on the clinical factors in the samples, 16 prognosis-related genes (*EFNB2*, *TSPAN7*, *INPP5A*, *VAMP2*, *CALML5*, *SNAI2*, *RHOBTB1*, *CKB*, *ATF7IP2*, *RIMS2*, *RCBTB2*, *YBX1*, *RAB27B*, *NFATC1*, *TCEAL4*, and *SLC16A3*) (Table 5) were screened using univariate and multivariate Cox regression analyses. An analysis then was performed of the correlation between the expression and methylation levels in 16 prognostic genes in TCGA and GSE62950 datasets (Supplementary Figure 1). Five clinical factors were identified: pathologic N (nodes), pathologic T (tumor), pathologic stage, new tumor, and radiation therapy. As shown in Table 6, pathologic N, pathologic T, pathologic stage, and new tumor were significantly correlated with prognosis. The Kaplan-Meier curves for the correlations between the 4 clinical factors and overall survival (OS) are shown in Supplementary Figure 2. A cluster analysis of the expression and methylation levels of the 16 prognosis-related genes and the 4 clinical factors revealed that the samples could be divided into 2 clusters. There were 160 and 175 samples in clusters 1 and 2, respectively (Figure 4). In addition, a chi-square test of sample clinical information in the 2 clusters revealed that pathologic N was significantly correlated with both clusters (P=0.0467) (Supplementary Table 1).

Construction and validation of the risk prediction model

A risk prediction model was constructed using the 16 prognosis-associated tumor marker genes identified by the Cox regression analysis. A Kaplan-Meier survival curve was used on the TCGA training set to assess the correlation between the risk groups and the prognosis for OS and recurrence. In OS prognosis, low-risk patients (167 samples) had a longer OS time compared with high-risk patients (168 samples) (Table 7). The

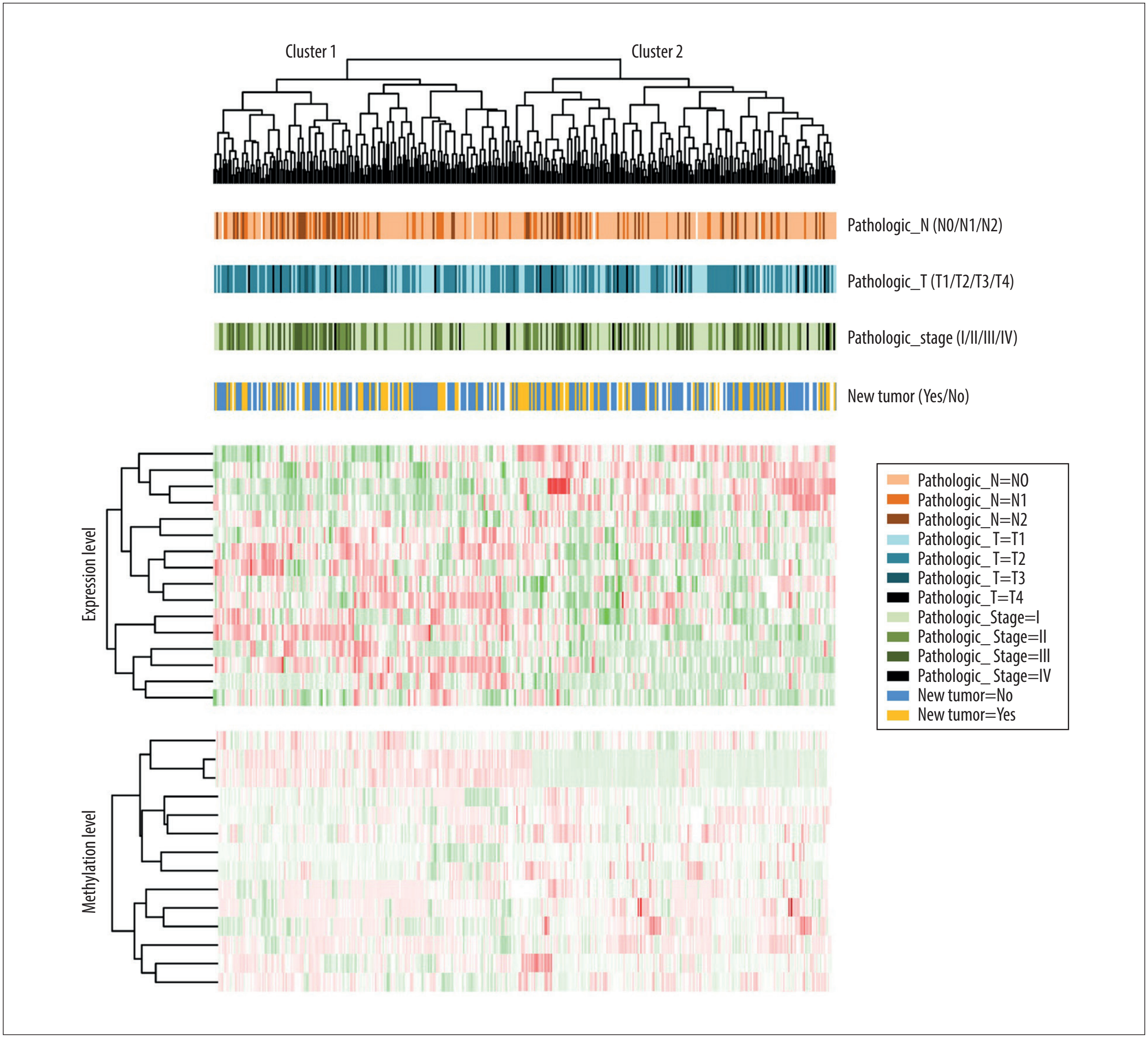


Figure 4. Bidirectional hierarchical cluster heatmaps based on 16 gene expression and methylation levels. The first line under the cluster tree represents pathologic N information, and the change from light orange to deep orange represents N0 to N2. The second line represents the pathologic T information, and the change from light blue to dark blue represents T1 to T4. The third line represents pathologic stage information, and the change from light green to dark green represents stages I to IV. The fourth line represents new tumor information, and the blue and gold represent the samples without and with new tumor, respectively.

P value for the correlation between the risk groups and OS prognosis was 3.961e-08. The Kaplan-Meier curve is shown on the left side of Figure 5A.

Based on the PI values, a receiver operating characteristic (ROC) curve was constructed. The area under the ROC curve (AUROC) for prognosis was 0.997 (Figure 5C; green curve). In the analysis of recurrence prognosis (260 samples), low-risk patients (130 samples) also had a longer time to relapse relative to the high-risk patients (130 samples) (Table 7). The P value for the correlation between the risk groups and prognosis for recurrence-free survival (RFS) was 3.961e-08 (Figure 5A; right) and the AUROC of the ROC curve was 0.985 (Figure 5C; blue curve).

The risk prediction model was validated in GSE37745 and the results were consistent with that in the training set. As shown in Figure 5B, the P value for the correlation between the risk groups and the prognosis for OS was 0.0091 (left) and between the risk groups and prognosis for recurrence was 0.0260 (right). The AUROC of ROC curve for OS and relapse prognoses were 0.979 (black curve) and 0.953 (red

Table 7. Prognostic time for different risk classification models of the TCGA and GSE37745 dataset.

		Overall survival time (months, mean±SD)		Recurrence-free survival time (months, mean±SD)	
		Low-risk	High-risk	Low-risk	High-risk
TCGA	Gene expression model	33.64±37.17	21.41±17.71	28.18±35.04	15.79±13.97
	Clinic factor model	29.03±35.46	27.01±27.76	25.74±32.68	19.32±22.79
	Combined model	33.55±40.09	22.33±17.78	28.33±35.87	16.31±14.57
GSE37745	Gene expression model	68.66±47.08	53.69±52.46	70.53±57.47	37.06±43.82
	Clinic factor model	84.51±62.38	54.51±50.67	62.53±55.77	30.66±39.29
	Combined model	84.50±62.38	54.51±50.69	62.53±55.77	30.66±39.29

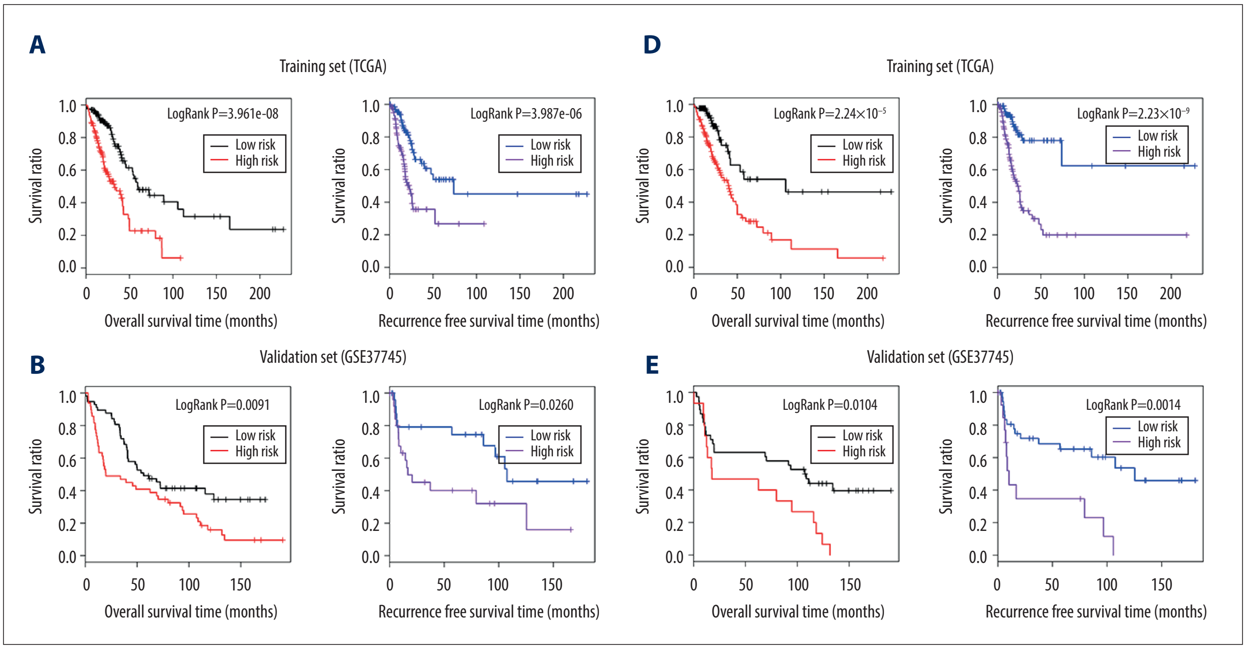
TCGA – The Cancer Genome Atlas.

curve), respectively (Figure 5C). Using the same methods, a risk prediction mode was constructed using the clinical factors (Figure 5D–5F) and both the tumor marker genes and the clinical factors (Figure 5G–5I). The OS and RFS for high- and low-risk groups are shown in Table 7.

Discussion

The present study integrated multiple LACC gene expression and methylation profile datasets and used meta-analysis to preliminarily screen out 265 genes whose expression levels were significantly influenced by methylation. Then 16 prognosis-related genes (*EFNB2*, *TSPAN7*, *INPP5A*, *VAMP2*, *CALML5*, *SNAI2*, *RHOBTB1*, *CKB*, *ATF7IP2*, *RIMS2*, *RCBTB2*, *YBX1*, *RAB27B*, *NFATC1*, *TCEAL4*, and *SLC16A3*) were elected using Cox regression analysis, which was then used successfully to construct a prognostic risk prediction model. In addition, we constructed a risk prediction model based on 4 clinical factors: pathologic N, pathologic T, pathologic stage, and new tumor. Finally, a comprehensive prognostic risk model that combined tumor marker genes and clinical factors was constructed and validated.

Of the 16 tumor marker genes, both calmodulin like 5 (*CALML5*) and inositol polyphosphate-5-phosphatase a (*INPP5A*) were involved in the hsa04070: phosphatidylinositol signaling system. Signaling by phosphorylated species of phosphatidylinositol regulates various cellular processes, such as cytoskeletal reorganization, membrane trafficking, and sex-dependent synaptic patterning [29,30]. Phosphatidylinositol 3-kinase (PI3K)



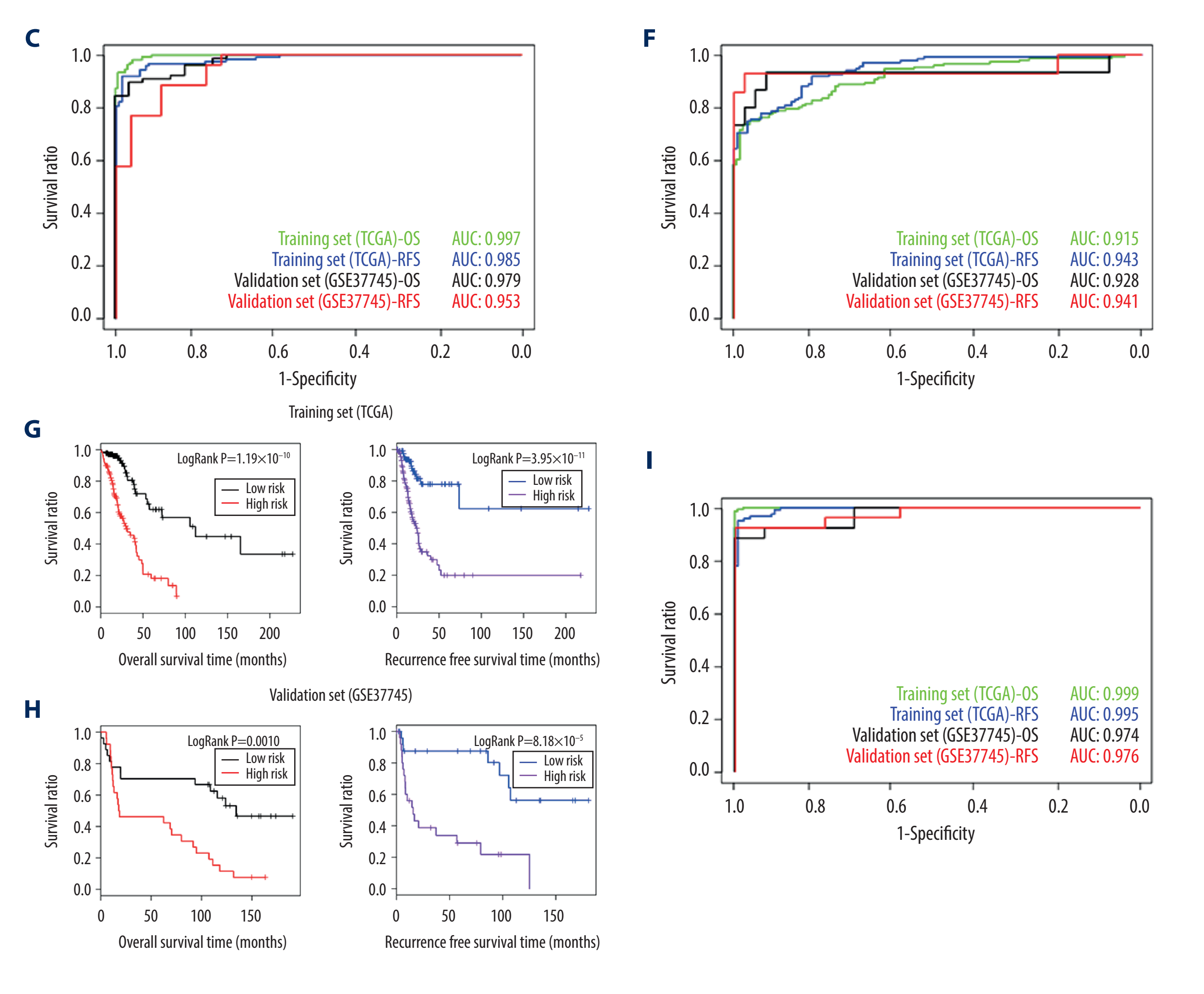


Figure 5. (**A**) The Kaplan-Meier curves for the risk prediction model based on tumor marker genes and OS prognosis (**left**) and recurrence prognosis (**right**) in TCGA training set. (**B**) The Kaplan-Meier curves for the risk prediction model based on tumor marker genes and OS prognosis (**left**) and recurrence prognosis (**right**) in the GSE37745 validation set. (**C**) AUROC curves for the prognosis prediction model and OS prognosis and recurrence prognosis in TCGA training set and the GSE37745 verification set. (**D**) The Kaplan-Meier curves for the risk prediction model based on clinical factors and OS prognosis (**left**) and recurrence prognosis (**right**) in TCGA training set. (**E**) The Kaplan-Meier curves for the risk prediction model based on clinical factors and OS prognosis (**left**) and recurrence prognosis (**right**) in the GSE37745 validation set. (**F**) The AUROC curves for the prognosis prediction model and OS prognosis and recurrence prognosis in TCGA training set and the GSE37745 verification set. (**G**) The Kaplan-Meier curves for the risk prediction model based on tumor marker genes combined with clinical factors and OS prognosis (**left**) and recurrence prognosis (**right**) in TCGA training set. (**H**) The Kaplan-Meier curves for the risk prediction model based on tumor marker genes combined with clinical factors and OS prognosis (**left**) and recurrence prognosis (**right**) in the GSE37745 validation set. (**I**) The AUROC curves for the prognosis prediction model and OS prognosis and recurrence prognosis in TCGA training set and the GSE37745 verification set. The green and blue curves in (**C, F, I**) represent the AUROC curves for OS prognosis and recurrence prognosis in TCGA and the black and red curves represent the AUROC curves of OS prognosis and recurrence prognosis in the GSE37745 verification set.

signaling is the most common phosphatidylinositol signaling in cancers, including those of the lung [31,32]. Specifically, INPP5A recently has been reported to be a prognostic marker for cutaneous squamous cell carcinoma [33]. In addition to *CALML5* and *INPP5A*, creatine kinase B (*CKB*) and solute carrier family 16 member 3 (*SLC16A3*) were also identified in function enrichment analysis. *CKB* was enriched in ion homeostasis-associated functions and *SLC16A3* was enriched in function associated with monocarboxylic acid transport. CKB is an enzyme involved in energy transduction pathways, and levels of it are low in colorectal cancer [34]. A recent study revealed that quantification of DNA methylation of specific CpG sites in the *SLC16A3* promoter had clinical potential for diagnosing and predicting prognosis of clear cell renal cell carcinoma [35]. Those findings, taken together with our results, lead us to speculate that *CALML5*, *INPP5A*, *CKB* and *SLC16A3* may be involved in the progression of LACC through different pathways, and they may serve as important markers of diagnosis and prognosis in LACC.

Among the 16 marker genes, ephrin B2 (*EFNB2*), tetraspanin 7 (*TSPAN7*), *INPP5A*, vesicle-associated membrane protein 2 (*VAMP2*), and *CALML5* had the lowest P values. *EFNB2* is a member of the ephrin family. The ephrin system is implicated in many cellular processes, such as cell proliferation, differentiation, and migration, as well as physiological or pathological angiogenesis [36]. It is also implicated in human cancers through autocrine or juxtacrine activation [37]. Coexpression of *EFNB2* and its receptor, Ephrin type-B receptor 4, has been reported in papillary thyroid carcinoma, glioblastoma, and uterine cervical and ovarian cancers [38–41]. Recently, Oweida et al. [42] suggested that overexpression of *EFNB2* can serve as a biomarker for patient prognosis. *TSPAN7*, a member of the transmembrane 4 superfamily, has been implicated in the development and progression of several cancers. It was first found to be strongly expressed in T-cell acute lymphoblastic leukemia [43]. Subsequent microarray analyses have demonstrated overexpression of *TSPAN7* in several solid tumors [44,45]. Research on the role of *TSPAN7* in LACC is rare. *VAMP2* is a member of the vesicle-associated membrane protein. The *VAMP2-NRG1* fusion gene promotes anchorage-independent colony formation of LACC cells, serving as a novel oncogenic driver of LACC [46]. Recently, Wang et al. demonstrated that miR-493-5p overexpression promotes cell apoptosis and inhibits the proliferation and migration of liver cancer cells by negatively regulating the expression of *VAMP2* [47], which indirectly indicates the important role that *VAMP2* plays in cancer. Taken together, all of these studies suggest that *EFNB2*, *TSPAN7* and *VAMP2*, may serve as prognostic makers in LACC.

Most of the other tumor marker genes we identified have been reported to be implicated in lung cancer or other human cancers. For instance, Sail family transcriptional repressor 2 (*SNAI2*) encodes a zinc-finger protein from the SNAI family of transcription factors [48]. *SNAI2* is amplified or interacts with specific oncogenes in many human cancers, including lung cancer [49,50]. *SNAI2* expression by cancer-associated fibroblasts is correlated with worse OS in NSCLC [51]. Rho-related BTB domain containing 1 (*RHOBTB1*), which belongs to the RhoBTB subfamily, has been proposed as a tumor suppressor [52]. Y-box binding protein-1 (*YBX1*) is upregulated in various cancers, including lung cancer, and serves as a new marker of lung cancer progression [53]. Hendrix et al. [54] found that *RAB27B*, a member of *RAS* oncogene family, regulates invasive tumor growth and metastasis of several breast cancer cell lines. Nuclear factor of activated T cells 1 (*NFATC1*) regulates many cancer-related functions, such as cell proliferation, migration, and angiogenesis. It also acts as an oncogene involved in some functions in cancer and induces a tumorigenic microenvironment [55]. Transcription elongation factor A (SII)-like 4 (*TCEAL4*) is downregulated in anaplastic thyroid cancer [56]. Therefore, these genes may have roles as key biomarkers in LACC.

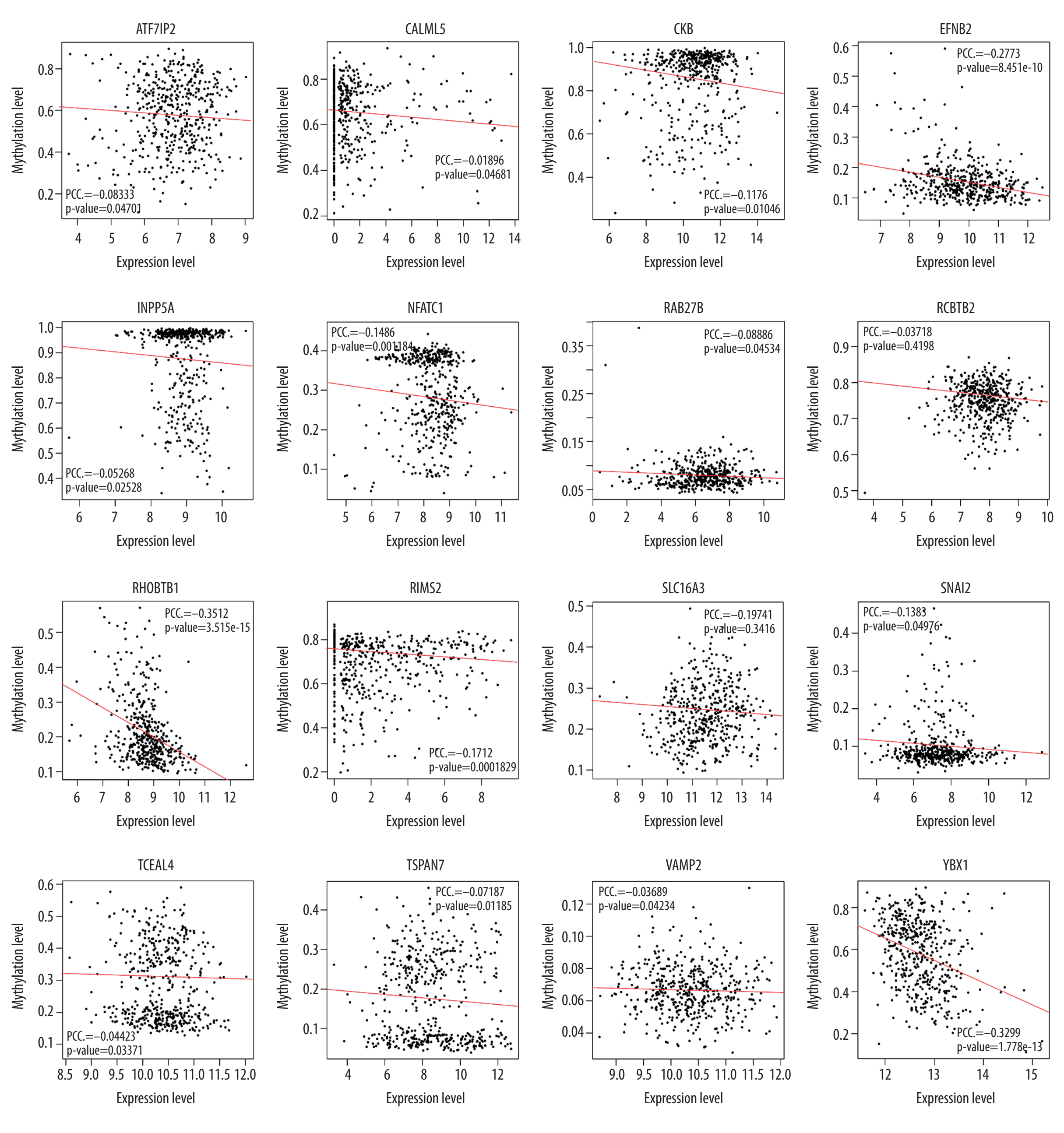
Conclusions

In the present study, we identified 16 tumor marker genes for LACC, based on analysis of multiple gene expression and methylation profiling datasets, and constructed an integrated risk prediction model that combined those tumor markers with clinical factors. The 16 genes we identified, such as *EFNB2*, *TSPAN7*, *INPP5A*, *VAMP2*, and *CALML5*, may serve as novel biomarkers in early diagnosis and prediction of prognosis of LACC.

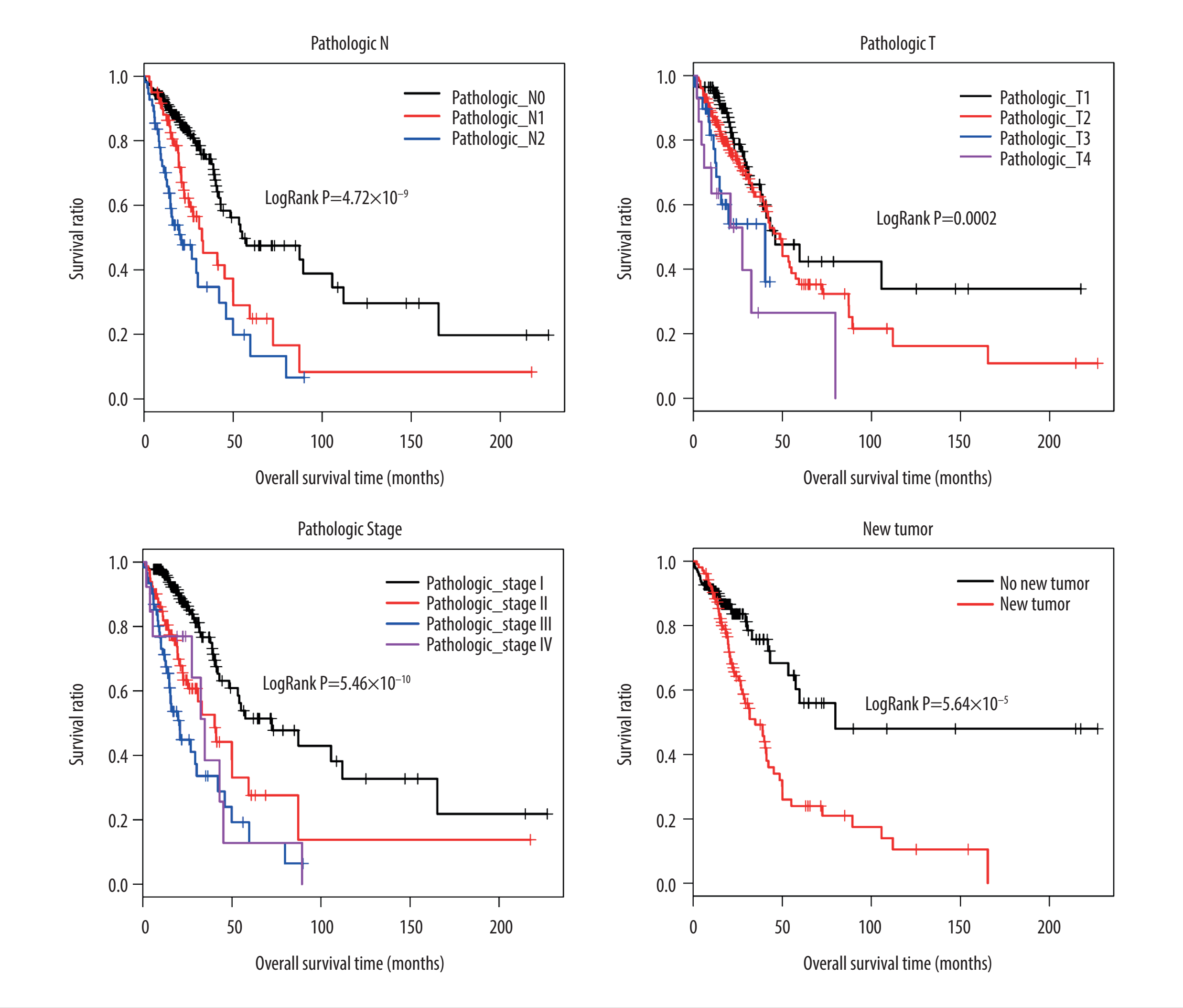
Conflict of interests

None.

Supplementary Data



Supplementary Figure 1. Analysis of the correlation between expression and methylation levels for 16 prognostic genes in TCGA and the GSE62950 dataset.



Supplementary Figure 2. The Kaplan-Meier curves for the correlations between the 4 clinical factors (pathologic N, pathologic T, pathologic stage, and new tumor) and overall survival.

Supplementary Table 1. Clinical and chi-square test information for samples in clusters 1 and 2.

Clinical characteristics	Cluster 1				Cluster 2				X-squared	P value
Pathologic N (N0/N1/N2)	93	36	28	–	121	24	27	–	5.4091	0.0467
Pathologic T (T1/T2/T3/T4)	48	93	14	5	63	87	15	9	2.8225	0.4198
Pathologic stage (I/II/III/IV)	82	42	31	5	98	39	30	8	1.5735	0.6654
New tumor (yes/no)	47	84	–	–	57	92	–	–	0.0823	0.7742

References:

1. Siegel R, Desantis C, Jemal A: Colorectal cancer statistics, 2014. Cancer J Clin, 2014; 64: 104–17
2. Zhou C: Lung cancer molecular epidemiology in China: Recent trends. Transl Lung Cancer Res, 2014; 3: 270–79
3. Imielinski M, Berger AH, Hammerman PS et al: Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell, 2012; 150: 1107–20
4. Goodgame B, Viswanathan A, Miller CR et al: A clinical model to estimate recurrence risk in resected stage I non-small cell lung cancer. Am J Clin Oncol, 2008; 31: 22–28
5. Vari S, Pilotto S, Maugeri-Saccà M et al: Advances towards the design and development of personalized non-small-cell lung cancer drug therapy. Expert Opin Drug Discov, 2013; 8: 1381–97
6. Somaiah N, Simon NG, Simon GR: A tabulated summary of targeted and biologic therapies for non-small-cell lung cancer. J Thorac Oncol, 2012; 7: S342–68
7. Wigle D, Jurisica I, Radulovich N et al: Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. Cancer Res, 2002; 62: 3005–8
8. Powell CA, Spira A, Derti A et al: Gene expression in lung adenocarcinomas of smokers and nonsmokers. Am J Respir Cell Mol Biol, 2003; 29: 157–62
9. Jiang H, Deng Y, Chen H et al: Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. BMC Bioinformatics, 2004; 5: 81
10. Beer D, Kardia S, Huang C et al: Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med, 2002; 8: 816–24
11. Herbst RS, Yano S, Kuniyasu H et al: Differential expression of E-cadherin and type IV collagenase genes predicts outcome in patients with stage I non-small cell lung carcinoma. Clin Cancer Res, 2000; 6: 790–97
12. Schneider P, Praeuer H, Stoeltzing O et al: Multiple molecular marker testing (p53, C-Ki-ras, c-erbB-2) improves estimation of prognosis in potentially curative resected non-small cell lung cancer. Br J Cancer, 2000; 83: 473–79
13. Lu Y, Lemon W, Liu PY et al: A gene expression signature predicts survival of patients with stage I non-small cell lung cancer. PLoS Med, 2006; 3: e467
14. Chen HY, Yu SL, Chen CH et al: A five-gene signature and clinical outcome in non-small-cell lung cancer. N Engl J Med, 2007; 356: 11–20
15. Potti A, Mukherjee S, Petersen R et al: A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. N Engl J Med, 2006; 355: 570–80
16. Slebos R, Kibbelaar R, Dalesio O et al: K-ras oncogene activation as a prognostic marker in adenocarcinoma of the lung. N Engl J Med, 1990; 323: 561–65
17. Horio Y, Takahashi T, Kuroishi T et al: Prognostic significance of p53 mutations and 3p deletions in primary resected non-small cell lung cancer. Cancer Res, 1993; 53: 1–4
18. Beer DG, Kardia SL, Huang C-C et al: Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med, 2002; 8: 816
19. Botling J, Edlund K, Lohr M et al: Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis, and tissue microarray validation. Clin Cancer Res, 2013; 19: 194–204
20. Parrish R, Spencer H: Effect of normalization on significance testing for oligonucleotide microarrays. J Biopharm Stat, 2004; 14: 575–89
21. Ritchie M, Phipson B, Wu D et al: limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res, 2015; 43: e47
22. Turan N, Ghalwash M, Katari S et al: DNA methylation differences at growth related genes correlate with birth weight: A molecular signature linked to developmental origins of adult disease? BMC Med Genomics, 2012; 5: 10
23. Kang D, Sibille E, Kaminski N, Tseng G: MetaQC: Objective quality control and inclusion/exclusion criteria for genomic meta-analysis. Nucleic Acids Res, 2012; 40: e15
24. Chang L, Lin H, Sibille E, Tseng G: Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. BMC Bioinformatics, 2013; 14: 368
25. Huang dW, Sherman B, Lempicki R: Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res, 2009; 37: 1–13
26. Huang dW, Sherman B, Lempicki R: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc, 2009; 4: 44–57
27. Wang P, Wang Y, Hang B et al: A novel gene expression-based prognostic scoring system to predict survival in gastric cancer. Oncotarget, 2016; 7: 55343–51
28. Goel M, Khanna P, Kishore J: Understanding survival analysis: Kaplan-Meier estimate. Int J Ayurveda Res, 2010; 1: 274–78
29. Toker A, Cantley LC: Signalling through the lipid products of phosphoinositide-3-OH kinase. Nature, 1997; 387: 673–76
30. Schwarz JM, Liang S-L, Thompson SM, McCarthy MM: Estradiol induces hypothalamic dendritic spines by enhancing glutamate release: A mechanism for organizational sex differences. Neuron, 2008; 58: 584–98
31. Krystal GW, Sulanke G, Litz J: Inhibition of phosphatidylinositol 3-kinase-Akt signaling blocks growth, promotes apoptosis, and enhances sensitivity of small cell lung cancer cells to chemotherapy. Mol Cancer Ther, 2002; 1: 913–22
32. Yip PY: Phosphatidylinositol 3-kinase-AKT-mammalian target of rapamycin (PI3K-Akt-mTOR) signaling pathway in non-small cell lung cancer. Transl Lung Cancer Res, 2015; 4: 165
33. Liang HJ, DiCaudo DJ, Schmidt JE et al: INPP5a expression as a prognostic marker in cutaneous squamous cell carcinoma (cSCC). J Clin Oncl, 2018; 36(15): e21567
34. Friedman DB, Hill S, Keller JW et al: Proteome analysis of human colon cancer by two-dimensional difference gel electrophoresis and mass spectrometry. Proteomics, 2004; 4: 793–811
35. Fisel P, Kruck S, Winter S et al: DNA Methylation of the SLC16A3 promoter regulates expression of the human lactate transporter MCT4 in renal cancer with consequences for clinical outcome. Clin Cancer Res, 2013; 19: 5170–81
36. Li X, Song C, Huang G et al: The coexpression of EphB4 and EphrinB2 is associated with poor prognosis in HER2-positive breast cancer. Oncotargets Ther, 2017; 10: 1735
37. Tang XX, Brodeur GM, Campling BG, Ikegaki N: Coexpression of transcripts encoding EPHB receptor protein tyrosine kinases and their ephrin-B ligands in human small cell lung carcinoma. Clin Cancer Res, 1999; 5: 455–60
38. Sharma GK, Dhillon VK, Masood R, Maceri DR: Overexpression of EphB4, EphrinB2, and epidermal growth factor receptor in papillary thyroid carcinoma: A pilot study. Head Neck, 2015; 37: 964–69
39. Alam SM, Fujimoto J, Jahan I et al: Overexpression of ephrinB2 and EphB4 in tumor advancement of uterine endometrial cancers. Ann Oncol, 2006; 18: 485–90
40. Tu Y, He S, Fu J et al: Expression of EphrinB2 and EphB4 in glioma tissues correlated to the progression of glioma and the prognosis of glioblastoma patients. Clin Translational Oncol, 2012; 14: 214–20
41. Alam SM, Fujimoto J, Jahan I et al: Coexpression of EphB4 and ephrinB2 in tumor advancement of uterine cervical cancers. Gynecol Oncol, 2009; 114: 84–88
42. Oweida A, Bhatia S, Hirsch K et al: Ephrin-B2 overexpression predicts for poor prognosis and response to therapy in solid tumors. Mol Carcinog, 2017; 56: 1189–96
43. Takagi S, Fujikawa K, Imai T et al: Identification of a highly specific surface marker of T-cell acute lymphoblastic leukemia and neuroblastoma as a new member of the transmembrane 4 superfamily. Int J Cancer, 1995; 61: 706–15
44. Chakraborty S: *In silico* analysis identifies genes common between five primary gastrointestinal cancer sites with potential clinical applications. Ann Gastroenterol, 2014; 27: 231–36
45. Wuttig D, Zastrow S, Füssel S et al: CD31, EDNRB and TSPAN7 are promising prognostic markers in clear-cell renal cell carcinoma revealed by genome-wide expression analyses of primary tumors and metastases. Int J Cancer, 2012; 131: E693–704
46. Jung Y, Yong S, Kim P et al: VAMP2–NRG1 fusion gene is a novel oncogenic driver of non-small-cell lung adenocarcinoma. J Thorac Oncol, 2015; 10: 1107–11

47. Wang G, Fang X, Han M et al: MicroRNA-493-5p promotes apoptosis and suppresses proliferation and invasion in liver cancer cells by targeting VAMP2. Int J Mol Med, 2018; 41: 1740–48

48. Casas E, Kim J, Bendesky A et al: Snail2 is an essential mediator of Twist1-induced epithelial mesenchymal transition and metastasis. Cancer Res, 2011; 71: 245–54

49. Hemavathy K, Ashraf SI, Ip YT: Snail/slug family of repressors: Slowly going into the fast lane of development and cancer. Gene, 2000; 257: 1–12

50. Atmaca A, Wirtz RW, Werner D et al: SNAI2/SLUG and estrogen receptor mRNA expression are inversely correlated and prognostic of patient outcome in metastatic non-small cell lung cancer. BMC Cancer, 2015; 15: 1–7

51. Andriani F, Leone G, Landoni E et al: SNAI2 expression by cancer-associated fibroblasts is a negative prognostic factor in non-small cell lung cancer. Cancer Res, 2014; 74: 2852

52. Xu RS, Wu XD, Zhang SQ et al: The tumor suppressor gene RhoBTB1 is a novel target of miR-31 in human colon cancer. Int J Oncol, 2013; 42: 676–82

53. Gessner C, Woischwill C, Schumacher A et al: Nuclear YB-1 expression as a negative prognostic marker in nonsmall cell lung cancer. Eur Respir J, 2004; 23: 14

54. Hendrix A, Maynard D, Pauwels P et al: Effect of the secretory small GTPase Rab27B on Breast cancer growth, invasion, and metastasis. J Natl Cancer Inst, 2010; 102: 866

55. Tripathi P, Wang Y, Coussens M et al: Activation of NFAT signaling establishes a tumorigenic microenvironment through cell autonomous and non-cell autonomous mechanisms. Oncogene, 2014; 33: 1840–49

56. Akaishi J, Onda M, Okamoto J et al: Down-regulation of transcription elogation factor A (SII) like 4 (TCEAL4) in anaplastic thyroid cancer. BMC Cancer, 2006; 6: 260

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]