Integrated computational analysis revealed hub genes in lung adenocarcinoma

Muhammad Jamal  
Department of Immunology, School of Basic Medical Sciences, Wuhan University, Wuhan 430071, P.R. China

Abdul Saboor Khan  
National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, P.R. China

Hina Iqbal Bangash  
State Key Laboratory of Agriculture Microbiology, Huazhong Agricultural University, Wuhan 430070, P.R. China

Tian Xie  
Department of Immunology, School of Basic Medical Sciences, Wuhan University, Wuhan 430071, P.R. China

Tianbao Song  
Department of Immunology, School of Basic Medical Sciences, Wuhan University, Wuhan 430071, P.R. China

Yufei Lei  
Department of Immunology, School of Basic Medical Sciences, Wuhan University, Wuhan 430071, P.R. China

Jingyuan Li  
Department of Immunology, School of Basic Medical Sciences, Wuhan University, Wuhan 430071, P.R. China

Haiyan Tan  
Department of Gastroenterology, Renmin Hospital of Wuhan University, Wuhan, Hubei 430071, P.R. China

Qiuping Zhang (✉ qpzhang@whu.edu.cn)  
Department of Immunology, School of Basic Medical Sciences, Wuhan University, Wuhan 430071, P.R. China

Method Article

Keywords: LUAD, DEGs, functional enrichment analysis, PPI, hub genes, diagnosis, prognosis, miRNA-hub gene network, Immunohistochemistry staining
Abstract

Background

Lung cancer (LUCA) is the leading cause of cancer-related morbidities and mortalities globally. Despite the recent advancements in lung cancer research, understanding of the molecular mechanism underlying LUCA tumorigenesis and prognosis remains suboptimal. This study aims to identify the candidate biomarkers and therapeutic genes in lung cancer.

Methods

In this study, gene expression profiles of GSE30219, GSE33532, GSE32863 and GSE43458 were downloaded from GEO. The differentially expressed genes (DEGs) in LUAD tissue and normal lung tissue with a p-value < 0.05 and a |log fold change (FC)| >1.0 were identified by GEO2R. For functional enrichment analysis of these DEGs, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses were performed with KOBAS and DAVID tools. Next, the candidate hub genes were filtered out with Cytoscape using CytoHubba plugin. These hub genes were validated by (the Cancer Genome Atlas) TCGA-based gene expression analysis, protein-protein network interaction (PPI) analysis, survival analysis. Moreover, the expression of these genes in cancer and normal tissue was assessed in the Human Protein Atlas (HPA) database. In addition, miRNA network of the hub genes was constructed. Finally, DGIdb database was used to check the drug-targeting potentials of the hub genes.

Results

A total of 332 overlapping differentially expressed genes (DEGs) including 73 upregulated and 259 downregulated, respectively were identified. GO analysis revealed that the DEGs were principally regulating various cancer-associated functions and pathways. The module analysis revealed 55 hub genes in 4 modules. The survival analysis through Kaplan-Meier (KM) plotter indicated that the altered expression of these genes resulted in the poor overall survival (OS) of LUCA patients. Moreover, these genes show a differential expression on both protein and mRNA level in cancer patient compared to the normal. In addition, in addition, 6 potential microRNAs (miRNAs) interacting with hub genes were identified. Finally, a list of 117 therapeutic small molecules was tabulated that could facilitate LUCA treatment.

Conclusions

The findings of this study may help in the development of novel and reliable biomarkers for diagnosis, prognosis and therapeutic intervention for LUAD.
Introduction

LUCA is the second leading cause of cancer-related mortality and second to prostate and breast cancer in terms of morbidities [1, 2]. According to American Cancer Society statistics, 131,880 and 235,760 lung cancer incidences and deaths were reported in 2021 [3]. There are numerous risk factors associated with LUCA, such as tobacco consumption, genetic susceptibility, poor diet and occupation exposure to air pollution [4]. Based on histopathological appearance, LUCA is classified into two major subtypes; non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). NSCLC accounts for 85% of the total lung cancer, which is further classified into three major histological subtypes; lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC) and large cell carcinoma (LCC) [5]. LUAD is the primary histological subtype of lung cancer in China [6]. Surgery is the first-line treatment and effective option for LUAD [7]. Despite the advancements in surgical procedures, and targeted and immunotherapies, the diagnosis and prognosis remain dismal with the five-year survival rate of 10–15% [8], which is due to its diagnosis at advance or metastatic stage [9].

Understanding the molecular mechanism underlying LUAD is critical to improve its diagnosis and prognosis by identifying new reliable biomarkers and therapeutic sites. Recently, integrated bioinformatics analysis of the gene expression microarray datasets revealed various hub genes in several cancers [10–12]. However, the latest biomarker-related studies are inadequate, and the DEGs findings may be inconsistent because of the complicated gene regulatory mechanism underlying LUCA tumorigenesis. Moreover, drug resistance remains a bottleneck, often resulting in treatment failure and reduced overall survival (OS) and progression-free survival (PFS). Nevertheless, RNA-sequencing or microarray-based global gene expression profiling has facilitated the researchers to identify genes that are crucial in LUCA pathology [13]. Using computational approaches, Mengwei et al., identified 249 DEGs including 113 upregulated and 136 downregulated genes [14], analysis of the genes with altered expression between the cancer and normal tissue may help in the development of effective biomarkers.

In this study, we performed a multistep analysis to identify candidate genes in LUAD through integrated bioinformatics analysis. To this end, the gene expression profiles in four different gene expression datasets (GSE30219, GSE33532, GSE32863, and GSE43458) containing 263 LUAD patients and 122 normal samples were downloaded from the GEO database, screened. The DEGs were subsequently analyzed using the integrated bioinformatics tools, including GO, KEGG, and PPI network construction. From the PPI network, hub genes were identified and screened for the diagnostic and prognostic potentials. Moreover, the expression of the hub genes was validated in TCGA dataset and in lung cancer cell lines. Furthermore, the miRNA network of the hub genes was constructed. In addition, hub genes targeted by the small molecular drugs were also identified.

Materials Methods

Cell culture:
Human lung cancer cell lines A549 and PC9, and normal epithelial cell line BEAS-2B were obtained from ATCC and cultured in DMEM (Biological Industries, USA), supplemented with 10% FBS (Biological Industries) and kanamycin. Cell culture was maintained in humidified chamber in 5% CO$_2$ at 37°C.

**GEO datasets collection and data processing:**

The NCBI-GEO database contains freely available the gene expression profiles. GEO datasets information of the lung cancer was obtained from the GEO database. We selected 4 lung cancer datasets including GSE30219, GSE33532, GSE32863 and GSE43458. Datasets meeting the inclusion criteria defined as; (i) human as the organism of investigation (ii) LUAD samples, (iii) dataset containing tumor and normal samples, (iv) array-based expression profiling, the dataset with normal lung tissue sample $\geq$15 were downloaded from the GEO database. GEO2R a web-based tool to use for the comparison of gene expression in two different groups of the same samples under the same experimental conditions (). We used GEO2R to analyze the selected LUAD datatests and the resultant data was downloaded in tables.

We screen the genes using cut-off criteria as $|\log_{2}FC|>1$ and adjusted p-value of $<$0.05 as statistically significant to identify the DEGs. The average of the expression level was considered for the genes with multiple probes. Locus-by-locus volcano plot was used to distribute the DEGs. Overlapping DEGs among all the four datasets were integrated and represented with the Venn diagram. Locus-by-locus volcano plot and Venn diagram was generated by gplot and Venn diagram packages in R program, respectively [15]. The heatmap of the top20 up-and downregulated genes were generated with Heml [16].

**GO and KEGG pathway enrichment analyses**

DAVID can explore the biological process (BP), molecular function (MF), cellular component (CC). For the functional annotation and pathway enrichment analysis of the DEGs, GO, and KEGG was used through the DAVID database. For pathway enrichment analysis, we defined a p-value of $<$0.05 as a cut-off criterion. The list of upregulated and downregulated DEGs were submitted to the DAVID tool, and the top 10 terms for each function were selected. Circos plot is a web-based visualization tool to identify and analyze genomes' similarities and differences [17]. In this study, the GO function of the top 20 upregulated and downregulated DEGs are represented as Circos plot.

**Protein-protein interaction network construction and modules analysis**

Protein interaction networks can decipher the possible molecular mechanism of genes involved in various cellular processing pathways. We performed the protein-protein Interaction (PPI) network construction by using STRING (version 11.0). A confidence score of $>0.4$ was set as a threshold. Cytoscape (version 3.7.2) was used for the visualization and analysis of the PPI network. In order to reveal the sub-network of the PPI network, Molecular Complex Detection (MCODE) plugin with default parameters including degree cut off $>$10, node cut off$=0.4$, k-core$= 2$ and max depth$=100$ was set as selection criteria. Moreover, functional pathway enrichment analysis of the DEGs in the modules was performed using the KOBAS database with default parameters.
Identification of hub genes

Cytohubba can predict vital nodes and subnetworks in a given biological network using several topological analysis strategies (Chin et al., 2014). Cytohubba plugin of Cytoscape was used to identify the candidate hub genes among the overlapped DEGs identified based on protein network. Maximal clique centrality (MCC) was selected among the other computing methods to explore hub genes modules.

Survival analysis of the DEGs

We screened all the DEGs for their prognostic potential using the Kaplan-Meier plotter database, which is an online tool to assess the prognostic potential of cancer-associated genes (more than 54 thousand genes) based on their expression level in 21 cancer types including breast cancer (6234 samples), gastric cancer (1440 samples), ovarian (2190) and lung cancer (3452). LUAD patients were classified into high-expression and low-expression group based on the gene expression pattern. The analysis is described as survival curves according to the hazard ratio (HR), 95% confidence interval (95% CI), and log-rank P-values and survival prognosiscurves. Moreover, the survival analysis of the hub genes was validated in UCALAN, which contain the available TCGA patient survival data based on RNAseq-based gene expression profiling. The diagnostic value of the hub genes was calculated using Wilson/Brown method with a 95% confidence interval.

Gene expression analysis

Gene expression profiling interactive analysis (GEPIA), is based on the analysis of RNA sequencing expression data obtained from TCGA and GTEx projects containing 483 LUAD and 347 normal samples. In our study, we used this database, to analyse the expression of the DEGs in different LUAD stages. ANOVA was used to assess the statistical significance of the gene expression variation. A Pr (>F) <0.05 was considered statistically significant. In addition, we obtained the gene expression profiles of LUAD patients and normal from the TCGA database available on (https://portal.gdc.cancer.gov/). The RNA-sequencing expression data from TCGA was filtered and the expression of the hub genes were assessed.

Immunohistochemistry staining

The Human Protein Atlas (HPA) aims at mapping and providing free access to all the human proteins expressed in cells, tissues and organs by the integration of data derived from various omics technologies, including antibody-based imaging, mass spectrometry-based proteomics, and transcriptomics and systems biology. In this study, we used HPA to examine the expression of the hub genes expressed in human LUAD and normal bronchial tissue.

Hub genes-drugs interaction network

Using Drug-Gene Interaction Database (DGIB), hub genes that were promisingly targeted by the hub genes the were selected and only the drugs approved by the Food and Drug Administration (FDA) were included.
STITCH database contains the known and predicted interaction between proteins and chemicals. In order to construct the hub gene-drug interaction network, the web-based tool named STITCH was applied.

**Construction of miRNA regulatory network**

Gene-Cloud Biotechnology Information (GCBI) is a versatile tool that predicts the expression level of genes in diseases and healthy individuals, and possible interactions of genes with non-coding RNA (ncRNA) and transcription factors. miRTarbase contains more than 360,000 miRNA-target interactions. These interactions are validated by reporter assays, western blot, microarray and next-generation sequencing [18]. We employed GCBI and miRTarbase to predict the miRNAs interacting with hub genes.

**RNA Extraction and quantitative real-time polymerase chain reaction (qRT-PCR):**

Total RNA was extracted from A549, PC9 and BEAS-2B cells using TRIzol reagent following the manufacturer instruction (Invitrogen, USA). 2 µg total RNA was converted into complementary DNA (cDNA) using RevertAid RT reverse transcription kit (Thermo Fisher Scientific). The SYBR-Green supermix kit (TaKara) was used to perform real-time PCR experiment to quantify the gene expression. The thermal cycling procedure was 95°C for 10 min, followed by 40 cycles of 95°C for 10 s and 60°C for 30 s. The relative expression levels of the hub genes were determined by 2-ΔΔCt method and normalized to housekeeping genes; GAPDH [19]. All the qPCR reactions were performed in triplicate. The sequence of the primers used in this study is mentioned in table S1.

**Statistical analysis**

Gene expression data of the related genes is presented as mean ± SD and analyzed using the paired t-test. All the analysis was performed with Graphpad Prism (8.0). A p-value of <0.05 was considered as statistically significant. The diagnostic value of the hub genes in LUAD was assessed using ROC curves and AUC obtained by Wilson/Brown method with a 95% confidence interval.

**Results**

**Identification and screening of DEGs between LUAD tissue and normal tissue in each dataset**

We employed a multistep strategy to explore candidate genes, which are important in LUAD pathogenesis by integrating bioinformatics tools (figure 1). After the preliminary search, we identified 4 datasets; GSE30219, GSE33532, GSE32863 and GSE43458 based on the inclusion criteria. GSE32863 and GSE43458 datasets contained only LUAD patient gene expression profiles, whereas only LUAD pathological subtype patients in GSE30219 and GSE33532 were screened. A total of 263 LUAD and 123 normal lung tissue samples were obtained in this study. The information of the dataset used in this study is summarized in table 1.

After screening the datasets based on the cut-off criteria, a set of 9014 DEGs including 3182 and 5832 up- and down-regulated genes were identified between LUAD and normal lung tissue (table 1). There were
2656 DEGs, including 857 upregulated and 1799 downregulated genes in GSE30219, 3756 DEGs, including 1362 upregulated and 2394 downregulated in GSE33532, 1026 DEGs, including 319 upregulated and 707 downregulated in GSE43458, and 1576 DEGs, including 644 upregulated and 932 downregulated genes in GSE32863 dataset. The volcano plots representing the distribution of the DEGs in each dataset are shown in figure 2A-D. Next, we further identified the genes, which were overlapped among all the four datasets. A total of 73 and 259 overlapping upregulated and downregulated genes, respectively were identified among all the datasets, which are represented in Venn diagrams (figure 2E and 2F, table 2). Moreover, among these commonly shared DEGs we identified the top 20 upregulated and downregulated genes and displayed on the heatmap (figure 2G).

Functional pathway enrichment analysis

GO and KEGG pathway enrichment analysis deciphers the function and relationship of the genes with various biological functions. To assess the functional roles of the 73 upregulated and 257 downregulated genes, DAVID database was used (table S2). Based on the p-value and cut-off criteria, the DEGs were classified into three functions; biological function, molecular function (MF) and cellular function (CF). The biological function (BF) category of the GO analysis revealed the enrichment of the upregulated genes in cellular and metabolic processes, cell division, tissue development and regulation of biological processes (figure S1A). Regarding the CF, DEGs were enriched in extracellular region part, proteinaceous extracellular matrix, cell-cell junction. (figure S1B). Moreover, the DEGs were enriched in protein and chromatin binding and catalytic activity in MF category of the GO analysis (figure S1C). Circos plot showing GO enrichment of the 20 upregulated genes illustrates the role of these genes in collagen catabolic and metabolic processes, extracellular structure organization and tissue development (figure 3A). Based on KEGG enrichment analysis, the upregulated genes significantly enriched in the cytoplasm and extracellular region (figure 3B). Moreover, the downregulated genes were enriched in multicellular organismal processes and regulation of biological processes, metabolic process, immune system (BF) (figure S1D) extracellular matrix, extracellular region part, (CF) (figure S1E), and binding and catalytic activity (MF) (figure S1F). Circos plot showing GO enrichment of the 20 downregulated genes shows the regulation of vasculature and system development, angiogenesis, cell motility and regulation of cell proliferation (figure 3C). Based on KEGG enrichment analysis, the downregulated genes significantly enriched in multicellular organismal processes, developmental processes and extracellular region (figure 3D).

Protein-protein interaction (PPI) network analysis

The establishment of PPI network interaction of these DEGs were constructed based on the information obtained from STRING database. A list of 330 DEGs were submitted to STRING database and a network was obtained, which consisted a total number of 935 edges and 292 nodes (figure S2A). Significant cluster module identification via MCODE-plugin in Cytoscape revealed 4 modules containing 55 genes in the whole PPI network (figure 4A, 4B, 4C and 4D). We calculated the top-ranked hub genes by using the MCC. The top 4 functional clusters of modules were selected (module 1, MCODE score=12; module 2,
MCODE score=9; module 3, MCODE score=5.0; module 4, MCODE score=5). The number of edges, nodes and genes in 4 modules are listed in table 3. DAVID database was used to perform the KEGG pathway analysis of the genes in modules (table S3). The genes in module 1 were involved in cell cycle, cell division, p53 signaling pathways. The genes in module 2 were enriched in HIF-1, TNF and PI3-Akt signaling pathways. The genes in module 3 were involved in vascular smooth muscle contraction, protein digestion and absorption, relaxin signaling pathway and pathways in cancer. Moreover, the genes in module 4 were enriched in adherens junction, acute myeloid leukemia, PPAR signaling pathway and FoxO signaling pathway. Next, the cytohubba plugin in cytoscape was applied to identify the hub genes in the modules and obtained 10 hub genes (KIF20A, MELK, CCNB2, CDC20, TOP2A, PECAM1, CDH5, IL6, ZBTB16 and RNF144B) for LUAD. Among the 10 hub genes, 5 genes were upregulated and 5 genes were downregulated. STRING database was used to construct the protein interaction network of the hub genes with each other and with neighboring genes (figure S2B and S2C). The interaction network of the hub genes comprised up of 10 nodes, 17 edges with a p-value of 3.57e-05. In addition, hub genes coexpression analysis revealed possible interaction of among the hub genes (figure 4E).

**Validation of the prognostic and diagnostic potentials of the hub genes**

In order to corroborate the importance of the and hub genes in LUAD, the prognostic potential of the hub genes was investigated by using KM plotter database. The overall survival (OS) analysis curves of the hub genes are represented in figure 5A-5J. Among the 10 hub genes, the upregulation of KIF20A [HR=1.66 (1.46-1.88)], MELK [HR= 1.61 (1.41-1.82)], CCNB2 [1. 99 (1.74-2.26)], CDC20 [1.82 (1.6-2.07)] and TOP2A [1. 55 (1.36-1.76)] predicted the poorer OS of the LUAD patients, whereas the lower expression of PECAM [HR=0.61 (0.52-0.72)], CDH5 [HR=0.69 (0.61-0.78)], LEPREL1 [HR= 0.76 (0.67-0.86)], RFN144B [HR=0.58 (0.49-0.7)] and ZBTB16 [HR=0.72 (0.62-0.7)] were significantly related to the poorer OS. Moreover, the upregulation of KIF20A, MELK, CCNB2, CDC20 and TOP2A predicted a poor progression-free survival (figure S6A-6E), whereas the lower expression of PECAM1, CDH5, ZBTB16 and RNF144B predicted the poorer PFS (figure S6F-6I).

In addition, we verified the expression and prognostic potentials of the 330 DEGs using GEPIA and KM Plot, respectively and found a set of 49 genes with significantly high expression and prognostic value (figure S3A). The hub genes with highest hazard rate (HR) and high expression including the UBE2T, TYMS, PRC1, CDC45, and TPX2 with HR of 2.23, 2.32, 1.91, 1.84 and 1.86, respectively, which indicated poor prognosis in LUAD patients (figure S3B-S3F and table 4). Similarly, a set of 109 genes with low expression and the good prognosis was obtained (figure S3G) including PTPRM, PRICKLE2, CPED, LPL, and DCN with an HR of 0.53, 0.48, 0.49, 0.65 and 0.61, respectively (figure S3H-S3L). Next, we determined the diagnostic value of the hub genes based on their TCGA-based gene expression profiles and found that these genes possess high diagnostic accuracy with AUC (area under the receiver operating characteristic curve) values of 0.98, 0.97, 0.98, 0.98, 0.99, 0.98, 0.98, 0.93, 0.87 and 0.96 value ranging for KIF20A, MELK, CCNB2, CDC20, TOP2A, PECAM1, CDH5, IL6, ZBTB16 and RNF144B, respectively (figure S4).
Validation of the expression of hub genes

In order to validate the expression of hub genes in TCGA, fragments per kilobase of transcript per million mapped reads (FPKM) gene expression profiles from the TCGA database available at (https://portal.gdc.cancer.gov/) were downloaded filtered and merged to produce one file, which contained the gene expression data for both LUAD and normal samples. The expression of the hub genes was assessed. The result showed that KIF20A, MELK, CCNB2, CDC20 and TOP2A genes were significantly upregulated in LUAD patients compared to the normal samples (figure 6A). However, PECAM1, CDH5, LEPREL1, ZBTB16 and RNF144B were downregulated in LUAD patients (figure 6B).

Next, the expression of hub genes was determined in two lung cancer cell lines; A549 and PC9, and normal bronchial epithelial cells BEAS-2B. We found that the expression of KIF20A, MELK, CCNB2, CDC20 genes was upregulated in lung cancer cell lines compared to the normal bronchial epithelium cell (figure 6C), whereas, the expression of PECAM1, CDH5, LEPREL1 genes was reduced in the lung cancer cell lines compared to BEAS-2B (figure 6D).

Verification of the expression of hub genes on protein level

In addition, the expression of the hub genes was assessed at protein level in the LUAD and normal lung tissue and using HPA database. The expression of KIF20A, MELK, CCNB2, CDC20 and TOP2A was highly expressed in LUAD tissue compared to the normal lung tissue (figure 7A-7E). Contrarily, a reduced expression of PECAM1, CDH5, LEPREL1, ZBTB16 and RNF144B in LUAD as compared to the normal lung tissue (figure 7F-7J). The clinical information of the immunohistochemistry (IHC) in human proteome atlas is available in table S4.

Hub gene-drug interaction network

Using DGIdb for exploring the drug-gene interaction, a list of 117 potential drugs for treating LUAD was identified (table 5). These drugs are all FDA-approved drugs. Among the hub genes, TOP2A was found to interact with 97/118 drugs, either as an inhibitor or in an unknown mechanism, followed by MELK, which interacted with 18 drugs (table 5). The downstream interaction networks of TOP2A and MELK was constructed with STITCH (figure 8A-8B). TOP2A might have a downstream effect on Ciprofloxacin, Dexrazoxane, Etoposide and Amsacrine (figure 8A), whereas MELK might have effect on Dovitinib, Nintedanib, Sunitinib and Staurosporine (figure 8B).

miRNA network analysis:

Given the role of miRNA in regulating biological processes, we identified the miRNAs targets of the hub genes using mirRTarbase and GCBI databases. The miRNA network was constructed and visualized with Cytoscape, which consisted of 9 hub genes and 165 interacting miRNAs. The interaction of the miRNAs with their corresponding hub genes are represented in figure 9A. Based on their degree of closeness, the top 10 molecules in the interaction network were filtered. RNF144B, ZBTB16 and CDH5, LEPREL1 and TOP2A with a degree score of 30,30,27,25 and 16, respectively were the top 5 miRNA targeting genes.
Moreover, hsa-miR-16-5p, hsa-miR-128-3p, hsa-let-7f-5p, hsa-let-7c-5p and hsa-let-7a-5p, were the top 5 interactive miRNAs with a score of 52, 47, 46, 45.95 and 45.96, respectively (figure 9B). The miRNA network of the 9 hub genes constructed by using GCBI database is shown in figure S5.

Expression of hub genes in different cancer stages

The expression of the identified hub genes in different lung cancer stages were assessed with GEPIA database. We observed a significant difference in the expression of KIF20A (F value= 6.03 Pr >F=0.00049), CDC20 (F value= 4.95 Pr >F=0.00217), MELK (F value= 4.56 Pr >F=0.00368), CCNB2 (F value= 7.43 Pr >F=7.16E-05), and TOP2A in different stages of lung cancer (F value= 2.88 Pr >F=0.003) (figure 10A-10E). However, the differential expression of PECAM1, CDH5, LEPREL1, ZBTB16 and RFN144B was insignificant (figure 10F-10J). The expression of CDC20, MELK and TOP2A decreased with continuous progression of LUAD (figure 10B, 10C and 10E). Whereas, the expression of KIF20 was higher in stage1 and stage 4 compared to stage 2 and stage 3 (figure 10A), and the expression of CCNB2 was highest in stage 2 (figure 10D).

Genetic alteration of the hub genes

Investigation of the mutation in genes associated with biological functions are important. Several mutations in these genes can lead to LUCA [20] We investigated the genetic alteration in the hub genes using Cbioportal and identified various alteration in 19% of patients or samples. The highest frequency of alteration was observed in MELK, CDC20, TOP2A, PECAM1 and CDH5 (7%, 3%, 2.6%, 2.6% and 1.3%), respectively. These alterations included missense mutation, truncating mutation, amplification, deep deletion and no alteration (figure 11A). Among the alterations, amplification accounted for the highest percentage of the alterations followed by deep deletion and mutations (figure 11B).

Discussion

Lung cancer contributes to the highest proportion of cancer-related deaths worldwide. LUAD is the most common histological type of lung cancer. Despite the remarkable improvements in surgical procedures and medical therapy made over the last few years, the overall survival of LUAD patients remains poorer. The cause of such high mortality is mainly due to diagnosis of LUAD patients in advanced stages, which is primarily due to the lack of effective diagnostic tools. The dysregulated expression of numerous genes has been reported in LUAD. Gene expression profiling of the lung cancer patients vs normal individuals has facilitated the modern research to identify the genes that are important in disease diagnosis, prognosis, and therapy. Recently, advancements in bioinformatics tools, a growing amount microarray and high throughput sequencing data have conferred a simple and sophisticated platform to discover genetic alteration, identify DEGs and understand the molecular mechanism of tumor diagnosis and prognosis, and therapy [21].

In this study, using bioinformatics analysis of the four GEO datasets of lung cancer, we identified 332 DEGs including 73 up- and 259 down-regulated genes. The DEGs with elevated expression were enriched
in cellular and metabolic processes, cell division, tissue development and regulation of biological processes. The downregulated DEGs were enriched in biological functions such as multicellular organismal processes and regulation of biological processes, metabolic process, immune system. Pathway enrichment analysis via KEGG and GO revealed the role of these genes in various biological pathways related to lung cancer including developmental process, immune system processing, cellular component organization, and regulation of biological processes, cellular and metabolic processes.

PPI network of a list of 330 DEGs was constructed using STRING database, which consisted a total number of 935 edges and 292 nodes and identified 4 modules containing 55 genes (table 3). The functional enrichment of modules genes revealed that these genes regulated cell cycle, cell division, p53 signaling pathways, vascular smooth muscle contraction, protein digestion and absorption, relaxin signaling pathway and pathways, adherens junction, acute myeloid leukemia, PPAR signaling pathway and FoxO signaling pathway. These findings also highlighted important clue to exploring the molecular interaction involved in LUAD progression.

Top ranked hub genes were selected using Cytoscape. Next, the expression level of these hub genes in cancer and normal lung tissue was verified using GEPIA database. The expression results were in line with that of the GEO database. The differential expression of these hub genes in different LUAD stages were also assessed. Owing to the significant differential expression of CDC20, MELK and TOP2A in the early and advance LUAD patients (figure 10), we postulate that this finding may answer why the increased expression of CDC20, MELK and TOP2A related to poorer PFS in early LUAD stage (figure S6B and C). Whereas RNF144B was associated with better OS and PFS in advanced stages in LUAD patients (figure 10 and figure S6I). In addition, these genes also possess high high diagnostic accuracy with AUC <0.9 (figure S4). Therefore, we proposed whether CDC20, MELK, TOP2A and RNF144B might be considered as biomarker in the diagnosis and prognosis of LUAD. We further identified the expression of these hub genes based on the TCGA gene expression data of the LUAD patients and normal samples, which further corroborated the importance of these hub genes in LUAD. The processes of tumor invasion and metastasis are governed by complex molecular mechanism. In this study, the dysregulated expression of hub genes was obvious LUAD tissues compared with normal lung tissue (figure 7), suggesting the possible role of these genes in tumorigenesis and progression. Despite many studies have described these hub genes related with diagnosis, prognosis and treatment of various cancers, the precise regulatory mechanism of LUAD tumorigenesis and progression remain elusive.

Compared with other studies, this study integrates various datasets with a larger sample size including 263 cancer and 122 normal lung tissues, respectively. Moreover, gene expression data of LUAD patient have been investigated in this study.

Kinesin-like family member 20A (KIF20A) plays a crucial role in organization of central spindle at anaphase and cytokinesis stages. It is an oncogene, which facilitate cancer cell proliferation and inhibiting apoptosis thus facilitating the malignant phenotype of LUAD[22] . The overexpression of KIF20A predicted the poor OS and PFS in LUAD patients in our study (figure 5B and figure S6A). Moreover,
the increased expression of KIF20A was found in lung cancer cell lines (figure 6C and D), meaning the tumorigenic role of KIF20A in LUAD. Maternal Embryonic Leucine Zipper Kinase (MELK) is a cell cycle-dependent kinase protein that has been reported in various cancers including SCLC[23]. However, the precise tumorigenic mechanism of MELK needs further studies. In our study, we found that the increased expression of MELK was related with poorer OS and PFS (figure 5C and figure S6B). In addition, we also found numerous potentially interacting miRNAs (figure S5). Cyclin B2 (CCNB2) involved in G2/M transition, has been associated with tumor degree, size, stage and metastasis as well as poor prognosis in NSCLC [24]. However, the role of CCNB2 in LUAD tumorigenesis is not well-elucidated. Type 2 topoisomerase alpha (TOP2A) encodes a nuclear enzyme called DNA Topoisomerase II, which is involved in DNA replication. TOP2A has been reported in many cancers including lung cancer [25]. The downregulation of TOP2A reduced the proliferation, migration and invasion abilities in lung cancer cells [25]. T2a and HER-2 of TOP2A in epithelial ovarian carcinoma was first reported by [26], whereas in this study, the alteration in TOP2A mainly included amplification, missense mutation and truncated mutations (figure 11). The cell cycle division cycle protein 20 homolog (CDC20), a vital cell division regulator has been implicated in lung cancer [27]. The high expression of CDC20 have been documented to be associated with poor prognosis in multiple cancers [28-30]. In our study, the higher expression of CDC20 favored a poorer OS and FPS, suggesting that the diagnostic potentials of CDC20 (figure 5E and figure S6E).

In addition to the upregulated hub genes, downregulated hub genes were also identified in the present study. Platelet endothelial cell adhesion molecule-1 (PECAM-1) a member of immunoglobulin superfamily, is expressed on the vascular endothelium and has been involved in the late progression of metastatic tumors. The microvessel density (MVD) of PECAM-1 was associated with tumor stages and a high PECAM-1 MVD had worse overall survival in either adenocarcinoma or EGFR mutation subgroups [31]. We found a reduced expression of PECAM-1 in the later LUAD stages compared to the initial stages and the reduced expression is related to poor PFS and OS (figure 10, figure 5G and figure S6F). The expression of cadherin 5 (CDH5) also known as vascular endothelial cadherin (VE-cadherin) is an angiogenic factor. The aberrant expression of VE-cadherin has been found to correlated with lymph node metastasis in NSCLC, suggesting its role in tumor metastasis[32]. However, the biomarker potentials of CDH5 in LUAD is not documented. Indeed, we evidenced a differential expression of CDH5 in LUAD tissues as compared to normal lung tissue (figure 6B) and the decrease expression was correlated with worse OS and PFS (figure 5H and figure S6G). Leprecan-like 1 (LEPREL1), also known as P3H2 is abundantly found in basement membrane [33]. The lower expression of LEPREL1 has been reported in hepatocellular carcinoma (HCC) without any correlation with the clinical parameters of HCC [34]. P3H3, a member of the prolyl hydroxylase family exhibits tumor suppressor activity in lung cancer. P3H3 is silenced by the DNA methylation and the ectopic expression of P3H3 circumvented cell proliferation, colony formation, metastasis and invasion of lung cancer [35]. We found the decreased expression of LEPREL1 in A549 and PC9 cell lines as compared to the BEAS cell line and in cancer patients as compared to healthy person (figure 6). Moreover, the lower expression of LEPREL predicted the poorer OS and PFS (figure 5I and figure S6J). Further studies are needed to explore whether LEPREL1 is an
independent factor in the diagnosis and/or development of LUAD. Zinc finger and BTB domain-containing 16 (ZBTB16) a transcription factor, which regulate cell proliferation, differentiation, development of organ, stem cell maintenance and development of innate immunity, has been reported both as a tumor suppressor and oncogene in various cancer [36]. However, its role in lung cancer is not well-explained. We found the decreased expression of ZBTB16 in TCGA-dataset and in lung cancer cell lines and the decreased expression was also associated with poor OS and PFS of lung cancer patients (figure 5K, figure 6 and figure S6H). RNF144B also known as PIR2, is an E3-ubiquitin ligase involved in the regulation of cell apoptosis and proliferation. It has been recognized as important biomarker in endometrial cancer [36, 37]. We observed a reduced expression of RNF144B in LUAD tissue, TCGA dataset and in lung cancer cell lines (figure 6). Moreover, the downregulation of RNF144B was correlated with poor OS of lung cancer patients (figure 5J and figure S6I). These findings suggest the biomarker potentials of PECAM1, CDH5, ZBTB16 and RNF144B in LUAD. Moreover, we further confirmed the differential expression of hub genes in LUAD tissue vs normal lung tissue using HPA (figure 7), which further corroborated the importance of these hub genes in LUAD.

In this study, a list of 150 potential drugs for treating LUAD was identified. Among the hub genes, TOP2A and MELK was potentially targeted by 96 and 17 drugs, respectively. Most of these drugs are inhibitors in nature. Only a few of the TOP2A and MELK targeting drugs have been used for the treatment of lung cancer such as etoposide (topoisomerase II inhibiting anti-cancer drug) has been used for treating NSLC[38], dexrazoxane (a catalytic inhibitor of DNA topoisomerase II that lack double stranded DNA break activity) possess with protective function against anthracycline-induced cardiotoxicity and extravasation injuries, has been administered with cyclophosphamide, adriamycin and vincristine (Advanced small cell lung cancer treated with CAV [39], sunitinib (a tyrosine kinase inhibitor possessing broad antitumor efficacy) [40] and nintedanib (inhibitor of receptor tyrosine kinase and nonreceptor tyrosine kinase) along with docetaxel is administered for NSCLC patients after first-line chemotherapy in European countries [41]. Further investigation and clinical trials are needed to explore the mechanism of these drugs that could be effective against lung cancer. Consequently, based on the findings of this study and aforementioned studies, these genes are crucial in LUAD tumorigenesis and we hope that future studies may concentrate on investigating the detailed mechanism of these hub genes, especially PECAM1, CDH5, ZBTB16, RNF144B, TOP2A and MELK.

In this study, numerous miRNAs interacting with hub genes have been identified. The top 5 interactive miRNAs include hsa-miR-16-5p, hsa-miR-128-3p, hsa-let-7f-5p, hsa-let-7c-5p and hsa-let-7a-5p. Frequent deletion or down-regulation of hsa-miR-16-5p has been found in LUAD and SCC. This miRNA is believed to target cyclins. The upregulation of hsa-miR-16-5p induces Rb-dependent cell cycle arrest in NSCLC [42]. hsa-miR-128-3p contributes to the development of chemoresistance and metastasis NSCLC patients by suppressing the expression of the inhibitors of Wnt/β-catenin and TGF-β pathways. Whereas the antagonism of hsa-miR-128-3p reverse metastasis and chemoresistance in NSCLC patients [43]. hsa-let-7f-5p hsa-let-7a-5p showed the lowest expression in NSCLC as compared to chronic obstructive pulmonary disease patients and healthy individuals [44]. However, the interaction of these miRNAs is based on the in-silico prediction and needs experimental validations. Therefore, in view of the vital role of
these 10 hub genes based on the findings of this study and the previously published studies described above, further research is needed to understand the precise molecular mechanisms in the tumorigenesis and prognosis of LUAD, especially exploring MELK, CCNB2, PECAM1, ZBTB16 and RNF144B.

Despite the interesting findings, there are still several shortcomings in the present study. Firstly, the data is obtained from the public databases and datasets, not produced by the authors. Secondly, the gene expression analysis was performed in TCGA and GEPIA databases, and in cell lines but not validated in the lung cancer patients. Moreover, due to the lack of functional genetic studies, we couldn’t investigate the impact of the hub gene-miRNA networks on the detection and treatment of LUAD. Nevertheless, the integrated bioinformatics approach used in this study provide more robust results compared to single dataset studies, and opens up new research avenues to investigate these hub genes in future.

In summary, we identified significant DEGs in lung cancer by integrated bioinformatics analysis. Consequently, these DEGs were screened out and 10 hub genes were identified, which were further validated in TCGA, GEPIA and HPA databases, and lung cancer cell lines. Moreover, a list of drugs interacting with the hub-genes were identified. Finally, candidate miRNAs interacting with hub genes were identified. These hub genes may pave the way for the development of novel and reliable biomarkers in LUAD.

Declarations

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable

CONSENT FOR PUBLICATION

Not applicable

COMPETING INTERESTS

The author disclose no competing interest.

AUTHOR CONTRIBUTIONS

MJ and QZ designed and conceived the study. AS and HB performed the bioinformatics data analysis. TS, ZH, and LY participated in figure design drafted the manuscript. TH and LY revised the language and critically revised the manuscript. All of the authors read and approved the final manuscript.

FUNDING

This work was supported by National Natural Science Foundation of China No. 81770180.

ACKNOWLEDGMENTS
We are grateful to Dr. Khaista Rahman, Research Assistant at the State Key Laboratory of Agricultural Microbiology, Huazhong Agricultural University Wuhan, China for critically reviewing and improving the language of the manuscript.

**AVAILABILITY OF THE DATA AND MATERIALS**

The databases used in this study for the analysis are publicly available on the following websites:

- GEO database (https://www.ncbi.nlm.nih.gov/geo/)
- KEGG (https://www.genome.jp/kegg/)
- DAVID (https://david.ncifcrf.gov/)
- KOBAS (http://kobas.cbi.pku.edu.cn/)
- STRING (https://string-db.org/)
- STITCH (http://stitch.embl.de/cgi/network.pl)
- Cytoscape (http://www.cytoscape.org/)
- Kaplan-Meier plotter (http://kmplot.com/analysis/index.php?p=service&cancer=lung)
- UACLAN (http://ualcan.path.uab.edu/index.html)
- GEPIA (http://gepia.cancer-pku.cn/)
- TCGA database (https://portal.gdc.cancer.gov/)
- GCB (https://www.gcbi.com.cn/gclib/html/index)
- miRTarbase (http://mirbase.cuhk.edu.cn/php/index.php).

**References**

1. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics*, 2019. CA: a cancer journal for clinicians, 2019. **69**(1): p. 7–34.
2. Jemal, A., et al., *Annual report to the nation on the status of cancer, 1975–2005, featuring trends in lung cancer, tobacco use, and tobacco control*. JNCI: Journal of the National Cancer Institute, 2008. **100**(23): p. 1672–1694.
3. Siegel, R.L., et al., *Cancer Statistics, 2021*. CA: A Cancer Journal for Clinicians, 2021. **71**(1): p. 7–33.
4. Malhotra, J., et al., *Risk factors for lung cancer worldwide*. European Respiratory Journal, 2016. **48**(3): p. 889–902.
5. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries*. CA: a cancer journal for clinicians, 2018. **68**(6): p. 394–424.
6. Ding, Y., et al., *Comparative study on the mutational profile of adenocarcinoma and squamous cell carcinoma predominant histologic subtypes in Chinese non-small cell lung cancer patients*. Thoracic cancer, 2020. **11**(1): p. 103–112.
7. Molina, J.R., et al. *Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship*. in *Mayo clinic proceedings*. 2008. Elsevier.
8. Herbst, R.S., D. Morgensztern, and C. Boshoff, *The biology and management of non-small cell lung cancer*. Nature, 2018. **553**(7689): p. 446–454.
9. Imielinski, M., et al., *Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing.* Cell, 2012. 150(6): p. 1107–1120.

10. Gao, X., et al., *Identification of key candidate genes and biological pathways in bladder cancer.* PeerJ, 2018. 6: p. e6036.

11. Liu, L., et al., *Identification of key genes and pathways of thyroid cancer by integrated bioinformatics analysis.* Journal of cellular physiology, 2019. 234(12): p. 23647–23657.

12. Wang, Y., et al., *Integrated bioinformatics analysis reveals key candidate genes and pathways in breast cancer.* Molecular medicine reports, 2018. 17(6): p. 8091–8100.

13. Sun, Z. and P. Yang, *Gene expression profiling on lung cancer outcome prediction: present clinical value and future premise.* Cancer Epidemiology and Prevention Biomarkers, 2006. 15(11): p. 2063–2068.

14. Ni, M., et al., *Identification of Candidate Biomarkers Correlated With the Pathogenesis and Prognosis of Non-small Cell Lung Cancer via Integrated Bioinformatics Analysis.* Frontiers in Genetics, 2018. 9(469).

15. Wang, Y., et al., *Genome-scale long noncoding RNA expression pattern in squamous cell lung cancer.* Scientific reports, 2015. 5: p. 11671.

16. Deng, W., et al., *HemI: a toolkit for illustrating heatmaps.* PloS one, 2014. 9(11): p. e111988.

17. Krzywinski, M.I., et al., *Circos: An information aesthetic for comparative genomics.* Genome Research, 2009.

18. Chou, C.-H., et al., *miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions.* Nucleic Acids Research, 2017. 46(D1): p. D296-D302.

19. Arocho, A., et al., *Validation of the 2-∆∆Ct calculation as an alternate method of data analysis for quantitative PCR of BCR-ABL P210 transcripts.* Diagnostic Molecular Pathology, 2006. 15(1): p. 56–61.

20. Carper, M.B. and P.P. Claudio, *Clinical potential of gene mutations in lung cancer.* Clinical and translational medicine, 2015. 4(1): p. 33.

21. Altman, R., *Current Progress in bioinformatics 2016.* Briefings in bioinformatics, 2016. 17(1): p. 1–1.

22. Zhao, X., et al., *Overexpression of KIF20A confers malignant phenotype of lung adenocarcinoma by promoting cell proliferation and inhibiting apoptosis.* Cancer medicine, 2018. 7(9): p. 4678–4689.

23. Inoue, H., et al., *Effective growth-suppressive activity of maternal embryonic leucine-zipper kinase (MELK) inhibitor against small cell lung cancer.* Oncotarget, 2016. 7(12): p. 13621.

24. Qian, X., et al., *CCNB2 overexpression is a poor prognostic biomarker in Chinese NSCLC patients.* Biomedicine & Pharmacotherapy, 2015. 74: p. 222–227.

25. Kou, F., et al., *TOP2A Promotes Lung Adenocarcinoma Cells' Malignant Progression and Predicts Poor Prognosis in Lung Adenocarcinoma.* Journal of Cancer, 2020. 11(9): p. 2496–2508.

26. Mano, M.S., et al., *Rates of topoisomerase II-alpha and HER-2 gene amplification and expression in epithelial ovarian carcinoma.* Gynecologic oncology, 2004. 92(3): p. 887–895.
27. Kato, T., et al., Overexpression of CDC20 predicts poor prognosis in primary non-small cell lung cancer patients. Journal of surgical oncology, 2012. 106(4): p. 423–430.

28. Kim, Y., et al., MAD2 and CDC20 are upregulated in high-grade squamous intraepithelial lesions and squamous cell carcinomas of the uterine cervix. International journal of gynecological pathology, 2014. 33(5): p. 517–523.

29. Moura, I.M., et al., High CDC 20 expression is associated with poor prognosis in oral squamous cell carcinoma. Journal of oral pathology & medicine, 2014. 43(5): p. 517–523.

30. Wu, W.-j., et al., CDC20 overexpression predicts a poor prognosis for patients with colorectal cancer. Journal of translational medicine, 2013. 11(1): p. 142.

31. Kuang, B.-h., et al., The prognostic value of platelet endothelial cell adhesion molecule-1 in non-small-cell lung cancer patients. Medical oncology, 2013. 30(2): p. 536.

32. Chen, H., et al., Expression and significance of VE-cadherin and E-cadherin in non-small cell lung cancer, 2011, Eur Respiratory Soc.

33. Järnum, S., et al., LEPREL1, a novel ER and Golgi resident member of the Leprecan family. Biochemical and biophysical research communications, 2004. 317(2): p. 342–351.

34. Wang, J., et al., LEPREL1 expression in human hepatocellular carcinoma and its suppressor role on cell proliferation. Gastroenterology research and practice, 2013. 2013.

35. Li, Y., et al., Collagen prolyl hydroxylase 3 has a tumor suppressive activity in human lung cancer. Experimental cell research, 2018. 363(1): p. 121–128.

36. Jin, Y., H.Z. Nenseth, and F. Saatcioglu, Role of PLZF as a tumor suppressor in prostate cancer. Oncotarget, 2017. 8(41): p. 71317.

37. Zhou, Q., et al., Pir2/Rnf144b is a potential endometrial cancer biomarker that promotes cell proliferation. Cell death & disease, 2018. 9(5): p. 1–10.

38. Furuse, K., Platinum/oral etoposide therapy in non-small cell lung cancer. Oncology, 1992. 49(Suppl. 1): p. 63–70.

39. Feldmann, J., et al. Advanced small cell lung cancer treated with CAV (cyclophosphamide + Adriamycin®+ vincristine) chemotherapy and the cardioprotective agent dexrazoxane (ADR-529, ICRF-187, Zinecard®). in Proc Ann Meet Am Soc Clin Oncol. 1992.

40. Gridelli, C., et al., Sorafenib and sunitinib in the treatment of advanced non-small cell lung cancer. Oncologist, 2007. 12(2).

41. Shiratori, T., et al., Effect of nintedanib on non-small cell lung cancer in a patient with idiopathic pulmonary fibrosis: A case report and literature review. Thoracic Cancer, 2020.

42. Bandi, N., et al., miR-15a and miR-16 are implicated in cell cycle regulation in a Rb-dependent manner and are frequently deleted or down-regulated in non–small cell lung cancer. Cancer research, 2009. 69(13): p. 5553–5559.

43. Cai, J., et al., Simultaneous overactivation of Wnt/β-catenin and TGFβ signalling by miR-128-3p confers chemoresistance-associated metastasis in NSCLC. Nature communications, 2017. 8(1):
44. Leidinger, P., et al., *High-throughput qRT-PCR validation of blood microRNAs in non-small cell lung cancer*. Oncotarget, 2016. 7(4): p. 4611.

**Tables**

Please see the supplementary files section to view the tables.