

Identification of key genes associated with lung adenocarcinoma by bioinformatics analysis

Science Progress

2021, Vol. 104(1) 1–18

© The Author(s) 2021

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0036850421997276

journals.sagepub.com/home/sci

Xinyu Wang^{1*}, Jiaojiao Yang^{2*} and Xueren Gao¹ 

¹School of Pharmacy, Yancheng Teachers' University, Yancheng, Jiangsu, China

²Department of Microbiology and Immunology, Shanxi Medical University, Taiyuan, Shanxi, China

Abstract

Lung adenocarcinoma (LUAD) is the most common histological type of lung cancer, comprising around 40% of all lung cancer. Until now, the pathogenesis of LUAD has not been fully elucidated. In the current study, we comprehensively analyzed the dysregulated genes in lung adenocarcinoma by mining public datasets. Two sets of gene expression datasets were obtained from the Gene Expression Omnibus (GEO) database. The dysregulated genes were identified by using the GEO2R online tool, and analyzed by R packages, Cytoscape software, STRING, and GPEIA online tools. A total of 275 common dysregulated genes were identified in two independent datasets, including 54 common up-regulated and 221 common down-regulated genes in LUAD. Gene Ontology (GO) enrichment analysis showed that these dysregulated genes were significantly enriched in 258 biological processes (BPs), 27 cellular components (CCs), and 21 molecular functions (MFs). Furthermore, protein-protein interaction (PPI) network analysis showed that PECAM1, ENG, KLF4, CDH5, and VWF were key genes. Survival analysis indicated that the low expression of ENG was associated with poor overall survival (OS) of LUAD patients. The low expression of PECAM1 was associated with poor OS and recurrence-free survival of LUAD patients. The cox regression model developed based on age, tumor stage, ENG, PECAM1 could effectively predict 5-year survival of LUAD patients. This study revealed some key genes, BPs, CCs, and MFs involved in LUAD, which would provide new insights into understanding the pathogenesis of LUAD. In addition, ENG and PECAM1 might serve as promising prognostic markers in LUAD.

*Xinyu Wang and Jiaojiao Yang contributed equally to the article.

Corresponding author:

Xueren Gao, School of Pharmacy, Yancheng Teachers' University, No. 2 South Road, Hope Avenue, Yancheng, Jiangsu 224007, China.

Email: gxr871230@126.com



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>)

which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

Keywords

Gene, lung adenocarcinoma, bioinformatics, biomarker

Introduction

Lung adenocarcinoma (LUAD) is the most common histological type of lung cancer, comprising around 40% of all lung cancer.¹ Although progress has been made in the diagnosis and treatment of LUAD, the 5-year survival rate for LUAD patients is still very poor, which is partly due to the fact that LUAD is often diagnosed at advanced stages involving disseminated metastatic tumors.² In addition, the pathogenesis of LUAD has not been completely elucidated until now. Therefore, it is necessary to have a better understanding of the molecular mechanism of occurrence and development of LUAD and find specific biomarkers with prognostic significance, which will help to improve the treatment and survival rate of LUAD.

In recent years, the combination of high-throughput gene expression detection technologies (such as RNA-seq and microarray) and bioinformatics has become an effective measure to provide new insights into the occurrence and development of diseases, including LUAD.^{3–5} Zheng et al.³ found that the expression level of SUV39H2 was up-regulated in LUAD tissues through bioinformatics analyses. SUV39H2 overexpression was significantly associated with its amplification and with shorter overall survival of LUAD patients. Subsequently, RNA-seq demonstrated that SUV39H2 might mediate tumorigenesis and metastasis of LUAD by regulating TPM4, OPTN, and STOM. Shi et al.⁴ developed a survival prediction scoring model based on LUAD RNA-seq data in The Cancer Genome Atlas (TCGA) database. The model including 31 lncRNAs could predict overall survival (OS) of LUAD patients with high accuracy. Su et al.⁵ identified key genes (such as CCL2, LYZ, and MMP2) and the essential miRNA, hsa-let-7d, related to LUAD brain metastases via microarray analysis of cDNA expression profiles. Chen et al.⁶ provided profiles of the tumor immune microenvironment that had prognostic value for patients with locally advanced LUAD based on targeted RNA-Seq data and bioinformatics methods. Zhu et al.⁷ revealed an important candidate gene (RN7SL494P) involved in both nodal metastasis and prognosis in LUAD by analyzing RNA-Seq data. These findings enhanced our understanding of genes involved in LUAD, but there were still hidden genes waiting to be revealed because of the complexity and heterogeneity of LUAD.

In the current study, we identified the dysregulated genes in lung adenocarcinoma based on gene expression profiles of 64 LUAD and 64 non-tumor lung tissues. Subsequent bioinformatic analyses were performed to explore biological processes (BPs), cellular components (CCs), molecular functions (MFs) and pathways enriched by the dysregulated genes and key genes involved in LUAD. The related datasets from TCGA and Genotype-Tissue Expression (GTEx) databases were used to confirm the expression levels of key genes in LUAD and assess the relationships between the expression levels of key genes and the survival of LUAD.

patients. Taken together, this study will provide useful guidance for further research on various genes associated with the occurrence and development of LUAD.

Materials and methods

Microarray datasets

Two sets of gene expression datasets (GSE32863 and GSE118370) were obtained from the Gene Expression Omnibus (GEO) database. Samples from GSE32863 included 13 male and 45 female. Thereinto nine patients were younger than 60 years old. Samples from GSE118370 included three male and three female. Thereinto three patients were younger than 60 years old. GSE32863 including 58 LUAD and 58 adjacent non-tumor lung tissues, were generated based on Illumina HumanWG-6 v3.0 expression beadchip (GPL6884) and contributed by Ite Laird-Offringa.⁸ The data had been processed by log2-transformation and Robust Spline Normalization (RSN) using the lumi package in R. GSE118370 including six invasive LUAD tissues and paired normal lung tissues, were generated based on Affymetrix Human Genome U133 Plus 2.0 Array (GPL570) and contributed by Xu et al.⁹ The data had been normalized by MAS 5.0 algorithm.

Identification of the dysregulated genes

GEO2R online tool (<https://www.ncbi.nlm.nih.gov/geo/geo2r/>) was used to identify the dysregulated genes in LUAD. The *t* test and Benjamini-Hochberg method were used to calculate the *p*-value and false discovery rate (FDR), respectively. The dysregulated genes were defined according to adjusted *p*-value (adj.P.Val) < 0.05 and |log2FC| > 1. The common dysregulated genes in GSE32863 and GSE118370 datasets were screened out by Venny 2.1 online tool (<http://bioinfogp.cnb.csic.es/tools/venny/index.html>).

Enrichment analysis of the dysregulated genes

To identify BPs, CCs, MFs and pathways closely related to LUAD, Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of the dysregulated genes were performed using clusterProfiler package in R.¹⁰ The adjusted *p*-value < 0.05 was considered as significant enrichment.

Protein-protein interaction (PPI) network analysis of the dysregulated genes

The PPI network of the common dysregulated genes in the two datasets was constructed using the STRING database (<https://string-db.org/>) with default parameter values (Evidence was selected as meaning of network edges; Textmining, experiments, databases, co-expression, neighborhood, gene fusion and co-occurrence

were selected as active interaction sources; Medium confidence (0.4) was selected as minimum required interaction score) and visualized using Cytoscape software.^{11,12} The degree of a node in PPI network was regarded as the number of interactions with other nodes. Key nodes (proteins/genes) in PPI network were selected according to degree >15.

Validation of the expression levels of key genes

To confirm the expression levels of key genes in LUAD, gene expression profiling interactive analysis (GEPIA) online tool (<http://gepia.cancer-pku.cn/>) was used to explore the related datasets in TCGA and GTEx databases, and analyze the expression levels of key genes in LUAD tissues compared with normal tissues. The method for differential analysis is one-way ANOVA. P -value <0.05 was considered as significant difference.

The relationships between the expression levels of key genes and the survival of LUAD patients

Based on TCGA database, GEPIA online tool was further used to analyze the relationships between the expression levels of key genes and the survival of LUAD patients. In survival analysis, the cut-off value of high/low expression groups was determined according to median value. P -value <0.05 was considered significant.

Development and evaluation of prognosis prediction model for LUAD patients

RNA-seq and survival data of LUAD patients were downloaded from UCSC Xena (<http://xena.ucsc.edu/>). Survival package in *R* was used to develop prognosis prediction model for LUAD patients. A nomogram predicting 3- and 5-year survival of LUAD patients was drawn by rms package. Receiver operating characteristic (ROC) curves was used to assess the discriminatory power of model and drawn by timeROC package. Area under curve (AUC) ≥ 0.7 was considered as an effective model.

Results

Identification and analysis of the dysregulated genes

In GSE118370 datasets, a total of 893 dysregulated genes including 226 up-regulated and 667 down-regulated genes in LUAD tissues were identified (Figure 1). In GSE32863 datasets, a total of 1263 dysregulated genes including 511 up-regulated and 752 down-regulated genes in LUAD tissues were identified (Figure 1). Venn diagrams showed that 275 dysregulated genes including 54 up-regulated and 221 down-regulated genes in LUAD tissues were overlapped in the two datasets (Figure 2 and Table 1).

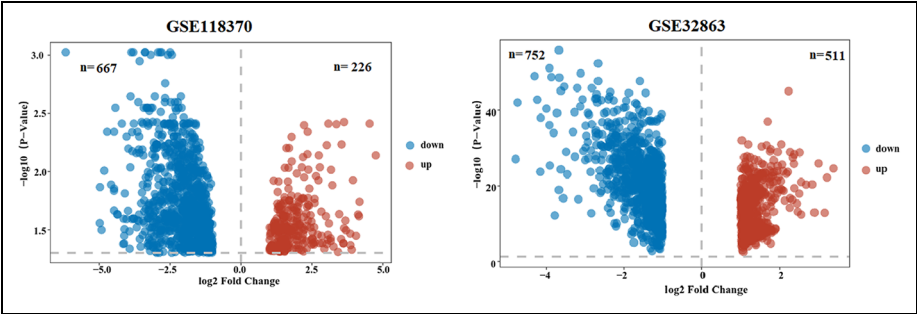


Figure 1. Volcano plot comparing all of the differentially expressed genes in two different datasets (The differentially expressed genes were selected according to adjusted p -value < 0.05 and $|\log_2FC| > 1$).

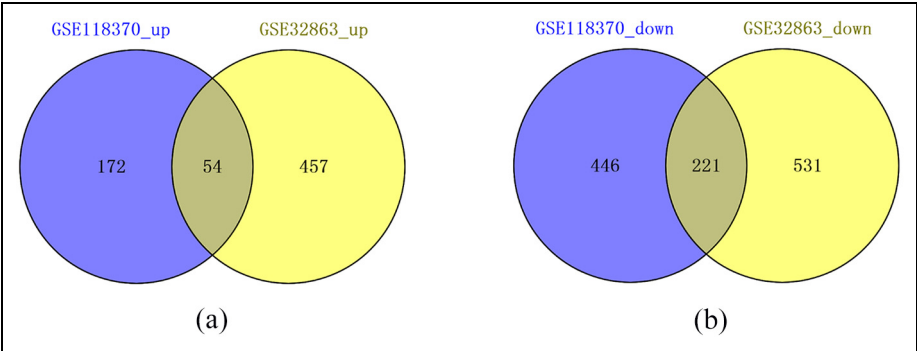


Figure 2. Venn diagrams showing the dysregulated genes in two sets of gene expression datasets: (a) the up-regulated genes in two datasets and (b) the down-regulated genes in two datasets.

The common dysregulated genes in LUAD tissues were further explored by GO and KEGG pathway enrichment analysis. The results based on GO enrichment analysis showed that these genes were markedly enriched in 27 CCs, 258 BPs, and 21 MFs. The top five enriched GO terms were as follows: CCs (cell-cell junction, extracellular matrix, collagen-containing extracellular matrix, membrane raft, membrane microdomain) (Figure 3(a)), BPs (extracellular structure organization, heart morphogenesis, morphogenesis of an epithelium, cardiac chamber morphogenesis, extracellular matrix organization) (Figure 3(b)) and MFs (transforming growth factor beta binding, glycosaminoglycan binding, extracellular matrix structural constituent, cytokine binding, growth factor binding) (Figure 3(c)). However, no pathway was significantly enriched by these common dysregulated genes.

Table 1. Common dysregulated genes in GSE32863 and GSE118370 datasets.

Gene	ID	log2FC(GSE32863)	log2FC(GSE118370)
CDH5	1003	-2.42	-2.50
RTKN2	219790	-1.11	-3.80
SLC6A4	6532	-1.92	-6.18
EDNRB	1910	-2.38	-3.38
CCM2L	140706	-1.94	-2.80
SEMA6A	57556	-2.18	-3.39
PTPRB	5787	-1.73	-2.85
CLEC1A	51267	-1.6	-2.58
GIMAP8	155038	-1.64	-3.19
RHOJ	57381	-1.45	-2.44
SDPR	8436	-2.72	-2.67
VIPRI	7433	-1.99	-3.04
ADARB1	104	-1.66	-2.68
PECAM1	5175	-2.27	-2.23
PRX	57716	-1.11	-3.62
TBX3	6926	-1.13	-3.43
CLDN18	51208	-4.79	-3.18
NOTCH4	4855	-1.15	-2.20
MME	4311	-2.27	-4.43
NDRG4	65009	-1.59	-2.83
TIMP3	7078	-2.37	-2.40
SPOCK2	9806	-2.93	-2.02
KANK3	256949	-1.51	-3.32
FHL1	2273	-2.46	-2.96
SASH1	23328	-1.68	-1.83
TCF21	6943	-2.67	-2.71
GRK5	2869	-1.99	-1.91
JAM2	58494	-1.88	-2.53
STXBP6	29091	-1.47	-3.78
GSTM5	2949	-1.06	-2.53
SH2D3C	10044	-1.93	-2.50
CD36	948	-2.9	-2.57
EMCN	51705	-1.7	-3.45
FOXF1	2294	-1.62	-2.71
TMEM74B	55321	-1.73	-2.62
LIMS2	55679	-1.54	-1.59
CAV2	858	-2.83	-1.95
SPTBN1	6711	-1.68	-2.35
TEK	7010	-2.71	-2.44
TMEM100	55273	-3.05	-4.47
FEZ1	9638	-2.17	-1.94
FAM107A	11170	-3.87	-2.91
CAV1	857	-3.55	-2.37
FGFR4	2264	-1.67	-2.37
IL7R	3575	-1.98	-2.23
EPB41L2	2037	-1.28	-1.70
NDRG2	57447	-1.58	-2.00

(continued)

Table 1. Continued

Gene	ID	log2FC(GSE32863)	log2FC(GSE118370)
ADCY4	196883	-1.32	-2.08
TBX2	6909	-1.18	-2.05
FAM65A	79567	-1.44	-1.47
ARHGEF6	9459	-1.6	-1.58
AGER	177	-4.31	-2.81
ANKRD29	147463	-1.62	-2.99
SEMA5A	9037	-1.26	-2.24
LAMA3	3909	-1.01	-2.68
CALCRL	10203	-1.96	-2.97
ACVRL1	94	-1.72	-2.22
HEG1	57493	-1.73	-2.07
SVEP1	79987	-2.44	-3.97
SIPRI	1901	-1.71	-2.07
RASIP1	54922	-1.93	-2.52
ADRB2	154	-2	-2.27
PTRF	284119	-1.66	-1.91
SLIT2	9353	-1.75	-2.66
CCDC50	152137	-1.33	-1.81
GBP4	115361	-1.43	-2.35
ABI3BP	25890	-2.21	-2.06
LAMA4	3910	-1.31	-3.03
FERMT2	10979	-1.36	-1.67
ECSCR	641700	-1.14	-1.97
C10orf67	256815	-1.21	-1.50
KLF4	9314	-2.23	-2.24
MSRB3	253827	-1.47	-1.90
GPX3	2878	-2.11	-2.23
NKG7	4818	-1.38	-1.92
TMEM47	83604	-1.64	-2.12
PKD4	5166	-2.2	-2.21
ID2	3398	-1.95	-1.65
GPC3	2719	-2.32	-2.18
MYL9	10398	-1.15	-1.76
CAMK2N1	55450	-1.15	-1.79
CA2	760	-2.24	-2.18
SLCO2A1	6578	-1.41	-2.31
LRRC32	2615	-1.75	-1.60
FGD5	152273	-1.81	-2.14
EPAS1	2034	-2.15	-2.19
ZNF331	55422	-1.03	-1.63
HHIP	64399	-1.1	-2.83
TGFBR3	7049	-2.2	-2.99
HSPC324	101928612	-1.43	-2.05
EML1	2009	-1.16	-2.22
FABP4	2167	-3.68	-4.82
CASP1	834	-1.24	-1.60
RGCC	28984	-2.74	-2.03

(continued)

Table 1. Continued

Gene	ID	log2FC(GSE32863)	log2FC(GSE118370)
PDE5A	8654	-1.8	-2.06
SMAD6	4091	-2.21	-2.68
CLEC14A	161198	-2.35	-1.86
DACH1	1602	-1.62	-2.57
UPK3B	80761	-1.54	-1.43
LTBP4	8425	-1.75	-2.08
SH3GL3	6457	-1.16	-3.36
B3GALNT1	8706	-1.21	-1.52
COL6A6	131873	-1.67	-2.32
PDLIM3	27295	-1.32	-2.41
STARD13	90627	-1.02	-2.11
STARD8	9754	-1.07	-1.80
VWF	7450	-2.23	-2.24
CDO1	1036	-1.54	-2.71
LMCD1	29995	-1.63	-1.92
MCEMP1	199675	-4.24	-2.23
RAMP3	10268	-2.27	-2.52
HSPB6	126393	-1.45	-2.38
CD93	22918	-2.3	-1.43
FMO2	2327	-3.12	-2.51
ESAM	90952	-1.83	-1.86
ARAP3	64411	-1.11	-1.31
LDB2	9079	-2.71	-2.35
TACC1	6867	-1.74	-1.39
LEPR	3953	-2.19	-1.96
SOX7	83595	-1.43	-2.58
ITM2A	9452	-2.37	-1.83
KLF13	51621	-1.2	-2.47
GHR	2690	-1.37	-3.10
DUOXA1	90527	-1.36	-1.93
LIMCH1	22998	-1.57	-1.27
AQP4	361	-2.52	-2.01
STX11	8676	-2.36	-2.44
GJA4	2701	-1.31	-1.22
PLPP3	8613	-2.02	-1.71
MYH11	4629	-1.99	-1.74
TM6SF1	53346	-1.14	-1.40
ANGPT1	284	-1.92	-2.03
CYYR1	116159	-2.01	-2.07
TCEAL2	140597	-1.61	-2.24
HOXA5	3202	-2.09	-1.83
LYVE1	10894	-2.69	-3.27
ZEB2	9839	-1.1	-1.47
ADGRL2	23266	-1.17	-2.84
SLIT3	6586	-1.62	-3.98
CPED1	79974	-1.73	-1.96
FGR	2268	-2.06	-1.66

(continued)

Table 1. Continued

Gene	ID	log2FC(GSE32863)	log2FC(GSE118370)
GNLY	10578	-1.27	-3.07
ETSI	2113	-1.27	-1.12
MYH10	4628	-1.48	-1.63
AHNAK	79026	-1.34	-1.06
GNG11	2791	-2.63	-2.62
NTNG1	22854	-1.14	-4.37
FCN3	8547	-3.91	-4.35
ILIRL1	9173	-1.13	-2.66
MFAP4	4239	-3.28	-2.36
COX4I2	84701	-1.41	-1.42
PPP1R14A	94274	-1.93	-2.29
ADH1B	125	-2.67	-3.25
PRKCH	5583	-1.24	-1.32
SPARCL1	8404	-2.66	-1.45
SCARA5	286133	-1.83	-2.86
CRYAB	1410	-2.21	-1.74
FXRD6	53826	-1.69	-1.27
KLF9	687	-1.54	-1.91
LRRN3	54674	-1.24	-1.83
COX7A1	1346	-2.21	-1.71
FZD4	8322	-1.67	-1.23
SOX17	64321	-1.14	-3.61
ENG	2022	-1.2	-1.75
APOL3	80833	-1.21	-2.58
VGLL3	389136	-1.04	-2.50
CD300LG	146894	-2.34	-4.98
TMEM204	79652	-1.54	-1.26
CELF2	10659	-1.86	-1.23
MAOB	4129	-1.25	-2.36
GFOD1	54438	-1.14	-1.56
MAL	4118	-1.95	-1.75
PODXL	5420	-1.08	-1.12
FAM189A2	9413	-1.84	-1.44
TNNC1	7134	-2.61	-3.16
SYNM	23336	-1.07	-1.32
CLIC5	53405	-2.39	-2.22
SLC19A3	80704	-1.73	-4.92
AXIN2	8313	-1.02	-2.83
GATA2	2624	-1.15	-1.40
ZNF366	167465	-1.12	-2.34
AOC3	8639	-1.1	-2.15
OLFML2A	169611	-1.21	-1.31
TSPAN18	90139	-1.11	-1.39
PIK3R1	5295	-1.09	-1.17
SBSPON	157869	-1.44	-1.48
ACADL	33	-1.18	-3.47
SLC39A8	64116	-1.96	-1.53

(continued)

Table 1. Continued

Gene	ID	log2FC(GSE32863)	log2FC(GSE118370)
JAML	120425	-1.68	-1.86
ALOX5	240	-2.3	-1.75
SFTPC	6440	-3.68	-3.57
CD52	1043	-2.71	-1.27
DKK3	27122	-1.51	-1.92
LHFP	10186	-1.67	-1.47
SOSTDC1	25928	-2.12	-1.97
ACSL4	2182	-1.18	-1.53
WFDC1	58189	-1.23	-1.65
ITPRIP	85450	-1.2	-1.23
PDZD2	23037	-1.18	-1.72
HIGD1B	51751	-2.52	-1.65
PPP1R3C	5507	-1.2	-1.78
PMP22	5376	-1.8	-1.43
FPR1	2357	-2.22	-3.29
EPB41L3	23136	-1.15	-1.30
OLFML1	283298	-1.58	-1.55
KANK2	25959	-1.53	-1.30
WISP2	8839	-1.78	-2.98
C1orf162	128346	-2.2	-1.29
GIMAP5	55340	-1.82	-1.31
HBEGF	1839	-2.1	-1.44
HYAL1	3373	-2.29	-1.83
DUOX1	53905	-2.07	-2.32
PLAC9	219348	-3.07	-1.77
HBB	3043	-4.15	-3.54
COL13A1	1305	-1.33	-2.06
DENND2A	27147	-1.22	-1.68
FBLN1	2192	-2.01	-1.72
C10orf54	64115	-1.57	-1.47
RGS5	8490	-1.09	-1.29
DPEP2	64174	-1.7	-1.74
ZAK	51776	-1.25	-1.15
BLACAT1	101669762	1.06	2.22
PROM2	150696	2.34	2.35
SPINK1	6690	3.16	4.75
COL10A1	1300	1.38	2.23
LAPTM4B	55353	1.23	2.41
TIMPI	7076	1.6	1.77
RASEF	158158	1.25	2.00
GOLM1	51280	1.82	1.53
SGPP2	130367	1.94	2.37
NQO1	1728	1.69	2.21
TMEM45B	120224	1.76	2.01
SULF1	23213	1.27	1.75
SPSB1	80176	1.11	1.44
AGR2	10551	1.26	2.40

(continued)

Table 1. Continued

Gene	ID	log2FC(GSE32863)	log2FC(GSE118370)
SOX4	6659	1.29	1.36
FAM83H	286077	1.44	2.83
ZNF750	79755	1.22	1.98
HOOK1	51361	1.79	1.51
NME1	4830	1.47	1.75
HABP2	3026	1.17	2.49
UBE2T	29089	1.3	1.96
ST14	6768	1.58	1.59
SMPDL3B	27293	1.32	2.08
NET1	10276	1.1	1.40
KCNK5	8645	1.24	1.27
CD24	100133941	1.03	2.58
KCNQ3	3786	1.23	1.62
ZDHHC9	51114	1.25	1.88
TLCD1	116238	1.3	1.91
AGRN	375790	1.64	1.07
KRT6A	3853	1.07	3.63
ALDH18A1	5832	1.01	1.48
SLC39A11	201266	1.35	1.60
SEMA4B	10509	1.85	1.75
TOP2A	7153	2.52	1.78
NEK2	4751	1.13	2.04
CEACAM6	4680	1.3	2.52
GJB2	2706	1.06	3.06
GYG2	8908	1.1	4.05
GFPT1	2673	1.26	1.62
P4HB	5034	1.09	1.55
SPDEF	25803	1.94	3.06
TXNDC17	84817	1.23	1.61
GDF15	9518	1.15	3.61
TFAP2A	7020	1.84	2.57
RAB25	57111	1.12	1.65
HMGB3	3149	2.13	1.94
FUT2	2524	1.36	1.08
GGCT	79017	1.2	1.61
FAM83A	84985	1.98	3.26
ELF3	1999	1.3	1.68
SERINC2	347735	2.27	1.95
TMPRSS4	56649	1.46	3.90
EPCAM	4072	1.27	1.37

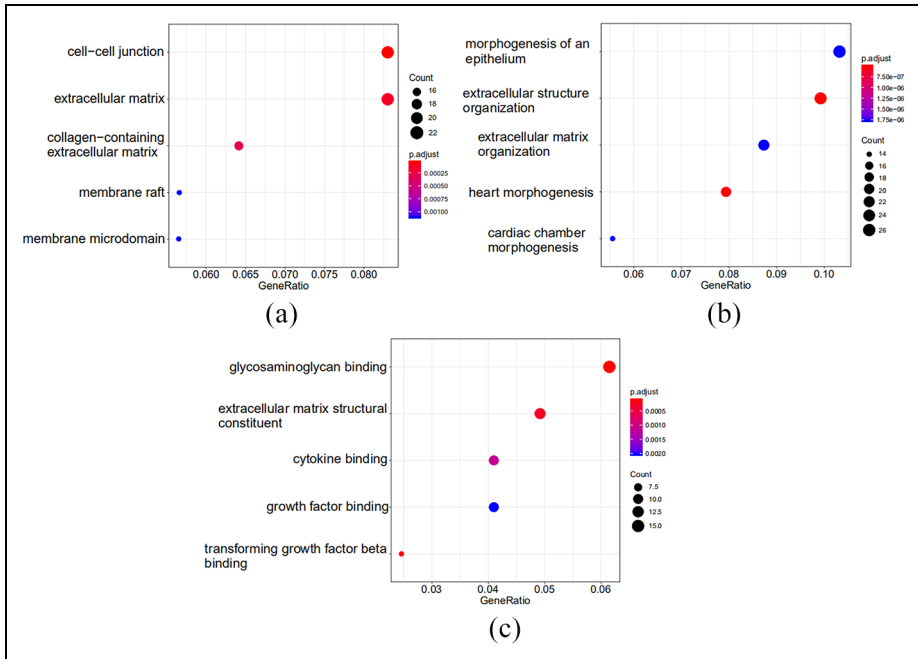


Figure 3. Bubble plots showing top five GO terms enriched by the common dysregulated genes: (a) cellular components (CCs), (b) biological processes (BPs), and (c) molecular functions (MFs).

As shown in Figure 4, there were 226 nodes and 463 edges in the PPI network. The average number of neighbors in the network was 4.097. Node degree analysis indicated that PECAM1 (platelet and endothelial cell adhesion molecule 1), CDH5 (cadherin 5), VWF (von Willebrand factor), ENG (endoglin), and KLF4 (Kruppel like factor 4) with degree >15 were key genes in the network.

Validation of the expression levels of key genes

By merging the related data in TCGA and GTEx databases, we found that the expression levels of PECAM1, CDH5, VWF, ENG, and KLF4 were significantly down-regulated in LUAD tissues compared with normal tissues, which was consistent with the current microarray results (Figure 5).

The relationship between the expression levels of key genes and the survival of LUAD patients

The relationship between the expression levels of PECAM1, CDH5, VWF, and ENG and the survival of LUAD patients was analyzed based on 478 patients. The

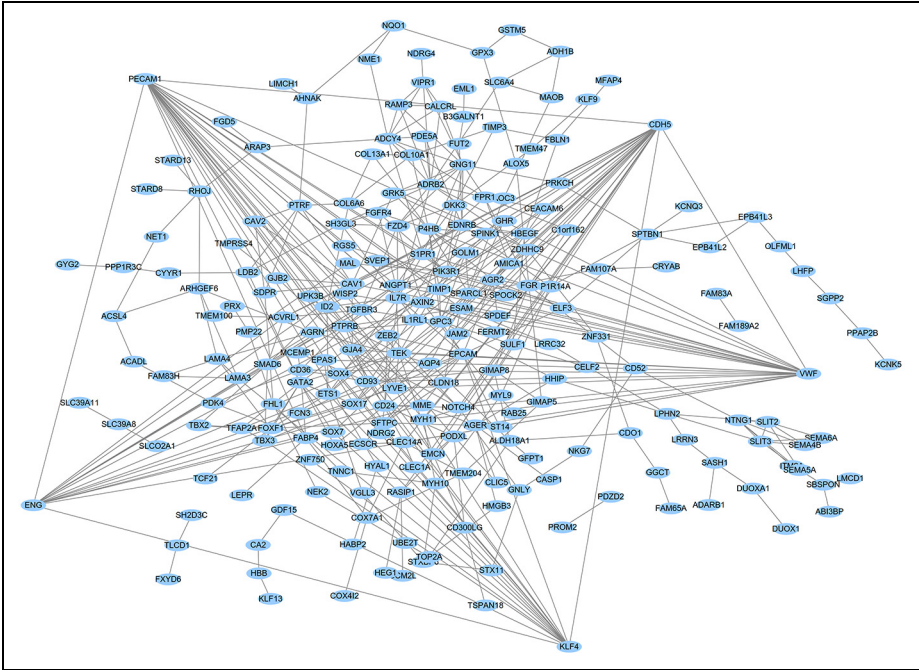


Figure 4. PPI network of the common dysregulated genes (The interaction relationship was constructed based on textmining, experiments, databases, co-expression, neighborhood, gene fusion and co-occurrence. The minimum required interaction score is 0.4).

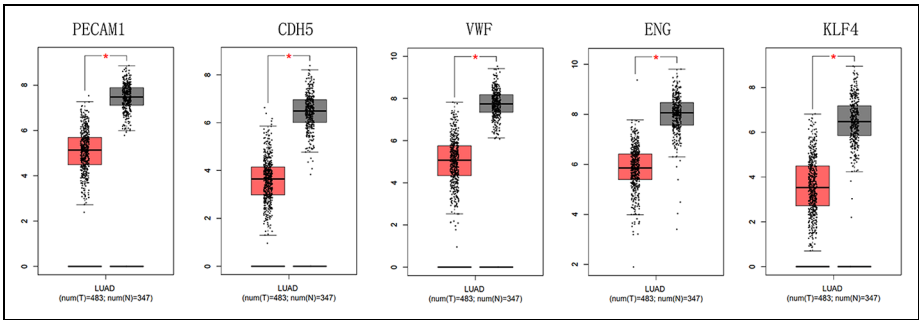


Figure 5. The expression levels of key genes in LUAD tissues (TCGA tumors vs (TCGA normal + GTEx normal); num: Number; T: tumor; N: normal; Ordinate values represent log2 (TPM + 1)).

relationship between the expression levels of KLF4 and the survival of LUAD patients was analyzed based on 477 patients. Results showed that low expression of ENG was associated with poor OS in LUAD patients. Low expression of

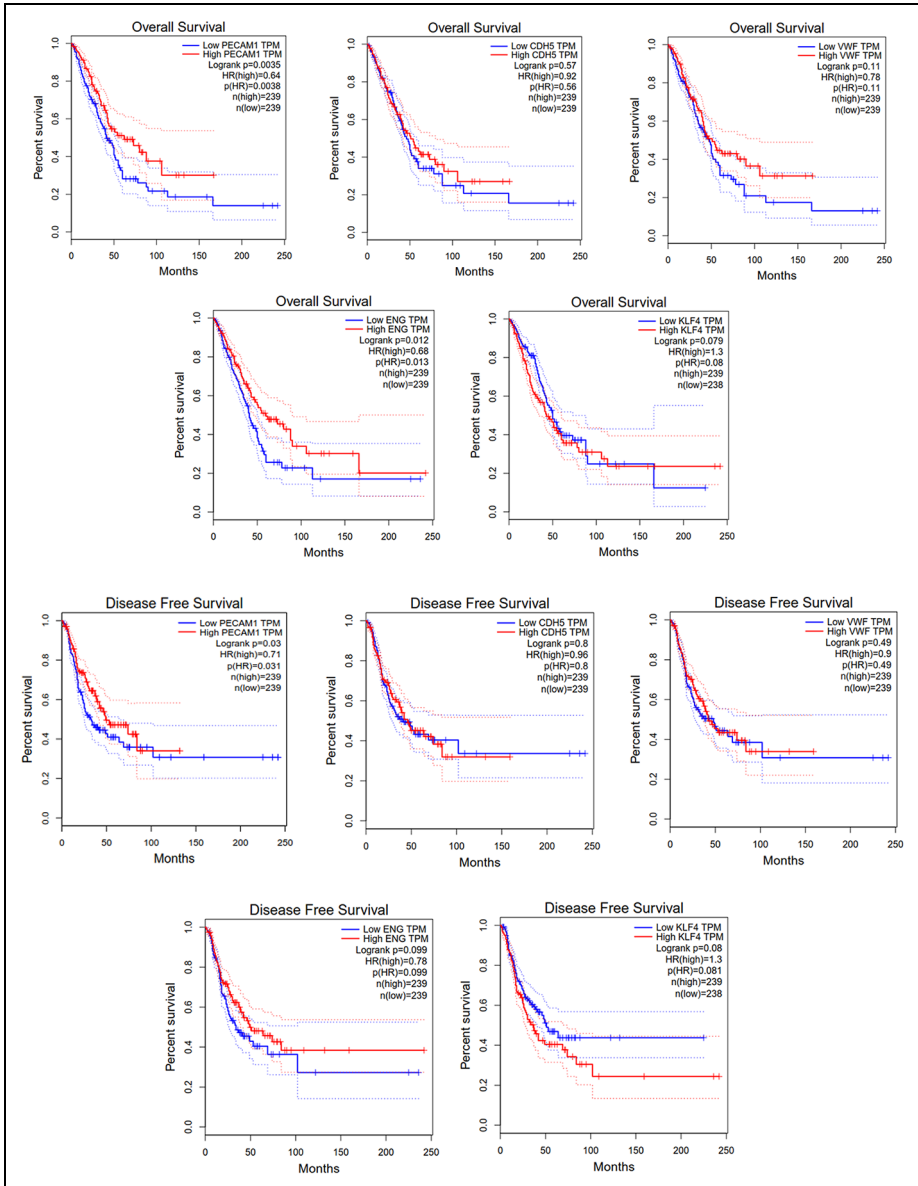


Figure 6. The relationship between the expression levels of key genes and the survival of LUAD patients (The hazards ratio was calculated based on Cox PH Model. The 95% confidence interval was added as dotted line).

PECAM1 was associated not only with poor OS but also with poor disease free survival (DFS) in LUAD patients (Figure 6).

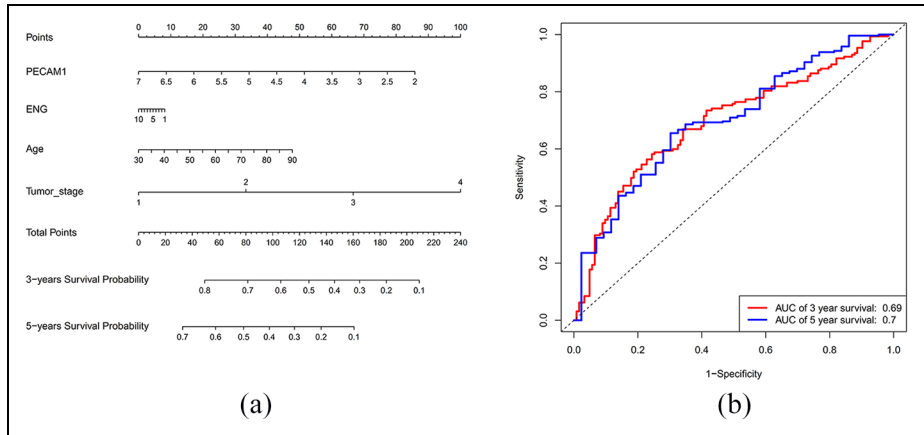


Figure 7. A nomogram predicting 3- and 5-year survival of LUAD patients and the receiver operating characteristic (ROC) curve analysis for the survival prediction model: (a) nomogram and (b) ROC curve.

Development and evaluation of prognosis prediction model for LUAD patients

Potential prognostic factors including age, tumor stage, ENG, and PECAM1 were used to develop prognosis prediction model for LUAD patients (Figure 7(a)). ROC curve showed that the model could effectively predict 5-year survival of LUAD patients (AUC = 0.7) (Figure 7(b)).

Discussion

With the development of high-throughput technologies, numerous datasets including mRNA, miRNA, and lncRNA expression profiles were generated and stored in public databases. Reanalysis of these datasets could obtain novel potential biomarkers of diseases. In the present study, we used bioinformatics methods and tools to retrieve the related datasets in GEO database and identify genes, BPs, CCs, and MFs closely related to LUAD. A total of two independent datasets (GSE32863 and GSE118370) were included in the current study. Differential gene expression analysis showed 275 common dysregulated genes (54 up-regulated and 221 down-regulated genes in LUAD tissues) in the two datasets. Using GO enrichment analysis, we found that these common dysregulated genes were significantly enriched in 27 CCs, 258 BPs, and 21 MFs, such as morphogenesis of an epithelium, transforming growth factor beta binding, cytokine binding, and glycosaminoglycan binding. By constructing the PPI network, a number of hub genes involved in LUAD were identified, including PECAM1, CDH5, VWF, ENG, and KLF4. PECAM1 gene located on chromosome 17q23.3 and encoded the protein which was found on the surface of monocytes, platelets, neutrophils, and some types of T-cells, and made up a large portion of endothelial cell intercellular junctions.^{13,14} Furthermore, the

protein was a member of the immunoglobulin superfamily and likely participated in leukocyte migration, angiogenesis, and integrin activation.^{15–17} CDH5 gene encoded a classical cadherin of the cadherin superfamily and played a role in endothelial adherens junction assembly and maintenance. The aberrant CDH5 expression had been observed in tumor cells of various malignancies, including lung cancer.^{18–21} VWF gene was located on chromosome 12p13.31 and encoded a plasma glycoprotein that played a critical role in primary hemostasis.²² In addition, several studies had also identified a series of additional biological functions for VWF. For instance, Starke et al.²³ described a novel role for VWF as a negative regulator of angiogenesis. Inhibition of VWF expression in endothelial cells increased vascular endothelial growth factor (VEGF) receptor-2 (VEGFR-2)-dependent proliferation and migration. VWF was important in modulating inflammatory responses, which was supported by data from several different animal models.^{24,25}

ENG gene encoded a homodimeric transmembrane protein which was a major glycoprotein of the vascular endothelium. O'Leary et al.²⁶ identified ENG as an epigenetically regulated tumor-suppressor gene in lung cancer. KLF4 gene encoded a protein that belonged to the Kruppel family of transcription factors. The encoded protein was thought to control the G1-to-S transition of the cell cycle following DNA damage by mediating the tumor suppressor gene p53. Li et al.²⁷ found that KLF4 was dramatically down-regulated in lung adenocarcinoma tissue and cell lines. Restoration of KLF4 inhibits migration and invasion of LUAD cells through suppressing MMP2.

In order to further confirm the expression levels of hub genes in the current PPI network, we analyzed the related data in TCGA and GTEx databases and found that these genes were also significantly down-regulated in LUAD tissues. Furthermore, survival analysis based on TCGA data suggested that the expression levels of PECAM1 and ENG were associated with survival of LUAD patients. Specifically, low expression of ENG was associated with poor OS of LUAD patients. Low expression of PECAM1 was associated not only with poor OS but also with poor DFS of LUAD patients. The prognosis prediction model based on age, tumor stage, ENG and PECAM1 could effectively predict 5-year survival of LUAD patients.

Although some key genes associated with lung adenocarcinoma had been identified and the prognosis prediction model had been constructed in the current study, these findings based on bioinformatics analysis should be further investigated experimentally.

Conclusions

The present study demonstrated that PECAM1, CDH5, VWF, ENG, and KLF4 were key genes in LUAD, and the expression levels of PECAM1 and ENG were associated with survival of LUAD patients, which would provide the foundation for the development of novel methods for the diagnosis and therapy of LUAD.


Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Xueren Gao  <https://orcid.org/0000-0003-2452-2340>

References

1. Zappa C and Mousa SA. Non-small cell lung cancer: current treatment and future advances. *Transl Lung Cancer Res* 2016; 5(3): 288–300.
2. Gao L, Qiu H, Liu J, et al. KLF15 promotes the proliferation and metastasis of lung adenocarcinoma cells and has potential as a cancer prognostic marker. *Oncotarget* 2017; 8(66): 109952–109961.
3. Zheng Y, Li B, Wang J, et al. Identification of SUV39H2 as a potential oncogene in lung adenocarcinoma. *Clin Epigenetics* 2018; 10(1): 129.
4. Shi X, Tan H, Le X, et al. An expression signature model to predict lung adenocarcinoma-specific survival. *Cancer Manag Res* 2018; 10: 3717–3732.
5. Su H, Lin Z, Peng W, et al. Identification of potential biomarkers of lung adenocarcinoma brain metastases via microarray analysis of cDNA expression profiles. *Oncol Lett* 2019; 17(2): 2228–2236.
6. Chen Y, Chen H, Mao B, et al. Transcriptional characterization of the tumor Immune microenvironment and its prognostic value for locally advanced lung adenocarcinoma in a Chinese population. *Cancer Manag Res* 2019; 11: 9165–9173.
7. Zhu X, Luo H and Xu Y. Transcriptome analysis reveals an important candidate gene involved in both nodal metastasis and prognosis in lung adenocarcinoma. *Cell Biosci* 2019; 9: 92.
8. Selamat SA, Chung BS, Girard L, et al. Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. *Genome Res* 2012; 22(7): 1197–1211.
9. Xu L, Lu C, Huang Y, et al. SPINK1 promotes cell growth and metastasis of lung adenocarcinoma and acts as a novel prognostic biomarker. *BMB Rep* 2018; 51(12): 648–653.
10. Yu G, Wang LG, Han Y, et al. ClusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 2012; 16(5): 284–287.
11. Szklarczyk D, Morris JH, Cook H, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* 2017; 45(D1): D362–D368.
12. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003; 13: 2498–2504.

13. Woodfin A, Voisin MB and Nourshargh S. PECAM-1: a multi-functional molecule in inflammation and vascular biology. *Arterioscler Thromb Vasc Biol* 2007; 27(12): 2514–2523.
14. Lertkietmongkol P, Liao D, Mei H, et al. Endothelial functions of platelet/endothelial cell adhesion molecule-1 (CD31). *Curr Opin Hematol* 2016; 23(3): 253–259.
15. Nakada MT, Amin K, Christofidou-Solomidou M, et al. Antibodies against the first Ig-like domain of human platelet endothelial cell adhesion molecule-1 (PECAM-1) that inhibit PECAM-1-dependent homophilic adhesion block in vivo neutrophil recruitment. *J Immunol* 2000; 164(1): 452–462.
16. DeLisser HM, Christofidou-Solomidou M, Strieter RM, et al. Involvement of endothelial PECAM-1/CD31 in angiogenesis. *Am J Pathol* 1997; 151(3): 671–677.
17. Cao G, O'Brien CD, Zhou Z, et al. Involvement of human PECAM-1 in angiogenesis and in vitro endothelial cell migration. *Am J Physiol Cell Physiol* 2002; 282(5): C1181–C1190.
18. Hung MS, Chen IC, Lung JH, et al. Epidermal growth factor receptor mutation enhances expression of cadherin-5 in lung cancer cells. *PLoS One* 2016; 11(6): e0158395.
19. Mao XG, Xue XY, Wang L, et al. CDH5 is specifically activated in glioblastoma stemlike cells and contributes to vasculogenic mimicry induced by hypoxia. *Neuro Oncol* 2013; 15(7): 865–879.
20. Fujita T, Igarashi J, Okawa ER, et al. CHD5, a tumor suppressor gene deleted from 1p36.31 in neuroblastomas. *J Natl Cancer Inst* 2008; 100(13): 940–949.
21. Metodieva SN, Nikolova DN, Cherneva RV, et al. Expression analysis of angiogenesis-related genes in Bulgarian patients with early-stage non-small cell lung cancer. *Tumori* 2011; 97(1): 86–94.
22. Sadler JE. Biochemistry and genetics of von Willebrand factor. *Annu Rev Biochem* 1998; 67: 395–424.
23. Starke RD, Ferraro F, Paschalaki KE, et al. Endothelial von Willebrand factor regulates angiogenesis. *Blood* 2011; 117(3): 1071–1080.
24. Hillgruber C, Steingraber AK, Pöppelmann B, et al. Blocking von Willebrand factor for treatment of cutaneous inflammation. *J Invest Dermatol* 2014; 134(1): 77–86.
25. Methia N, André P, Denis CV, et al. Localized reduction of atherosclerosis in von Willebrand factor-deficient mice. *Blood* 2001; 98(5): 1424–1428.
26. O'Leary K, Shia A, Cavicchioli F, et al. Identification of Endoglin as an epigenetically regulated tumour-suppressor gene in lung cancer. *Br J Cancer* 2015; 113(6): 970–978.
27. Li S, Huang L, Gu J, et al. Restoration of KLF4 inhibits invasion and metastases of lung adenocarcinoma through suppressing MMP2. *J Cancer* 2017; 8(17): 3480–3489.

Author biographies

Xinyu Wang is a College Student. Her research interest focuses on bioinformatics analysis and the pathogenesis of lung adenocarcinoma.

Jiaojiao Yang, Master of Medicine, is mainly engaged in research into the pathogenesis of cancer.

Xueren Gao, Doctor of Medicine, is mainly engaged in research into the pathogenesis of cancer.