Identification of molecular markers associated with diagnosis and prognosis in lung adenocarcinoma by bioinformatics analysis

Zimeng Wei  
Wuhan University

Linnan Zang  
Wuhan University

Min Zhao (minzhao@whu.edu.cn)  
Wuhan University

Research Article

Keywords: Lung Adenocarcinoma, bioinformatics, differentially expressed genes, prognosis

DOI: https://doi.org/10.21203/rs.3.rs-79373/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Background: Lung adenocarcinoma (LUAD) is the main histological subtype of lung cancer. However, the molecular mechanism underlying LUAD is not yet clearly defined, but elucidating this process in detail would be of great significance for clinical diagnosis and treatment.

Methods: Gene expression profiles were retrieved from Gene Expression Omnibus database (GEO), and the common differentially expressed genes (DEGs) were identified by online GEO2R analysis tool. Subsequently, the enrichment analysis of function and signaling pathways of DEGs in LUAD were performed by gene ontology (GO) and The Kyoto Encyclopedia of Genes and Genomics (KEGG) analysis. The protein-protein interaction (PPI) networks of the DEGs were established through the Search Tool for the Retrieval of Interacting Genes (STRING) database and hub genes were screened by plug-in CytoHubba in Cytoscape. Afterwards, the miRNAs and the hub genes network was constructed via miRWalk. Finally, receiver operating characteristic (ROC) curve and Kaplan-Meier plotter were performed to analyze the diagnosis and prognosis efficacy of hub genes.

Results: A total of 312 DEGs were identified, including 74 up-regulated and 238 down-regulated genes. GO analysis results showed that DEGs were mainly enriched in biological processes including composition of extracellular matrix, regulation of angiogenesis and so on. KEGG analysis results revealed DEGs were mainly enrolled in cell adhesion signaling pathway. Subsequently, 10 hub genes, CDC20, CENPF, TPX2, TOP2A, KIAA0101, CDCA7, ASPM, ECT2, UBE2T and COL1A1, were identified. And TOP2A, CDCA7, TPX2 and COL1A1 showed strong relationships with each other and the miRNAs nearby in miRNAs-mRNA network obtained by miRWalk website. Finally, all these 10 hub genes were found significantly related to the diagnosis and prognosis of LUAD (p<0.05).

Conclusions: Our results suggested that TOP2A, CDCA7, TPX2 and COL1A1 might present predictive value for the development and prognosis in LUAD, and might be used as potential molecular markers for the diagnosis and treatment of LUAD.

Introduction

Lung cancer is the leading cause of cancer morbidity and mortality in the world, with 2.1 million new cases and 1.8 million deaths predicted in 2018 [1]. Although the great advance in surgical and molecular target therapies, the prognosis of lung cancer is still dismal [2]. According to histopathology characteristics, lung cancer can be divided into two main types: non-small cell lung cancer (NSLC) and small cell lung cancer (SCLC), with NSLC being accounted for 85% of lung cancer with lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC) and large cell carcinoma (LCC) three main subtypes [3]. LUAD has been ranked the first cancer mortality in china [4],and receiving an operation remains a primary method, unfortunately the 5-year survival rate is around 10–15% [5]. Lacking sensitive and effective early diagnosis biomarkers and tending to metastasis and rapid dissemination are deemed
to contribute to the poor prognosis. Therefore, looking for reliable biomarkers or drug targets is of great importance for improving the diagnosis and prognosis of LUAD.

High-throughput screening technique which is based on RNA sequencing or microarray gene expression profiles has been widely used for gene detection and diseases analysis. Academic can rapidly and precisely get the knowledge of key genes associated with serious diseases like cancer via a large number of biological information data mining [6]. For instance, Huang et al. used integrated bioinformatics analysis of RNA sequencing data and superenhancer catalogs to identify superenhancer-associated circRNAs and further found that loss of superenhancer-regulated circNfix promoted cardiac regenerative repair and functional recovery [7]. As for lung cancer, compound CA-5f was considered as a novel late-stage autophagy inhibitor with potent anti-tumor effect against NSLC via bioinformatics technique [8]. The discovery and the lucubration of these key genes obtained through bioinformatics analysis contribute to the discovery of effective biomarkers, which can be applied to the diagnosis, treatment and prognosis of tumors.

The aim of our study was to perform multiple analyses towards key genes of LUAD through integrated bioinformatics methods. First of all, we downloaded three microarray datasets containing mRNA expression data from Gene Expression Omnibus (GEO) and the differentially expressed genes (DEGs) were screened out via GEO2R online tool. Next, by virtue of Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), functional annotation and enrichment analysis about these DEGs were performed, and then we established protein-protein interaction (PPI) network and obtained the hub genes in LUAD. Then, the miRNAs and the hub genes network was constructed to select and identify the candidate biomarkers. Receiver operating characteristic (ROC) diagnosis and Kaplan-Meier survival analysis were carried out to provide support for the diagnosis, targeted therapy and prognosis of LUAD. Last but not least, we generated candidate biomarkers’ expression violin plots based on patient pathological stage.

## Materials And Methods

*Acquisition and processing of GEO data.* 3 LUAD datasets were obtained from GEO database (available online: [http://www.ncbi.nlm.nih.gov/geo](http://www.ncbi.nlm.nih.gov/geo)) [9], including the gene expression profiles GSE118370 (expression profiling by array, 6 LUAD samples, 6 normal samples), GSE32863 (expression profiling by array, 58 LUAD samples, 58 normal samples) and GSE43458 (expression profiling by array, 80 LUAD samples, 30 normal samples). The GSE118370 dataset (Submission date: Jan 23, 2019) was based on the GPL570 platform (Affymetrix Human Genome U133 Plus 2.0 Array, Affymetrix, Inc). The GSE32863 dataset (Submission date: Mar 21, 2012) was based on the GPL6884 platform (Illumina HumanWG-6 v3.0 expression beadchip). The dataset GSE43458 (Submission date: Aug 06, 2013) was based on the GPL6244 platform ([HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array [transcript (gene) version]). DEGs were screened and identified via GEO2R online tools (available online: [https://www.ncbi.nlm.nih.gov/geo/geo2r/](https://www.ncbi.nlm.nih.gov/geo/geo2r/)) with the cut-off |log2FC|≥1 and adj. p < 0.05. The numbers of DEGs we got from GSE118370, GSE32863 and GSE43458 were 3302, 1584 and 1025, respectively. DEGs
co-occurred in all 3 datasets were integrated together and showed in a Venn diagram (available online: http://bioinformatics.psb.ugent.be/webtools/Venn/).

**GO and KEGG Pathway enrichment analyses and visualization.** Functional annotations and pathway enrichment analyses on DEGs were carried out through Enrichr website (available online: https://amp.pharm.mssm.edu/Enrichr) [10], which included the visual results of GO and KEGG enrichment analyses. GO can be used to study gene-related biological process (BP), molecular function (MF) and cellular component (CC), while KEGG can simulate and analyze the pathways that DEGs were enriched in.

**PPI network and acquisition of hub genes.** PPI network can explain the possible mechanisms of the genes involved in intracellular processing. So the STRING (available online: https://stringdb.org/, Version11.0) [11], an online tool was first applied to establish a PPI network of DEGs, which was then imported into Cytoscape software (version 3.6.1) [12]. In addition, the plug-in CytoHubba in Cytoscape was used to search and analyze the hub genes that were generated through Maximal Clique Centrality (MCC) score.

**The construction of miRNA interaction network.** MiRWalk website (available online: http://mirwalk.umm.uni-heidelberg.de/) [13] was used to establish a miRNA-mRNA regulatory network of the hub genes we obtained from CytoHubba plug-in.

**Diagnosis and prognosis analyses of hub genes.** Receiver operating characteristic (ROC) can evaluate the diagnostic efficacy of each index and the larger the area under curve (AUC), the higher the diagnostic efficacy [14]. We used SPSS (version 24) to gather statistics and analyze the case samples in GSE43458 dataset. In addition, the Kaplan-Meier plotter (available online: http://kmplot.com/) [15], a online database for survival analysis, was used to analyze the prognosis value of the hub genes. The hazard ratio (HR) with 95% confidence intervals and logrank P value were calculated and displayed on the diagram.

**Expression violin plots about candidate biomarkers based on patient pathological stage.** In order to detect the relationship between patient pathological stage and candidate biomarkers’ expression, Gene Expression Profiling Interactive Analysis (GEPIA, available online: http://gepia.cancer-pku.cn/) [16] was utilized.

**Results**

**Identification and screening of DEGs.** Three microarray datasets: GSE118370, GSE32863 and GSE43458 were obtained from GEO database, including 144 samples from LUAD patients and 94 samples of normal lung tissues. The differentially expressed genes (DEGs) were identified by online GEO2R analysis tool with the cut-off criterion of |log2FC|≥1 and adj. p < 0.05, and got 3302, 1584 and 1025 DEGs from each dataset, respectively. Afterwards, a Venn diagram (Fig. 1) was used to get an intersection of DEGs from the 3 datasets above, acquiring a total of 312 DEGs with a consistent expression, in which there were 74
up-regulated genes and 238 down-regulated genes (Table 1) in the LUAD samples compared to the normal samples.

<table>
<thead>
<tr>
<th>Up- or Down-regulated</th>
<th>Number</th>
<th>Differentially Expressed Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up-regulated</td>
<td>74</td>
<td>GDF15, TPX2, C15orf48, IGFBP3, COL1A1, ADGRF1, METTL7B, ANKR22, SLC50A1, TMPRSS4, SPP1, CLDN10, PLPP2, TIMP1, MMP12, HABP2, UBE2T, SLC2A1, EPCAM, ECT2, MMP11, FAM83A, CDH3, CRABP2, CEACAM5, FUT3, LDL1, TNFRSF21, TFAP2A, SULT1, GCNT3, COL1A1, LAD1, FUT2, KIAA0101, AGR2, SPINK1, S100P, HMGB3, DSP, GOLM1, PROM2, MUC16, CST1, SERINC2, ETVO, LGSN, PYCR1, CDCA7, SFN, CDC20, COL3A1, ASPM, OCIAD2, CD24, TCN1, SLC7A5, IL37, NQO1, PSAT1, SPDEF, TOP2A, GJB2, THBS2, SMPDL3B, PCP4, AIM2, COL1A1, LCN2, CTHRC1, F8X032, ATP10B, CENPF, MMP9</td>
</tr>
<tr>
<td>Down-regulated</td>
<td>238</td>
<td>RTKN2, FRAS1, MSRB3, AQPO, NOSTRIN, PAPSS2, A2M, HIGD1B, ANX3, TEK, GRK5, VGLL3, C10orf67, GHR, MS4A2, ZNF385B, SLC02A1, LAMP3, BCHE, STARD13, ID1, PTRF, SPTBN1, TSPAN7, LRRK2, AFAP1L1, ACADL, AHNAK, FHL1, KLF4, PLSCR4, NDRG2, TCF21, GLDN, NEDD4L, LDLR, SFTPPL, OLR1, GLPR2, AGER, HOXAX5, CD93, MSRI, WIF1, RAMBP3, CLC5, NPNT, LLAMA, ICAM2, PDK4, GBP4, EFEMP1, EMCN, ADARB1, LMO7, STARD8, LRRC32, SASH1, ESAM, SCGB1A1, CA4, TACC1, FM03, LYVE1, GIMAP4, P3H2, PHACTR2, FEZ1, PDE5A, CPA3, TBX3, PDZD2, LHFP, ADIRF, HBB, HEY1, NTNG1, GNG11, RHOJ, SRPX, TSPAN18, ZNF106, SLC39A8, SFTPC, MYL9, TGFB3, LPL, CAT, PTPRM, NDRG4, SOSTDC1, SBSPON, GPC3, ADCY4, FPR2, PGM5, LMCD1, FOSSB, MARCO, GSTM5, DCN, EPAS1, SLIT2, CX3CR1, FZD4, PCOLCE2, CYBRD1, C2orf40, AOX1, HSD17B6, GKN2, CYP4B1, SMAD6, AFF3, HSPB8, SLC6A4, TMEM204, GIMAP8, CDH13, ACVRL1, TIMP3, DACH1, EML1, REL, HBEGF, CRYAB, CD36, ETS2, PTPRB, ADGRL2, PPARG, AQPO, MT1M, TNNC1, IL1RL1, MCEMP1, CB2, KLF6, EMP1, ABCA8, SLC19A3, AOC3, TMEM47, ZBTB16, EDNRB, SLIT3, FAM167A, MYH10, STX11, MYH11, DUSP1, RGCC, VWF, ABI3BP, ARR81, VSIG4, CYYR1, MS4A15, HPGD, C7, S100A8, HEG1, IL6, CPED1, PRICKLE2, ANGPT1, SEMA6A, TMEM100, DUOX1, FERMT2, CD34, KIAA1324L, DPT, MFAP4, PECAM1, PRG4, MAOB, B3GALNT1, CDO1, SEPP1, SDPR, EPB41L2, ADAMTS1, CLEC14A, ANOS1, C8B, STXB6, S1PR1, GIMAP7, LIMCH1, SPOCK2, RNF144B, LDB2, CALCRL, RECK, CAV1, CLEC1A, JAM2, SPARCL1, CA2, ANKR229, ADH1A, METTL7A, ITM2A, FGDC5, FAM189A2, LRRN3, LGI3, IL7R, FMO2, SOCS2, ACTG2, TMOD1, FGF8, P2, ARHGEF6, IL33, PLLP, HHIP, PGC, ADH1B, NEDD9, FABP4, FCN3, PDLIM3, FAM107A, GPM6B, SEMA5A, RAMP2, VIPR1, PLAC8, ANKR21, MAMDC2, CLDN18, CDH5, OLFML1, COL6A6, FBLN5</td>
</tr>
</tbody>
</table>
**GO analysis and KEGG pathway analysis of DEGs.** To evaluate the function of DEGs, we performed functional annotations and pathway enrichment analyses via GO and KEGG. According to the result of KEGG, DEGs emerged a high enrichment in the pathways of cell adhesion, protein digestion and absorption, tyrosine metabolism, ECM receptor interaction and so on (Fig. 2A). GO analysis showed that DEGs mainly participated in the biological processes (BP) of the extracellular matrix organization, regulation of angiogenesis, negative regulation of blood vessel morphogenesis and so on (Fig. 2B). The molecular function (MF) included amyloid-beta binding, the activation of protein homodimerization, the activation of metalloendopeptidase inhibitors, the binding of low-density lipoprotein and so on (Fig. 2C). The analysis of cellular component (CC) emerged that these genes mainly encoded the components on the plasma membrane (Fig. 2D).

**Analysis of PPI network and identification of hub genes.** Exploring the protein-protein interaction in organisms can help to study molecular mechanisms of some diseases from a perspective of system. We built a PPI network via STRING website (Fig. 3A), then the interaction data were imported into Cytoscape, and then CytoHubba plug-in was applied to confirm 10 hub genes according to the scores of MCC. The results were shown in Fig. 3B. They were CDC20, CENPF, TPX2, TOP2A, KIAA0101, CDCA7, ASPM, ECT2, UBE2T and COL1A1, and their full names and functions were shown in Table 2.
<table>
<thead>
<tr>
<th>No.</th>
<th>Up or down regulation</th>
<th>Name</th>
<th>Full name</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>↑</td>
<td>CDC20</td>
<td>Cell division cycle 20</td>
<td>CDC20 is a regulatory protein interacting with several other proteins at multiple points in the cell cycle.</td>
</tr>
<tr>
<td>2</td>
<td>↑</td>
<td>CENPF</td>
<td>Centromere protein F</td>
<td>CENPF is a component of the nuclear matrix during the G2 phase of interphase</td>
</tr>
<tr>
<td>3</td>
<td>↑</td>
<td>TPX2</td>
<td>TPX2 microtubule nucleation factor</td>
<td>TPX2 can bind to and increases the activity of a protein kinase, an enzyme which phosphorylates a protein.</td>
</tr>
<tr>
<td>4</td>
<td>↑</td>
<td>TOP2A</td>
<td>DNA topoisomerase II alpha</td>
<td>TOP2A is involved in processes such as chromosome condensation, chromatid separation, and the relief of torsional stress.</td>
</tr>
<tr>
<td>5</td>
<td>↑</td>
<td>KIAA0101</td>
<td>PCLAF(PCNA clamp associated factor)</td>
<td>KIAA0101 can prevent doxorubicin-induced apoptosis by inhibiting p53 activation</td>
</tr>
<tr>
<td>6</td>
<td>↑</td>
<td>CDCA7</td>
<td>Cell division cycle associated 7</td>
<td>Overexpression of this gene is found to enhance the transformation of lymphoblastoid cells</td>
</tr>
<tr>
<td>7</td>
<td>↑</td>
<td>ASPM</td>
<td>Assembly factor for spindle microtubules</td>
<td>ASPM is essential for normal mitotic spindle function in embryonic neuroblasts.</td>
</tr>
<tr>
<td>8</td>
<td>↑</td>
<td>ECT2</td>
<td>Epithelial cell transforming 2</td>
<td>The expression of this gene is elevated with the onset of DNA synthesis and remains elevated during G2 and M phases.</td>
</tr>
</tbody>
</table>
9  ↑  UBE2T  Ubiquitin-conjugating enzyme E2T  The protein encoded by UBE2T catalyzes the covalent attachment of ubiquitin to protein substrates.
10  ↑  COL1A1  Collagen type I alpha 1 chain  COL1A1 encodes the pro-alpha1 chains of type I collagen whose triple helix comprises two alpha1 chains and one alpha2 chain.

miRNA regulatory network of hub genes. Through inputting 10 hub genes into miRWalk website, we obtained the miRNAs regulatory network related with the hub genes (Fig. 4). The blue spots in the center were the hub genes, and what surrounded them were the miRNA which might regulate the hub genes. As were shown in Fig. 4, these four genes: TPX2, TOP2A, CDCA7 and COL1A1 had stronger interactions and formed regulation networks with the miRNAs nearby, in which the central miRNAs has-miR-6129, has-miR-6759-5p, has-miR-6510-5p and so on had two more connections with the four genes above. miR-145 [17], hsa-let-7d-5p [18] and so on also proved to have associations with the occurrence and the development of tumors.

Diagnosis and prognosis analyses of hub genes. Based on the GSE43458 dataset, SPSS software was used to analyze the diagnostic efficacy of these 10 hub genes. The results showed that the AUC of all the hub genes were above 0.800 (Fig. 5), CDC20 (AUC=0.940), CENPF (AUC=0.935), TPX2 (AUC=0.929), TOP2A (AUC=0.954), KIAA0101 (AUC=0.914), CDCA7 (AUC=0.927), ASPM (AUC=0.902), ECT2 (AUC=0.896), UBE2T (AUC=0.978) and COL1A1 (AUC=0.948). Among them, the AUC of TOP2A and UBE2T were above 0.950, indicating that as diagnosis indexes, TOP2A and UBE2T would have better efficacy. Kaplan-Meier prognosis analysis has been performed, finding that all the hub genes especially UBE2T (HR=2.23 [1.87-2.66], logrank p < 1E-16) and TPX2 (HR=1.86 [1.64-2.12], logrank p < 1E-16), had significant relations with the prognosis of LUAD patients, as what were shown in Fig. 6.

Expression profiles analyses about four candidate biomarkers. As for the four candidate biomarkers, TPX2, TOP2A, CDCA7 and COL1A1, we further explored their relationship between expression profiles and patient’s pathological stage. In Fig.7, we found that all these four genes, especially COL1A1, had an increase trend in multiple pathological stages of LUAD, and with the progression of tumors (from stage 1 to stage 4), the expression of these candidate biomarkers increased significantly.

Discussion

In this study, 312 DEGs were identified including 74 up-regulated and 238 down-regulated genes in three datasets from the GEO database. Furthermore, through Enricher online web tool, we visualized the outcomes derived from Gene Ontology and KEGG pathway enrichment analysis. As for the biological processes, these DEGs were enriched in extracellular matrix organization, regulation of angiogenesis, negative regulation of blood vessel morphogenesis, and negative regulation of angiogenesis and so on.
As for the molecular functions, the DEGs were mainly enriched in amyloid-beta binding, protein homodimerization activity, metalloendopeptidase inhibitor activity, low-density lipoprotein particle binding and calcium ion binding. For the cellular components, the DEGs showed enrichment in the integral component of plasma membrane. Angiogenesis is closely related to the occurrence and progression of cancers [19] and the formation of extracellular matrix is also associated with tumor metastasis and invasion [20, 21]. According to KEGG pathway enrichment analysis, tyrosine metabolism was found to be significant in LUAD. Li et al. revealed that activation of tyrosine metabolism in CD13+ cancer stem cells may drive relapse in hepatocellular carcinoma by means of generating nuclear acetyl-CoA to acetylate and stabilize Foxd3, and allowing CD13+ cancer stem cells to sustain quiescence and resistance to chemotherapeutic agents [22]. Nitration of protein tyrosine has proved to be involved in a variety of biological processes, including signal transduction, protein degradation, energy metabolism, mitochondrial dysfunction, enzyme inactivation, immunogenic response, cell apoptosis and cell death, and plays an important role in the occurrence and metastasis of lung cancer [23]. Therefore, the signaling pathway of tyrosine metabolism was expected to be a potential drug therapy target for LUAD.

Next, DEGs PPI network was constructed via the STRING online database and Cytoscape software. By virtue of “CytoHubba” plug-in, the top ten hub genes, CDC20, CENPF, TPX2, TOP2A, KIAA0101, CDCA7, ASPM, ECT2, UBE2T, COL1A1 were identified and they were all found up-regulated in LUAD. In addition, miRWalk, an online analysis tool, was used to construct a network of miRNAs associated with the regulation of these genes, and finally four genes (TOP2A, CDCA7, TPX2 and COL1A1) were found to show strong associations with each other, which could be considered as new effective targets to improve the prognosis of LUAD patients.

Topoisomerase II (TOP2) has been clarified to have crucial functions, including DNA replication, transcription and chromosome segregation, and more and more active anticancer drugs targeted it [24]. TOP2 contains two types of isozymes: TOP2A and topoisomerase II beta (TOP2B) [25], TOP2A is the only enzyme able to cleave and re-ligate the double-strand backbone of DNA, which is indispensable for DNA replication, transcription, and repair [26, 27]. Ejlertsen et al. [26] showed that TOP2A was a direct molecular target of anthracyclines that can improve the sensitivity of anthracycline-containing chemotherapy in high-risk breast cancer patients. In malignant peripheral-nerve sheath tumor, TOP2A was the most overexpressed gene compared with benign neurofibromas [28]. High expression of TOP2A was found to be correlated to worse overall survival (OS) in all non-small-cell lung cancer and lung adenocarcinoma patients, but not in lung squamous cell carcinoma patients [29, 30]. It has also been reported that miRNA's being associated with TOP2A plays an important role in lung cancer, for example, down-regulation of miRNA-144-3p whose potential target was TOP2A, was highly enriched in various key pathways like the protein digestion and absorption and the thyroid hormone signaling pathways in non-small cell lung cancer from the comprehensive meta-analysis [31].

Cell division cycle associated 7, CDCA7, was identified as a c-Myc responsive gene, and behaved as a direct c-Myc target gene. Overexpression of CDCA7 was found to enhance the transformation of lymphoblastoid cells, and it complements a transformation-defective Myc Box II mutant, suggesting its
involvement in c-Myc-mediated cell transformation [32]. In quite a number of tumors, such as hepatocellular carcinoma [33], colorectal cancer [34], lymphoma [35], breast cancer [36], CDCA7 was all reported up-regulated and might be a potential prognostic factor and therapeutic target. Wang et al. have found that CDCA7 could promote lung adenocarcinoma proliferation via regulating the cell cycle and silencing CDCA7 inhibited cell proliferation in LUAD through G1 phase arrest and induction of apoptosis, which implied that CDCA7 might be identified as a potential therapeutic target for new biomarkers and LUAD [37].

TPX2, which is also known as DIL2 or p100, uses two flexibly linked elements (‘ridge’ and ‘wedge’) in a novel interaction mode to simultaneously bind across longitudinal and lateral tubulin interfaces [38, 39]. In ovarian cancer, it can promote the proliferation and migration of human ovarian cancer cells by regulating PLK1 expression [40]. Except for these function, in various cancers can TPX2 also control bladder cancer cell’s proliferation and invasion via TPX2-p53-GLIPR1 regulatory circuitry [41], regulate the PI3K/AKT signaling pathway to facilitate hepatocellular carcinoma [42], interactive with miRNA such as miR-485-3p [43], miR-361-5p [44], miR-335-5p [45], miR-216b [46] and so on. Zhou et al. have verified that TPX2 can activate the epithelial-mesenchymal transition process and promote both the expression and activities of matrix metalloproteinase (MMP)2 and MMP9 in non-small cell lung cancer (NSCLC), which means TPX2 promotes the metastasis and malignant progression of NSCLC and could thus serve as a marker of poor prognosis in NSCLC [47].

Some studies have confirmed that COL1A1-related miRNAs were involved in the regulation of different kinds of tumors, including LUAD [18, 48]. Zhang et al. found that COL1A1 was positively correlated with NOTCH3 expression, and the miR-150 /NOTCH3/COL1A1 axis might be involved in EGFR-TKI resistance in LUAD, which provided a potential target for LUAD therapy and could be used as a prognostic target [49]. Hsa-let-7d-5p has been proved that it can regulate cell cycle including both G1/S and G2/M cell cycle phase transitions and telomere maintenance in human lung fibroblasts and is related to cellular senescence [18]. Hsa-miR-483-3p can also play an important role in lung cancer caused by radon exposure by regulating certain signaling pathways such as proliferative protein kinases (MAPK) and reactive oxygen species (ROS) [50].

Followed by, ROC curve analysis of these 10 hub genes was conducted based on dataset GSE43458, and we found that they all had certain diagnostic efficacy, which TOP2A ranked first. The results of Kaplan-Meier survival analysis also showed that these hub genes were significantly correlated with the prognosis of LUAD patients, suggesting that they may be potential prognostic biomarkers of LUAD. Lastly, we evaluated the expression patterns of TOP2A, CDCA7, TPX2 and COL1A1 in the main pathological stages and demonstrated that the expression of these four genes in tumors increased with the increase of the pathological stage of LUAD patients.

In summary, our study indicated that TOP2A, CDCA7, TPX2 and COL1A1 had strong relationships with each other and also owned better diagnosis and prognosis efficacy in LUAD. The mechanism and their mutual regulation network are worthy of further research and experiments. Anyway, all of our analyses
may provide some useful direction into the potential biomarkers and molecular mechanism of the occurrence and development of LUAD.

**Abbreviations**

LUAD, lung adenocarcinoma; GEO, Gene Expression Omnibus; DEGs, differentially expressed genes; GO, Gene Ontology; KEGG, The Kyoto Encyclopedia of Genes and Genomics; PPI, The protein-protein interaction; STRING, Search Tool for the Retrieval of Interacting Genes; NSLC, non-small cell lung cancer; SCLC, small cell lung cancer; LUSC, lung squamous cell carcinoma; LCC, large cell carcinoma; BP, biological process; MF, molecular function; CC, cellular component; ROC, receiver operating characteristic; AUC, area under curve; HR, hazard ratio.

**Declarations**

**Acknowledgements**

Not applicable.

**Funding**

This study was supported by grants provided by the Teaching Reform Research Project of Medical School, Wuhan University No.2020004 and Open Experimental Project of Wuhan University (2020).

**Availability of data and materials**

All data generated or analyzed during this study are included in this published article.

**Authors’ contributions**

All authors contributed to the study concept and design, as well as the interpretation of the data. MZ designed and managed the whole research; ZMW analyzed the data and wrote the main manuscript text; LNZ revised the manuscript. All authors read and approved the final version of the manuscript.

**Ethics approval and consent to participate**

Not applicable.

**Patient consent for publication**

Not applicable.

**Competing interests**

The authors declare that they have no competing interests.
References


Figures

Figure 1

The expression of DEGs in GSE32863, GSE43458 and GSE118370 via Venn diagrams.
The KEGG and GO enrichment analyses about 312 DEGs using Enricher online tool.
Figure 3

The PPI network constructed by STRING online database and the identification of 10 hub genes.

Figure 4

The interaction network between hub genes and its related miRNAs via the website miRWalk.
**Figure 5**

The ROC curves for 10 hub genes. AUC, area under the receiver operating characteristic curve.
Figure 6

The Kaplan-Meier plotter survival analyses of 10 hub
Figure 7

The expression violin plots of candidate biomarkers based on patient pathological stage using GEPIA