COL1A1, COL1A2, COL3A1 and DCN are Potential Biomarkers to Predict Progression from Smokers to Suffering from Lung Adenocarcinoma

Jixin Chen
Guangzhou University of Traditional Chinese Medicine: Guangzhou University of Chinese Medicine
Shuqi Chen
Artemisinin Research Center, Guangzhou University of Chinese Medicine, Guangzhou, Guangdong 510405, P.R. China
Feiye Wang
The Second Clinical College, Guangzhou University of Chinese Medicine, Guangzhou, Guangdong 510405, P.R. China
Sumei Wang (wangsumei198708@163.com)
The Second Clinical College, Guangzhou University of Chinese Medicine, Guangzhou, Guangdong 510405, P.R. China. Department of Oncology, Clinical and Basic Research Team of TCM Prevention and Treatment of NSCLC, Guangdong Provincial Hospital of Chinese Medi https://orcid.org/0000-0002-2011-4019
Wanyin Wu
The Second Clinical College, Guangzhou University of Chinese Medicine, Guangzhou, Guangdong 510405, P.R. China. Department of Oncology, Clinical and Basic Research Team of TCM Prevention and Treatment of NSCLC, Guangdong Provincial Hospital of Chinese Medi

Research

Keywords: lung adenocarcinoma, smoker, biomarker, bioinformatic analysis, gene expression

Posted Date: September 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-76976/v1

Abstract

Background: To explore novel related genes and potential biomarkers that predict progression from smokers to lung adenocarcinoma (LA).

Methods: Three datasets from GEO (Gene Expression Omnibus) database were used to identify differentially expressed genes (DEGs) between LA tissue (LAT) and normal tumor adjacent tissue (TAT). The overlap of DEGs could be found and enriched in gene oncology (GO) and pathways to discover the potential biological mechanisms. Protein-protein interaction (PPI) network was applied to find the relationship among proteins. Survival analysis contributed to the definiteness of key genes. The expression of key genes in LA patients who smoke was verified. Furthermore, genetic alterations, co-expression and pathways of key genes were explored. To obtain more information, key genes were further analyzed in immune infiltration, drug target and the distribution of single cell in LA.

Results: 245 DEGs were revealed in 3 datasets from GEO. In Kaplan Meier plotter, we found that high expression of COL1A1, COL1A2 and COL3A1 was associated with poorer survival while low expression of DCN was contributed to poorer survival in LAs who smoke. Thus, three up-regulated genes (COL1A1, COL1A2, COL3A1) and one down-regulated gene (DCN) were defined as key genes. Their genetic alterations were more common in female LA smokers and co-expression genes/proteins of them mainly functioned at extracellular matrix. Furthermore, COL1A1, COL1A2, COL3A1 genes had a common targeted drug called Collagenase clostridium histolyticum (DB00048) and DCN gene had a targeted drug called Tromethamine (DB03754). In the Single Cell Expression Atlas of EMBL-EBL, COL3A1 gene was specifically highly expressed in female LA patients with brain metastasis.

Conclusions: COL1A1, COL1A2, COL3A1 and DCN could be regarded as novel potential biomarkers that predict progression from smokers to lung adenocarcinoma.

Background

According to the investigation of the World Health Organization (WHO), the new cases of lung cancer around the world were 2093876 and its death was 1761007 in 2018. It is well known that cigarette smoking is a main reason leading to lung cancer. In the year 2011, cigarette smoking accounted for 80.2% of lung cancer deaths among adults (≥ 35 years old) in the United States[1]. Besides, smoking status even affected the curative effect of lung adenocarcinoma (LA) and altered the transcriptome of non-involved lung tissue in LA[2, 3]. People consider that the major pathological type of smoking-related lung cancer was squamous carcinoma while adenocarcinoma was rare, previously. With advances in diagnostic techniques and improved cigarette design, the proportion of lung adenocarcinoma among smokers has increased obviously in recent decades[4, 5]. In the large population-based cohorts, the percentage of adenocarcinoma with current smokers ranged from 30–42% in the United States [6]. Considering the development trend, we selected lung adenocarcinoma as the research object.

It has been reported that smoking could shorten the survival of LA patients with TP53 mutation and impact the DNA methylation level at genome-wide[7, 8]. Besides, it was found that aberrant activation of mast cells and $CD4^+$ memory T cells also contributed to LA development and progression. However, the signaling pathways and driver genes in smoking-associated lung adenocarcinoma remain uncertain [9]. In the last few decades, we mainly focused on the effect of smoking between lung cancer patients and normal people. But why non-smokers also suffered from lung cancer? Besides, we also knew that some people who smoke had no lung cancer during their whole life. Therefore, a further understanding of the biological characteristics and differences between cancerogenic and normal smokers may provide some references for the prevention of lung cancer and the treatment of patients who smoke.

To explore the potential mechanism and biomarkers of smokers who had lung adenocarcinoma, we compared the lung adenocarcinoma tissue (LAT) with normal tumor adjacent tissue (TAT) in smokers by analyzing datasets from GEO. The overlap of DEGs could be found and their gene oncology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were analyzed to discover the potential biological mechanism. Protein-protein interaction (PPI) network was also applied to observe the relationship among proteins. Hub genes were concluded by Mcode (Molecular Complex Detection) and cytohubba modules in Cytoscape (a public bioinformatics software, version 3.7.1). By using Kaplan Meier plotter, survival analysis was analyzed to define key genes. Information from The Cancer Genome Atlas (TCGA) and ONCOMINE database were used to explore genetic alterations and to verify the expression of key genes in LA patients who smoke. Furthermore, the pathways and co-expression of key genes were explored. To obtain more information, key genes were further analyzed in immune infiltration, drag target and the distribution of single cell in LA.

Methods

Datasets selection

Based on the pre-retrieval results of objective microarray data, the GEO database was identified as the source of datasets. GEO is a comprehensive public database containing high-throughput gene expression data for multiple species and diseases, as well as chip and microarray data [10]. We selected lung adenocarcinoma patients who smoke as study subjects, and compared gene expression differences between LAT and TAT to find predictive biomarkers. Datasets containing the following criteria for patient inclusion and exclusion may be included. Inclusion criterion: 1. Patients whose genes were sequenced from lung tissue were pathologically diagnosed as lung adenocarcinoma; 2. Smoking status included current or former smoking status, unlimited years of smoking; 3.The microarray data type was limited to gene sequencing data; 4.The age and gender of the patients were not limited. Exclusion criterion: 1. Datasets that couldn't be analyzed online using GEO2R;2. Gene symbol of datasets cannot be extracted or converted.

Identification of DEGs

Smoking samples of each dataset were grouped into “normal” and “cancer”, and analyzed by GEO2R to search the top 250 DEGs with default setting. The results were downloaded and managed in EXCEL software. According to the criterion of |logFC|>1 and adjusted *P*-value (adj.*P*) < 0.01, the DEGs of each dataset were further screened and divided into up-regulated and down-regulated genes by the positive and negative of logFC. Venny2.1 was applied to find the overlap of up-regulated and down-regulated DEGs in all datasets.

Functional enrichment analysis

In order to elucidate the functional profiles and pathways of the DEGs, we used WebGestalt (WEB-based Gene SeT AnaLysis Toolkit) to obtain the enriched biological process (BP), cellular component (CC), molecular function (MF) and KEGG pathway. $P < 0.05$ was considered statistically significant. The results were presented in bubble diagrams with the ggplot2 and dplyr packages in R 3.6.3 language.

PPI network construction and module analysis

Firstly, the PPI network of DEGs was presented using the STRING online database, in which the moderate confidence (an interaction with a combined score > 0.4) was statistically significant [11]. Then, the MCODE plug-in in Cytoscape software (version 3.7.1) was used to cluster the PPI network based on topological principles to determine the closely related regions and the most significant modules in PPI network was found out (degree cutoff = 2, node score cutoff = 0.2, k-core = 2, max depth = 100) [12, 13].

Hub genes selection

In the identification of hub genes, Cytohubba plug-in of Cytoscape plays an important role in ranking the nodes in PPI network based on the different network characteristics of 11 topological analysis methods [14]. In this study, we sorted the top 10 hub genes from high to low according to the four topological methods of MCC (maximum clique centrality), MNC (maximum neighborhood component), Degree and EPC (edge percolated component), and identified the common genes as hub genes. In addition, based on the EGA, TCGA and GEO databases, an online tool Kaplan Meier-plotter was applied to analyze overall survival influence in the hub genes in lung adenocarcinoma patients who smoke [15]. We defined these genes with overall survival significance ($P < 0.05$) as key genes, showing potential roles to be used as predictive biomarkers.

Key genes analysis in lung adenocarcinoma who smoke

Oncomine online database was used to conduct third-party verification of key gene expression in lung adenocarcinoma patients who smoke to improve reliability [16]. Besides, the cBioPortal online tool is used to study the genetic changes in the key genes in lung adenocarcinoma patients who smoke [17]. The online analysis tool GeneMANIA was used to build a network including co-expression and pathways of these key genes. This database contains complete datasets from GEO, BioGRID and other databases as well as genome datasets for specific functions of organisms [18].

Further analysis of key genes

To analyze the relationship between tumor and immune system, TISIDB online website was used to discover the situations of immune invasion of these genes among lung adenocarcinoma patients, based on the interaction among 28 kinds of lymphocytes. Meanshile, drug targets were also analyzed in TISIDB. Moreover, the distribution of key genes in lung adenocarcinoma tissue cells was described through the Single Cell Expression Atlas of EMBL-EBL in order to obtain more meaningful information.

Results

245 DEGs were identified between LAT and TAT

Three datasets about gene expression of LAT and TAT were found from GEO including GSE10072 [19], GSE31547 and GSE32863 [20]. Specifically, based on GPL94 ([HG-U133A] Affymetrix Human Genome U133A Array), the GSE10072 dataset included 58 LAT samples (42 smokers) and 49 TAT samples (34 smokers). The GSE31547 dataset, also relied on GPL94, contained 30 LAT samples (19 smokers) and 20 TAT samples (15 smokers). The GSE32863 dataset, while depended on GPL6884 (Illumina HumanWG-6 v3.0 expression beadchip), contained 58 LAT samples (29 smokers) and 30 TAT samples (28 smokers). Totally, our study included 90 LAT smoking samples and 77 TAT smoking samples to study DEGs. Using software including GEO2R and EXCEL, a total of 2402 DEGs(669 in GSE10072, 433 in GSE31547, and 1300 in GSE32863) containing 245 overlapping DEGs were extracted from three datasets by comparing the differences between LAT and TAT in smokers. More specific, there were 196 down-regulated genes (Fig. 1A) and 49 up-regulated genes (Fig. 1B) between LAT and TAT.

Enrichment analysis of the 245 DEGs

To reveal the biological functions of the 245 DEGs, we performed functional annotation and pathway enrichment analysis using the WebGestalt online tool. On the one hand, the main results (FDR < 0.01) of GO and KEGG enrichment of 49 up regulated DEGs were as follows: 1. BP: scollagen fibril organization, extracellular matrix organization, extracellular structure organization and tissue development (Fig. 2A). 2. MF: extracellular matrix structural constituent, extracellular matrix structural constituent conferring tensile strength, platelet-derived growth factor binding, structural molecule activity, protease binding and growth factor binding (Fig. 2B). 3. CC: fibrillar collagen trimer, banded collagen fibril, extracellular matrix, collagen-containing extracellular matrix, complex of collagen trimers, endoplasmic reticulum lumen, collagen trimer, extracellular matrix component (Fig. 2C). KEGG: Protein digestion and absorption and ECM-receptor interaction (Fig. 2D). On the other hand, the main results (FDR < 0.01) of GO and KEGG enrichment of 196 down regulated DEGs were as follows: 1. BP: biological adhesion, cell adhesion, cardiovascular system development, vasculature development, blood vessel development, blood vessel morphogenesis, angiogenesis, circulatory system development, tube development and regulation of inflammatory response (Fig. 2E). 2. MF: amyloid-beta binding, peptide binding, glycosaminoglycan binding, signaling receptor binding, amide binding, cytokine binding, transforming growth factor beta binding, growth factor binding, low-density lipoprotein, particle binding and identical protein binding

(Fig. 2F). 3. CC: extracellular matrix, secretory granule, cell surface, intrinsic component of plasma membrane, collagen-containing, extracellular matrix, integral component of plasma membrane, cytoplasmic vesicle part, secretory vesicle, secretory granule membrane and cytoplasmic vesicle membrane (Fig. 2G). 4. KEGG: Complement and coagulation cascades (Fig. 2H).

Twenty genes were selected by using PPI network

The STRING database was used to predict the potential relationship of 245 DEGs at the protein level (combined score > 0.4). The PPI network was exported from the STRING database and interpreted by Cytoscape software, including 218 nodes and 753 edges (Fig. 3A). In addition, according to the established MCODE plug-in parameters, the most important PPI network module was selected, which was composed of 20 nodes and 76 edges (Fig. 3B).

COL1A1, COL1A2, COL1A3 and DCN were identified as key genes

In order to further identify the hub genes, Cytohubba plug-in was used to sort the top 10 nodes according to the established four topological analysis methods (Table 1). A total of 9 genes (IL6, SPP1, COL1A1, COL1A2, COL3A1, VWF, CTGF, COMP and DCN) were identified as hub genes, respectively. To identify key genes, Kaplan Meier plotter was applied to perform survival analysis. The significant genes are potential to be used as predictive biomarkers. In LA smokers, we found that high expression of COL1A1 ($P = 0.016$), COL1A2 ($P = 0.015$) and COL3A1 ($P = 0.04$) was associated with poorer survival while low expression of DCN ($P = 0.012$) was contributed to poorer survival. However, other genes made no sense in survival analysis (Fig. 4). Thus, COL1A1, COL1A2, COL1A3 and DCN were identified as key genes.

Table 1 | Hub genes for highly differentiated expressed genes ranked in Cytohubba plugin of Cytoscape

Rank	MCC	MNC	Degree	EPC
1	COL3A1	IL6	IL6	IL6
2	COL1A2	COL3A1	COL3A1	SPP1
3	COL1A1	COL1A2	COL1A2	COL1A1
4	SPP1	COL1A1	COL1A1	COL1A2
5	IL6	VWF	VWF	COL3A1
6	VWF	SPP1	SPP1	VWF
7	DCN	**ADRB2**	**ADRB2**	CTGF
8	CTGF	DCN	DCN	COMP
9	COMP	COMP	COMP	DCN
10	**ADRB2**	CTGF	CTGF	**IGFBP3**

The bold genes are non-shared genes.

The probable mechanisms of COL1A1, COL1A2, COL1A3 and DCN alterations in LAT

To verify the expression of the key genes, 3 studies from Oncomine online database were used to further validate the expression of these 4 genes (Fig. 5A). Comparing to TAT in smokers, COL1A1, COL1A2 and COL3A1 were highly expressed significantly ($P < 0.001$), while DCN was lowly expressed significantly ($P < 0.001$), in LAT. All key genes had obvious expressive dissimilarities in smokers between LA samples and normal lung tissue samples, which was in accordance with the results from GEO database. Regarding the genetic alteration, 4 hub genes were altered in 65 (12.8%) of 508 lung adenocarcinoma patients who smoke (TCGA, firehose legacy) and only samples with genetic alteration were presented in Fig. 5B. Among the 4 genes, COL1A1 and COL1A2 were mainly altered by amplification. However, COL3A1 was mainly altered by missense mutation. Besides, DCN was mainly altered by deep deletion and missense mutation and its genetic alteration seemed to occur in female. There was no obvious correlation among the four genetic alterations. However, in Fig. 5C, genetic alterations of these key genes significantly occurred in female LA smokers by Chi-squared test (P-value = 1.913×10^{-3}, FDR = 0.0421). Including co-expression (84.90%) in 325 studies and pathways (15.10%) in 6 studies, an interaction network for the 4 genes with 20 proteins/genes was generated via GeneMANIA. Based on the synthesis score, the ranking order from high to low was CD36, ITGA11, COL6A3, COL5A2, LUM, COL5A1, SPARC, TGFB3, MXRA5, JUNB, POSTN, FN1, FBN1, VEGFD, THBS2, ITGB1, COL6A1, CDH11, GTF3A and FAP. According to the FDR value, the top seven credible functions of this network were extracellular matrix organization, extracellular

structure organization, extracellular matrix, proteinaceous extracellular matrix, extracellular matrix disassembly, collagen and extracellular matrix part (Fig. 5D).

COL1A1, COL1A2, COL3A1 and DCN genes could be applied as predictive biomarkers in LA

We defined COL1A1, COL1A2, COL3A1 and DCN genes as key genes and further analyses were performed. Spearman correlation test was conducted to explore the correlation between genes and lymphocyte infiltration. We found that among 517 LA samples, the infiltration abundance of 28 kinds of lymphocytes was almost positively correlated with the expression of the 4 key genes [21]. For the up-regulated genes, the top 5 correlations of COL1A1, COL1A2, COL3A1 had better consistency and they were central memory CD8 T cell, regulatory T cell, memory B cell, natural killer cell, and natural killer T cell. For the down-regulated genes, the top 5 correlations of DCN were mast cell, type 1 T helper cell, natural killer cell, macrophage and regulatory T cell (Fig. 6A). Based on the information collected from DrugBank database, COL1A1, COL1A2, COL3A1 genes had a common targeted drug called Collagenase clostridium histolyticum (DB00048) and DCN gene had a targeted drug called Tromethamine (DB03754) (Fig. 6B). Via single cell RNA sequencing of 176 lung adenocarcinoma patient-derived cells (t-SNE Perplexity = 15) [22], we analyzed the cell distributions in diseases (Fig. 7A), sex (Fig. 7B) and metastatic sites (Fig. 7C). Furthermore, the expression levels of COL1A1 (Fig. 7D), COL1A2 (Fig. 7E), COL3A1 (Fig. 7F) and DCN (Fig. 7G) in these cell distributions were also presented. By comparing these seven results, we found that COL1A1 was more expressed in male LA patients, COL1A2 was more expressed in female LA patients with brain metastases. Excitingly, COL3A1 was specifically expressed highly in female LA patients with brain metastases.

Discussion

In lung cancer, lots of researches have shown that many signaling pathways could be activated by smoke, including β adrenergic receptor-mediated (β-ARs), nicotinic acetylcholine receptor-mediated (nAChRs), nuclear factor-KB (NF-KB), epidermal growth factor receptor (EGFR) and gamma aminobutyric acid (GABA) signaling pathways [23]. However, understanding the mechanisms that lead to lung adenocarcinoma in smokers remains a hard work. At the genetic level, smokers and nonsmokers were accurately identified in LA based on a support vector machine (SVM) classification model constructed from 27 characteristic genes with significant enrichment of cancer proteoglycan and Ras signaling pathway [24]. It was suggested that 7 mRNAs including CYP17A1, PKHD1L1, RPE65, NTSR1, FETUB, IGFBP1 and G6PC might be used as prognostic indicators in smokers with LA [9]. However, very few researchers focus on the development from normal lung tissue to LA tissue in smokers. Therefore, our research is committed to discover potential biomarkers that predict progression from smokers to lung adenocarcinoma.

In this study, a total of 245 DEGs (196 down-regulated genes and 49 up-regulated genes) were identified. In GO and KEGG enrichment analysis of 196 down-regulated genes, we found that these genes exist in

the extracellular matrix and the endoplasmic reticulum lumen in the cytoplasm. They are mainly involved in the biological process of the construction of extracellular matrix and tissues development through protein digestion and absorption and extracellular matrix receptor action pathways, catalyzing the synthesis of extracellular matrix structural components, as well as the combination of growth factors and proteases. However, in GO and KEGG enrichment analysis of 49 up-regulated genes, we found that existing in extracellular matrix, cytoplasmic vesicles, secretory vesicles and plasma membrane, these genes catalyze the binding of various molecules through the complement system pathway, which is mainly involved in cell or molecule adhesion, blood vessel formation and development, and the regulation of inflammatory response. Thus, it has been seen that 245 DEGs mainly acted on the tumor microenvironment which is essentially composed of genetically abnormal cells surrounded by blood vessels, fibroblasts, immune cells, stem cells and extracellular matrix (ECM) [25]. The complement pathway is a type of innate immunity that mainly supplements immunoglobulin and enhances the ability of immune cells to clear by promoting inflammation and attacking pathogen cell membranes [26]. In the tumor microenvironment, complement regulates both pro-tumor and anti-tumor pathways. It has been proved that complement activation via a C3a receptor pathway mediates lung cancer progression while RNA interference with CD59 synthesis can inhibit the growth and metastasis of lung adenocarcinoma cells [27, 28]. Dysregulated complement activation is a key link between inflammation, the suppression of antitumour immune responses and the promotion of tumorigenesis [29]. Furthermore, It was suggested that complement inhibitors or activators combined with targeted therapy or immunotherapy have promising prospects in the treatment of lung adenocarcinoma [26].

After a series of screening, COL1A1, COL1A2, COL3A1 and DCN genes were defined as key genes. COL1A1 (Collagen Type I Alpha 1 Chain) and COL1A2 (Collagen Type I Alpha 2 Chain) are protein coding genes that encode the pro-chains of type I collagen (COL1) which is related to osteogenesis imperfect [30]. COL3A1 (Collagen Type III Alpha 1 Chain) is another kind of collagen coding gene and its mutation is responsible for Ehler-Danlos syndrome type IV [31]. DCN, also called Decorin, is a protein coding gene and its mutation was regarded as a main aetiological agent of Congenital stromal corneal dystrophy (CSCD) [32]. Genetic alterations of these key genes were more common in female LA smokers. The network of co-expression and pathway indicated that these 20 genes/proteins were mainly functioned in extracellular matrix (especially collagen). CD36 was the most relevant pathway gene/protein with COL1A1 and COL1A2. CD36 could mediate the related pathway to inhibit angiogenesis which is the basis of tumor growth and metastasis [33]. Thus, we showed that COL1A1 and COL1A2 might promote tumor progression by inhibiting CD36 related pathways.

Regarding the immune infiltration analysis, we found that COL1A1, COL1A2, COL3A1 mainly regulate central memory CD8 T cell, regulatory T cell, memory B cell, natural killer (NK) cell, and natural killer T (NKT) cell while DCN mainly regulates mast cell, type 1 T helper cell, natural killer cell, macrophage and regulatory T cell. These immune cells act different roles in tumor immune microenvironment. Central memory CD8 T cell plays an important role in immunocytotherapy that it was obtained from the patient's body, proliferated in vitro and then transferred back to the body to achieve the anti-tumor effect [34]. For

patients who were unresponsive to PD-L1 or cytotoxic T lymphocyte-associated protein 4 (CTLA-4), regulatory T cell was accelerated to express by antibody-mediated depletion of immune checkpoint 4-1BB in order to modulate an antitumor immune response [35]. Memory B cells secreted high level of immunoglobulin against tumor antigens that accumulated in regional lymph nodes partly caused by PD-L1 blockade [36]. Studies have shown that patients with low NK cells activity had an increased risk of lung cancer, and injecting multiple allogeneic NK cells tended to have a better prognosis [37, 38]. Based on hematopoietic stem cell-engineer, invariant NKT cell, a potent immune cell for targeting cancer, changed its disadvantage of low level in cancer patients and developed an original therapy proved with long-term effect and no toxicity in vivo [39]. Mast cells can both anti-tumor through tumor infiltration can directly affect the proliferation and invasion of tumor cells and promote tumor through the establishment of the tumor microenvironment and regulating tumor cell immune response and whether anti-tumor or promote tumor depends on cancer types, tumor progression and the location of immune cells in tumor [40, 41]. Particularly, it was suggested that abnormal activation of mast cells led to lung immune dysfunction in smokers, which also contributes to tumor development and progression [42]. Type 1 T helper cell mainly secretes cytokines Interferon-γ (IFN-γ) which is also beneficial to tumor cells such as facilitating tumor growth, altering immune resistance of tumor and promoting immunosuppressive tumor microenvironment [43, 44]. The ability of macrophage to mount an effective antitumor response was governed by metabolism meanwhile its metabolism can be actively reprogramed by the tumor microenvironment via metabolites, cytokines or other signaling mediators so that the anticancer effect of macrophages was reduced[45]. Thus, inhibition of this reprogramming process has become a new approach for tumor therapy.

The immune environment infiltrated by these genes has potential value in tumor development and treatment. It is the importance of these genes in tumors has prompted us to look for relevant targets to inhibit them. In drug target analysis, COL1A1, COL1A2, COL3A1 genes had a common targeted drug called Collagenase clostridium histolyticum (DB00048) and DCN gene had a targeted drug called Tromethamine (DB03754). However, both of them haven't been applied in cancer therapy. Furthermore, single cell sequencing data was applied to explore expression specificity of these genes. It was displayed that COL1A1 gene more often expressed in male LA patients while COL1A2 and DCN genes more often expressed in female LA patients with brain metastases and COL3A1 gene was specific high expression in female LA patients with brain metastases. Via IHC, Liu Y et al proved that COL1A1 and COL3A1 were significantly high expression in brain tumor tissue metastasized from LA [46]. Due to the specific expression of COL3A1, it has the potential to predict brain metastases in LA. However, no studies have described the relationship between other genes (COL1A2 and DCN) and brain metastases.

Although potential biomarkers were found in our study, there are some deficiencies as followed. Firstly, limited to a clinical condition of smoker, we only choose ONCOMINE data to verify the expression key genes and didn't analyze the correlation of key genes and tumor stages. Both of them can further enhance veracity and richness in our study. Secondly, due to the limited experimental conditions, we didn't conduct experimental verification of the key genes, and some of these genes were only verified by the

experimental results of published literature. Thirdly, the concrete functions and mechanisms of how key genes induce LA are still unclear. Thus, more studies are needed to clarify those questions.

Conclusions

By bioinformatics analysis of the gene expression profile of smokers in LA, four core molecules were identified potentially associated with the pathogenesis from normal to LA in smokers, including 3 up regulated genes (COL1A1, COL1A2, COL3A1) and a down regulated gene (DCN). Besides, their genetic alterations were more common in female LA smoker and co-expression genes/proteins of them mainly functioned in extracellular matrix. Especially, COL3A1 is a potential biomarker to predict brain metastasis in lung adenocarcinoma. These key genes may be regarded as novel potential biomarkers that predict progression from normal to LA in smokers. However, due to the limitations of this study, further studies are needed to elucidate the biological function of these genes in the progression from normal to LA in smokers.

Abbreviations

LA, lung adenocarcinoma; GEO, Gene Expression Omnibus; DEGs, differentially expressed genes; LAT, LA tissue; TAT, tumor adjacent tissue; GO, gene oncology; PPT, Protein-protein interaction; WHO, World Health Organization; KEGG, Kyoto Encyclopedia of Genes and Genomes; TCGA, The Cancer Genome Atlas; BP, biological process; CC, cellular component; MF, molecular function; MCC, maximum clique centrality; MNC, maximum neighborhood component; EPC, edge percolated component; β-ARs, β adrenergic receptor-mediated; nAChRs, nicotinic acetylcholine receptor-mediated; NF-KB, nuclear factor-KB; EGFR, epidermal growth factor receptor; GABA, gamma aminobutyric acid; SVM, support vector machine; ECM, extracellular matrix; COL1A1, Collagen Type I Alpha 1 Chain; COL1A2, Collagen Type I Alpha 2 Chain; COL1, pro-chains of type I collagen; COL3A1, Collagen Type III Alpha 1 Chain; DCN, Decorin; CSCD, Congenital stromal corneal dystrophy; NK, natural killer; NKT, natural killer T cell; CTLA-4, cytotoxic T lymphocyte-associated protein 4; IFN-γ, Interferon-γ.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Publicly available datasets and online tools were analyzed and used in this study. http://www.ncbi.nlm.nih.gov/geo; http://www.ncbi.nlm.nih.gov/geo/geo2r;https://bioinfogp.cnb.csic.es/tools/venny/index.html; http://www.webgestalt.org/;https://string-db.org/; http://www.oncomine.com;https://www.cbioportal.org/; http://genemania.org/;http://kmplot.com/analysis/; http://cis.hku.hk/TISIDB/index.php; https://www.ebi.ac.uk/gxa/sc/home.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by grants from the National Natural Science Foundation of China (81974543, 81903991), the Guangzhou science and technology plan project (201804010149, 202002030155), the Guangdong Natural Science Foundation of China (2019A1515011362), the Major Research Projects in First-class Disciplines of Guangzhou University of Chinese Medicine (A1260619111001), the Chinese medicine science and technology research project of Guangdong Provincial Hospital of Chinese Medicine (YN2019MJ09), and the Guangdong Provincial Key Laboratory of Clinical Research on Traditional Chinese Medicine Syndrome (ZH2020KF03).

Authors' contributions

Jixin Chen (gzucmcjx@126.com) is responsible to design and initiate this study and prepare the manuscript. Feiye Wang (15350737060@163.com) is involved in data acquisition. Shuqi Chen (1134338593@qq.com) is responsible for the statistical analysis. Sumei Wang (wangsumei198708@163.com) is responsible to design this study and edit the manuscript. Wanyin Wu (wwanyin@126.com) is responsible for concept and initiation of this study.

Acknowledgements

Not applicable.

References

1. Siegel RL, Jacobs EJ, Newton CC, Feskanich D, Freedman ND, Prentice RL, Jemal A. Deaths Due to Cigarette Smoking for 12 Smoking-Related Cancers in the United States. JAMA Intern Med. 2015;175:1574–6.
2. Hotta K, Kiura K, Takigawa N, Kuyama S, Segawa Y, Yonei T, Gemba K, Aoe K, Shibayama T, Matsuo K, et al. Sex difference in the influence of smoking status on the responsiveness to gefitinib monotherapy in adenocarcinoma of the lung: Okayama Lung Cancer Study Group experience. J Cancer Res Clin Oncol. 2009;135:117–23.

3. Pintarelli G, Noci S, Maspero D, Pettinicchio A, Dugo M, De Cecco L, Incarbone M, Tosi D, Santambrogio L, Dragani TA, Colombo F. Cigarette smoke alters the transcriptome of non-involved lung tissue in lung adenocarcinoma patients. Sci Rep. 2019;9:13039.
4. Thun MJ, Lally CA, Flannery JT, Calle EE, Flanders WD, Heath CW Jr. Cigarette smoking and changes in the histopathology of lung cancer. J Natl Cancer Inst. 1997;89:1580–6.
5. Sakao Y, Miyamoto H, Oh S, Takahashi N, Inagaki T, Miyasaka Y, Akaboshi T, Sakuraba M. The impact of cigarette smoking on prognosis in small adenocarcinomas of the lung: the association between histologic subtype and smoking status. J Thorac Oncol. 2008;3:958–62.
6. Wakelee HA, Chang ET, Gomez SL, Keegan TH, Feskanich D, Clarke CA, Holmberg L, Yong LC, Kolonel LN, Gould MK, West DW. Lung cancer incidence in never smokers. J Clin Oncol. 2007;25:472–8.
7. Aisner DL, Sholl LM, Berry LD, Rossi MR, Chen H, Fujimoto J, Moreira AL, Ramalingam SS, Villaruz LC, Otterson GA, et al. The Impact of Smoking and TP53 Mutations in Lung Adenocarcinoma Patients with Targetable Mutations-The Lung Cancer Mutation Consortium (LCMC2). Clin Cancer Res. 2018;24:1038–47.
8. Bakulski KM, Dou J, Lin N, London SJ, Colacino JA. DNA methylation signature of smoking in lung cancer is enriched for exposure signatures in newborn and adult blood. Sci Rep. 2019;9:4576.
9. Zhou D, Sun Y, Jia Y, Liu D, Wang J, Chen X, Zhang Y, Ma X. Bioinformatics and functional analyses of key genes in smoking-associated lung adenocarcinoma. Oncol Lett. 2019;18:3613–22.
10. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30:207–10.
11. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43:D447–52.
12. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011;27:431–2.
13. Bandettini WP, Kellman P, Mancini C, Booker OJ, Vasu S, Leung SW, Wilson JR, Shanbhag SM, Chen MY, Arai AE. MultiContrast Delayed Enhancement (MCODE) improves detection of subendocardial myocardial infarction by late gadolinium enhancement cardiovascular magnetic resonance: a clinical validation study. J Cardiovasc Magn Reson. 2012;14:83.
14. Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol. 2014;8(Suppl 4):11.
15. Li T, Gao X, Han L, Yu J, Li H. Identification of hub genes with prognostic values in gastric cancer by bioinformatics analysis. World J Surg Oncol. 2018;16:114.
16. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia. 2004;6:1–6.
17. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci

Signal. 2013;6:pl1.

18. Montojo J, Zuberi K, Rodriguez H, Kazi F, Wright G, Donaldson SL, Morris Q, Bader GD. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. Bioinformatics. 2010;26:2927–8.
19. Landi MT, Dracheva T, Rotunno M, Figueroa JD, Liu H, Dasgupta A, Mann FE, Fukuoka J, Hames M, Bergen AW, et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. PLoS One. 2008;3:e1651.
20. Selamat SA, Chung BS, Girard L, Zhang W, Zhang Y, Campan M, Siegmund KD, Koss MN, Hagen JA, Lam WL, et al. Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. Genome Res. 2012;22:1197–211.
21. Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, Hackl H, Trajanoski Z. Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. Cell Rep. 2017;18:248–62.
22. Kim KT, Lee HW, Lee HO, Kim SC, Seo YJ, Chung W, Eum HH, Nam DH, Kim J, Joo KM, Park WY. Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. Genome Biol. 2015;16:127.
23. Wen J, Fu JH, Zhang W, Guo M. Lung carcinoma signaling pathways activated by smoking. Chin J Cancer. 2011;30:551–8.
24. Liu Y, Ni R, Zhang H, Miao L, Wang J, Jia W, Wang Y. Identification of feature genes for smoking-related lung adenocarcinoma based on gene expression profile data. Onco Targets Ther. 2016;9:7397–407.
25. Becker JC, Andersen MH, Schrama D, Thor Straten P. Immune-suppressive properties of the tumor microenvironment. Cancer Immunol Immunother. 2013;62:1137–48.
26. Kleczko EK, Kwak JW, Schenk EL, Nemenoff RA. Targeting the Complement Pathway as a Therapeutic Strategy in Lung Cancer. Front Immunol. 2019;10:954.
27. Kwak JW, Laskowski J, Li HY, McSharry MV, Sippel TR, Bullock BL, Johnson AM, Poczobutt JM, Neuwelt AJ, Malkoski SP, et al. Complement Activation via a C3a Receptor Pathway Alters CD4(+) T Lymphocytes and Mediates Lung Cancer Progression. Cancer Res. 2018;78:143–56.
28. Lin MC, Shen CH, Chang D, Wang M. Inhibition of human lung adenocarcinoma growth and metastasis by JC polyomavirus-like particles packaged with an SP-B promoter-driven CD59-specific shRNA. Clin Sci (Lond). 2019;133:2159–69.
29. Berraondo P, Minute L, Ajona D, Corrales L, Melero I, Pio R. Innate immune mediators in cancer: between defense and resistance. Immunol Rev. 2016;274:290–306.
30. Marini JC, Forlino A, Bächinger HP, Bishop NJ, Byers PH, Paepe AD, Fassier F, Fratzl-Zelman N, Kozloff KM, Krakow D, et al. Osteogenesis imperfecta. Nature Reviews Disease Primers. 2017;3:17052.
31. Pepin M, Schwarze U, Superti-Furga A, Byers PH. Clinical and genetic features of Ehlers-Danlos syndrome type IV, the vascular type. N Engl J Med. 2000;342:673–80.

32. Mellgren AE, Bruland O, Vedeler A, Saraste J, Schönheit J, Bredrup C, Knappskog PM, Rødahl E. Development of congenital stromal corneal dystrophy is dependent on export and extracellular deposition of truncated decorin. Invest Ophthalmol Vis Sci. 2015;56:2909–15.
33. Best B, Moran P, Ren B. VEGF/PKD-1 signaling mediates arteriogenic gene expression and angiogenic responses in reversible human microvascular endothelial cells with extended lifespan. Mol Cell Biochem. 2018;446:199–207.
34. Casati A, Varghaei-Nahvi A, Feldman SA, Assenmacher M, Rosenberg SA, Dudley ME, Scheffold A. Clinical-scale selection and viral transduction of human naïve and central memory CD8 + T cells for adoptive cell therapy of cancer patients. Cancer Immunol Immunother. 2013;62:1563–73.
35. Freeman ZT, Nirschl TR, Hovelson DH, Johnston RJ, Engelhardt JJ, Selby MJ, Kochel CM, Lan RY, Zhai J, Ghasemzadeh A, et al. A conserved intratumoral regulatory T cell signature identifies 4-1BB as a pan-cancer target. J Clin Invest. 2020;130:1405–16.
36. Kamata T, Yoshida S, Takami M, Ihara F, Yoshizawa H, Toyoda T, Takeshita Y, Nobuyama S, Kanetsuna Y, Sato T, et al. Immunological features of a lung cancer patient achieving an objective response with anti-programmed death-1 blockade therapy. Cancer Sci. 2020;111:288–96.
37. Lin M, Liang S, Wang X, Liang Y, Zhang M, Chen J, Niu L, Xu K. Percutaneous irreversible electroporation combined with allogeneic natural killer cell immunotherapy for patients with unresectable (stage III/IV) pancreatic cancer: a promising treatment. J Cancer Res Clin Oncol. 2017;143:2607–18.
38. Choi SI, Lee SH, Park JY, Kim KA, Lee EJ, Lee SY, In KH. Clinical utility of a novel natural killer cell activity assay for diagnosing non-small cell lung cancer: a prospective pilot study. Onco Targets Ther. 2019;12:1661–9.
39. Zhu Y, Smith DJ, Zhou Y, Li YR, Yu J, Lee D, Wang YC, Di Biase S, Wang X, Hardoy C, et al. Development of Hematopoietic Stem Cell-Engineered Invariant Natural Killer T Cell Therapy for Cancer. Cell Stem Cell. 2019;25:542–57.e549.
40. Derakhshani A, Vahidian F, Alihasanzadeh M, Mokhtarzadeh A, Lotfi Nezhad P, Baradaran B. Mast cells: A double-edged sword in cancer. Immunol Lett. 2019;209:28–35.
41. Khazaie K, Blatner NR, Khan MW, Gounari F, Gounaris E, Dennis K, Bonertz A, Tsai FN, Strouch MJ, Cheon E, et al. The significant role of mast cells in cancer. Cancer Metastasis Rev. 2011;30:45–60.
42. Li X, Li J, Wu P, Zhou L, Lu B, Ying K, Chen E, Lu Y, Liu P. Smoker and non-smoker lung adenocarcinoma is characterized by distinct tumor immune microenvironments. Oncoimmunology. 2018;7:e1494677.
43. Adorini L. Interleukin-12, a key cytokine in Th1-mediated autoimmune diseases. Cell Mol Life Sci. 1999;55:1610–25.
44. Mojic M, Takeda K, Hayakawa Y. The Dark Side of IFN-γ: Its Role in Promoting Cancer Immunoevasion. *Int J Mol Sci* 2017, 19.
45. Mehla K, Singh PK. Metabolic Regulation of Macrophage Polarization in Cancer. Trends Cancer. 2019;5:822–34.

46. Liu Y, Carson-Walter EB, Cooper A, Winans BN, Johnson MD, Walter KA. Vascular gene expression patterns are conserved in primary and metastatic brain tumors. J Neurooncol. 2010;99:13–24.

Figures

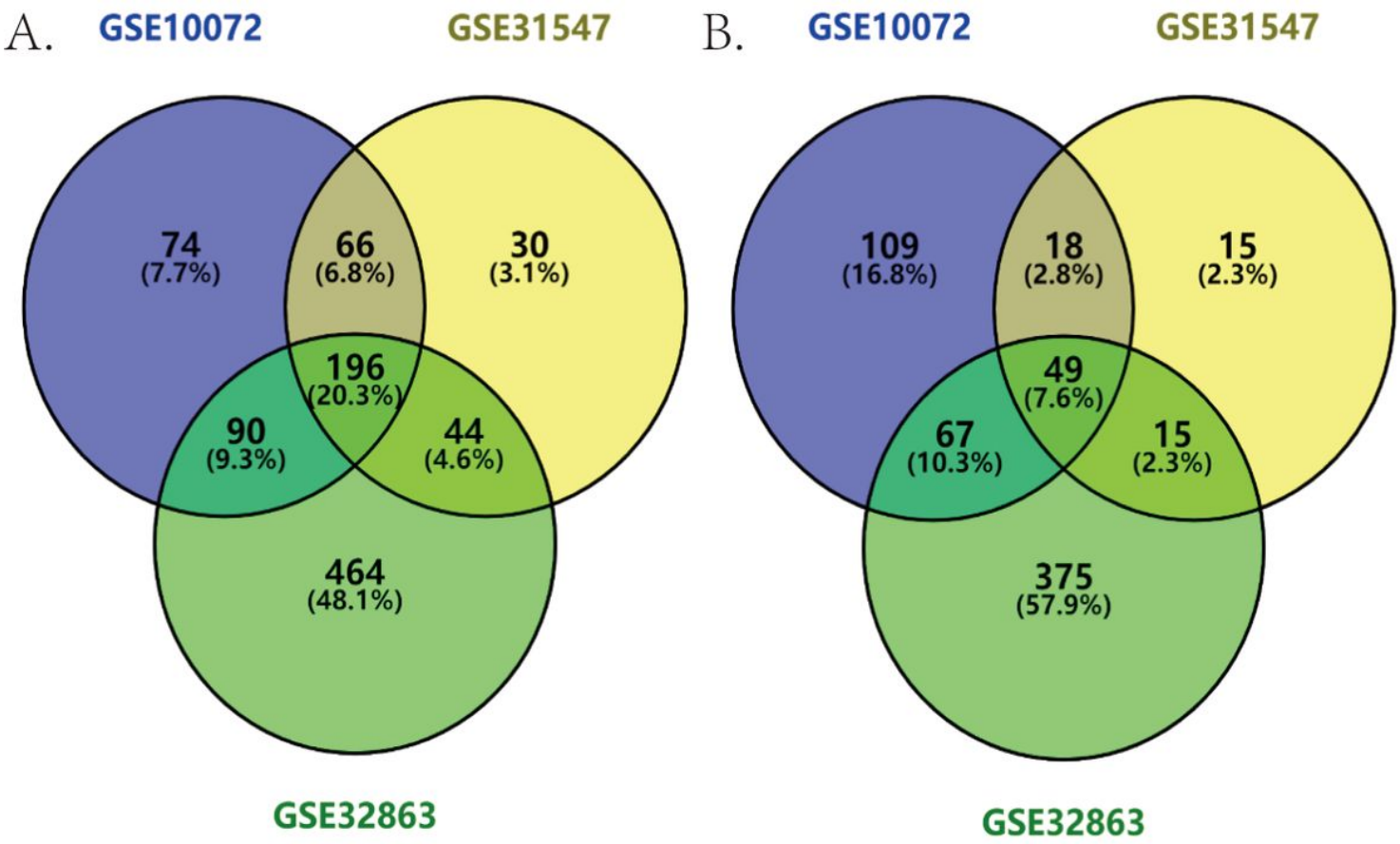

Figure 1

A.Venn diagram of overlapping 196 down regulated DEGs from GSE10072, GSE31547 and GSE32863 datasets;B.Venn diagram of overlapping 49 up regulated DEGs from GSE10072, GSE31547 and GSE32863 datasets.

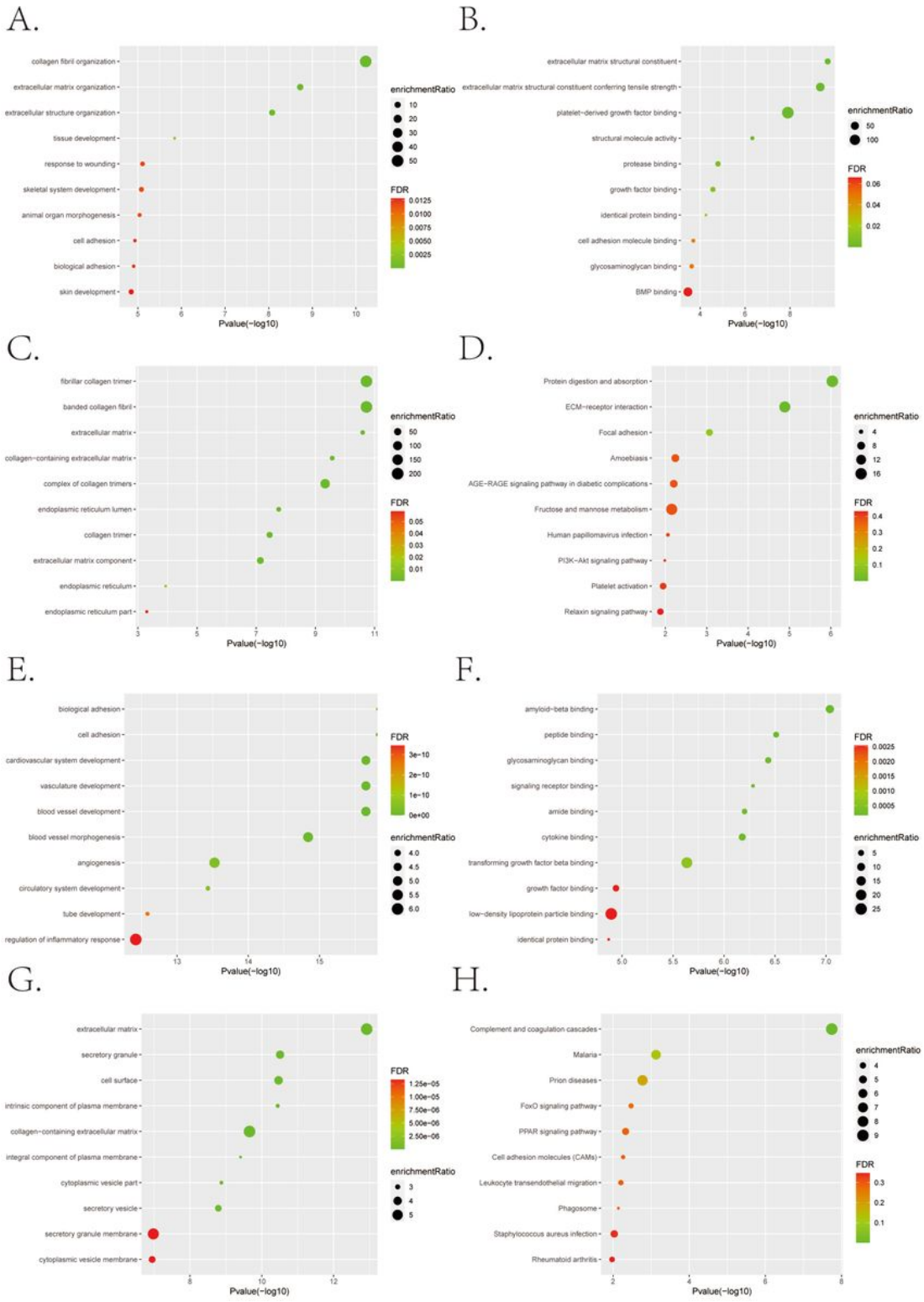

Figure 2

GO and KEGG analysis of the overlapping DEGs in LA. A. Biological process in up regulated DEGs; B. Molecular function in up regulated DEGs;C.Cellular component in up regulated DEGs; D. KEGG pathway in up regulated DEGs; E. Biological process in down regulated DEGs; F. Cellular component in down regulated DEGs; G. Molecular function in down regulated DEGs; H. KEGG pathway in down regulated DEGs.

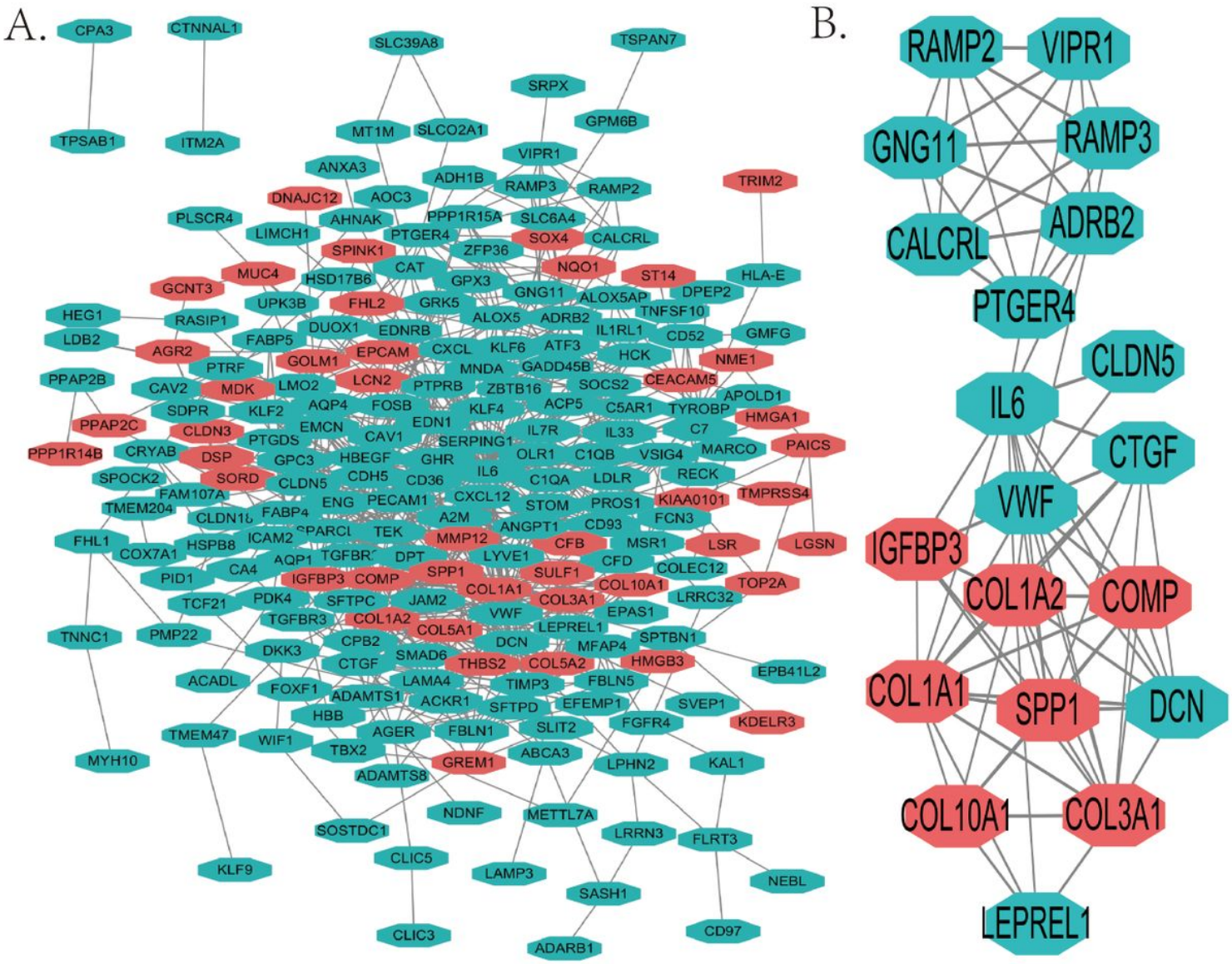

Figure 3

A.PPI network constructed via STRING and Cytoscape including 218 nodes and 753 edges;B. the most important PPI network module was obtained using MCODE in Cytoscape, consisted of 20 nodes and 76 edges (Red color represents up regaluted genes and green color represents down regaluted genes).

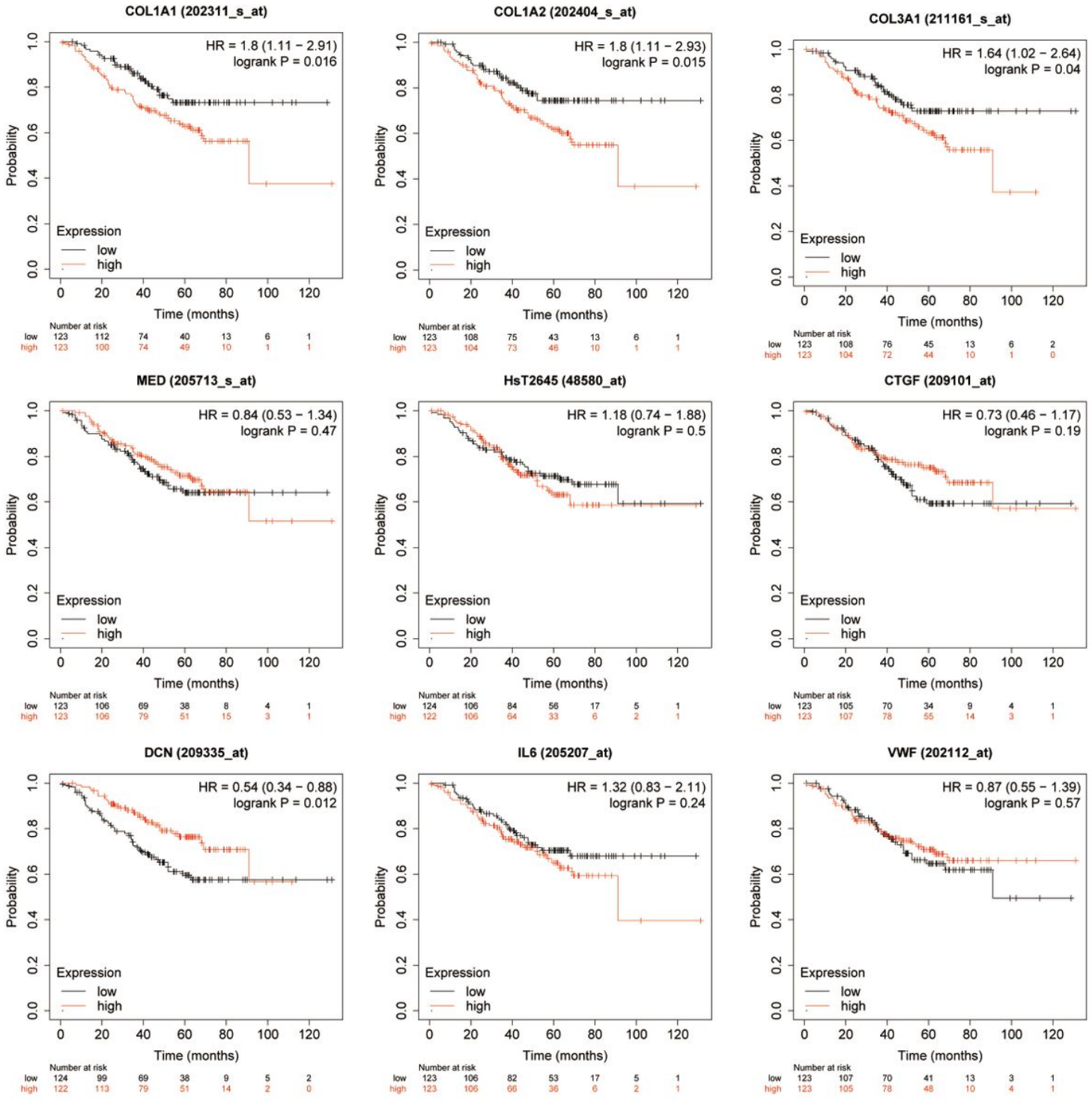

Figure 4

The survival analysis of hub genes in LA somkers via Kaplan Meier plotter (MED=COMP, HsT2645=SPP1).

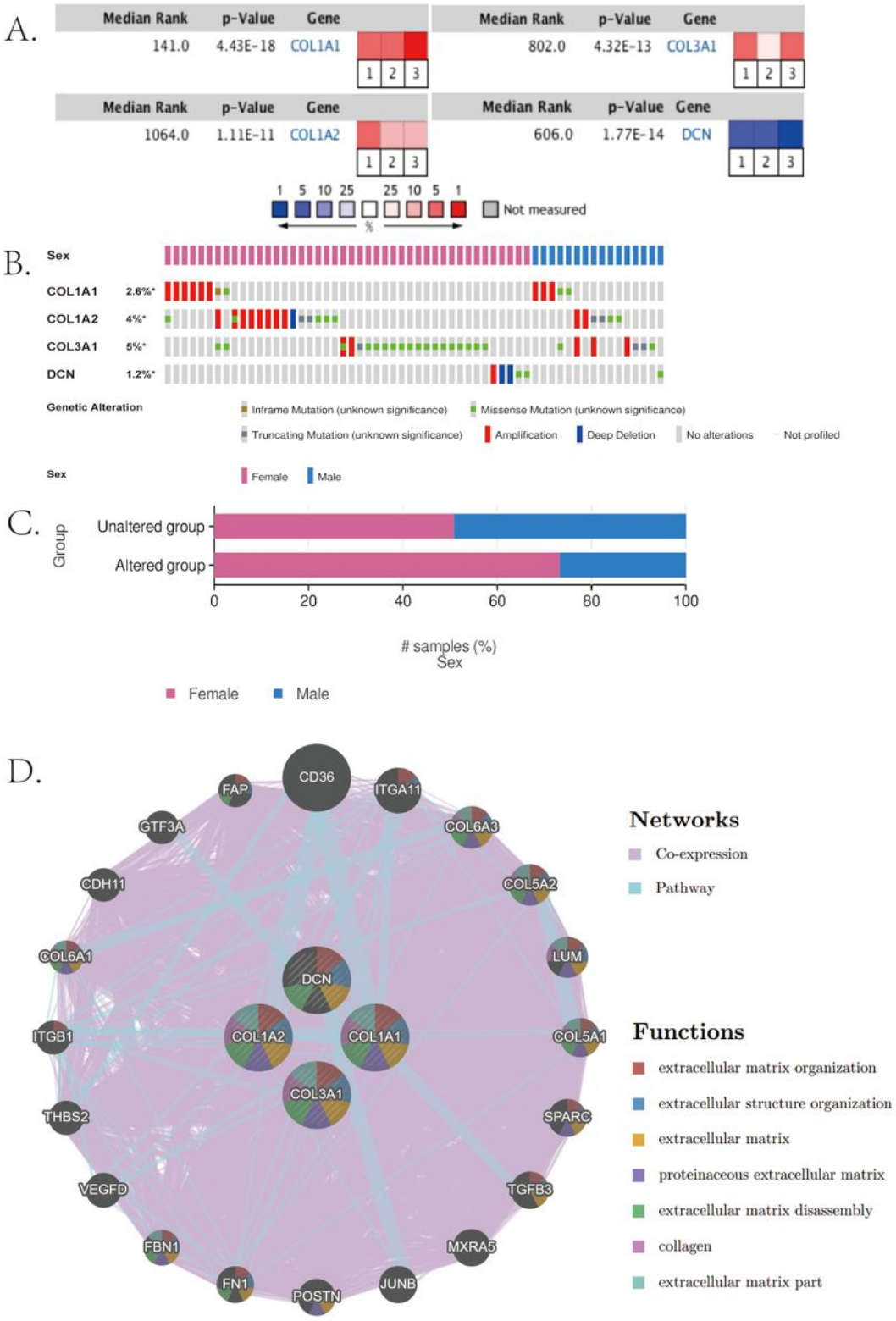

Figure 5

A. Heat maps of key gene expression in clinical LAT vs normal TAT in smokers via 3 studies in Oncomine database: 1. Lung Adenocarcinoma vs. Normal Landi Lung, PloS ONE, 2008; 2. Lung Adenocarcinoma vs. Normal Okayama Lung, Cancer Res, 2012; 3. Lung Adenocarcinoma vs. Normal Selamat Lung, Genome Res, 2012. B. Genetic alteration of key genes in 508 LA patients who smoke via cBioPortal. C. The relationship between altered genes and sex in genetic alteration analysis. D. The co-expression and

pathway networks of key genes via GENEMANIA (The area of genes/proteins represents synthesis score in co-expression and pathway. The thicker the line, the smaller the FDR value of the correlation).

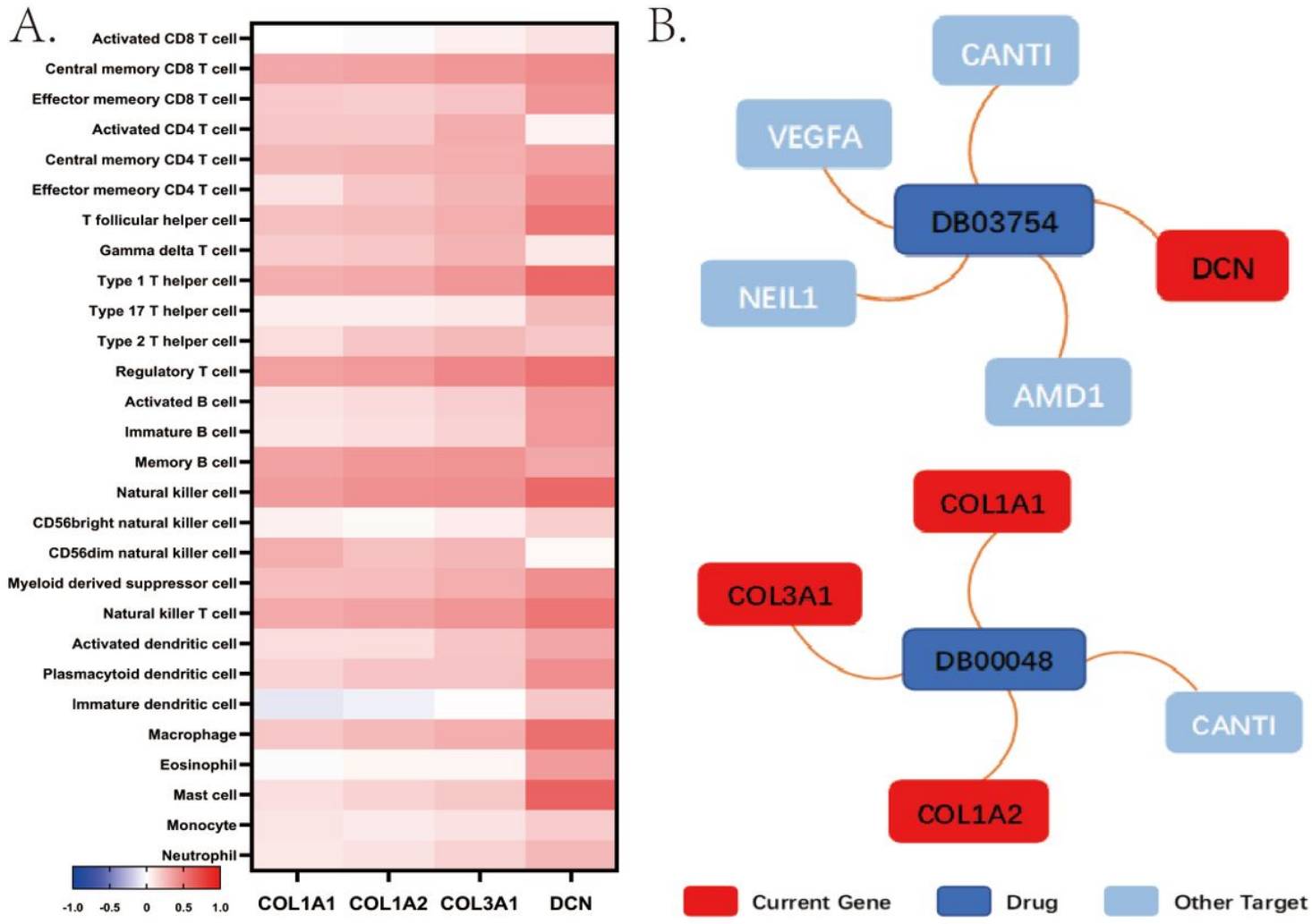

Figure 6

A. The correlation of 4 key genes (COL1A1, COL1A2, COL3A1 and DCN) and 28 types of lymphocytes in TISIDB with Spearman correlation test. B. Drug targets analysis of 4 key genes in TISIDB based on DrugBank database.

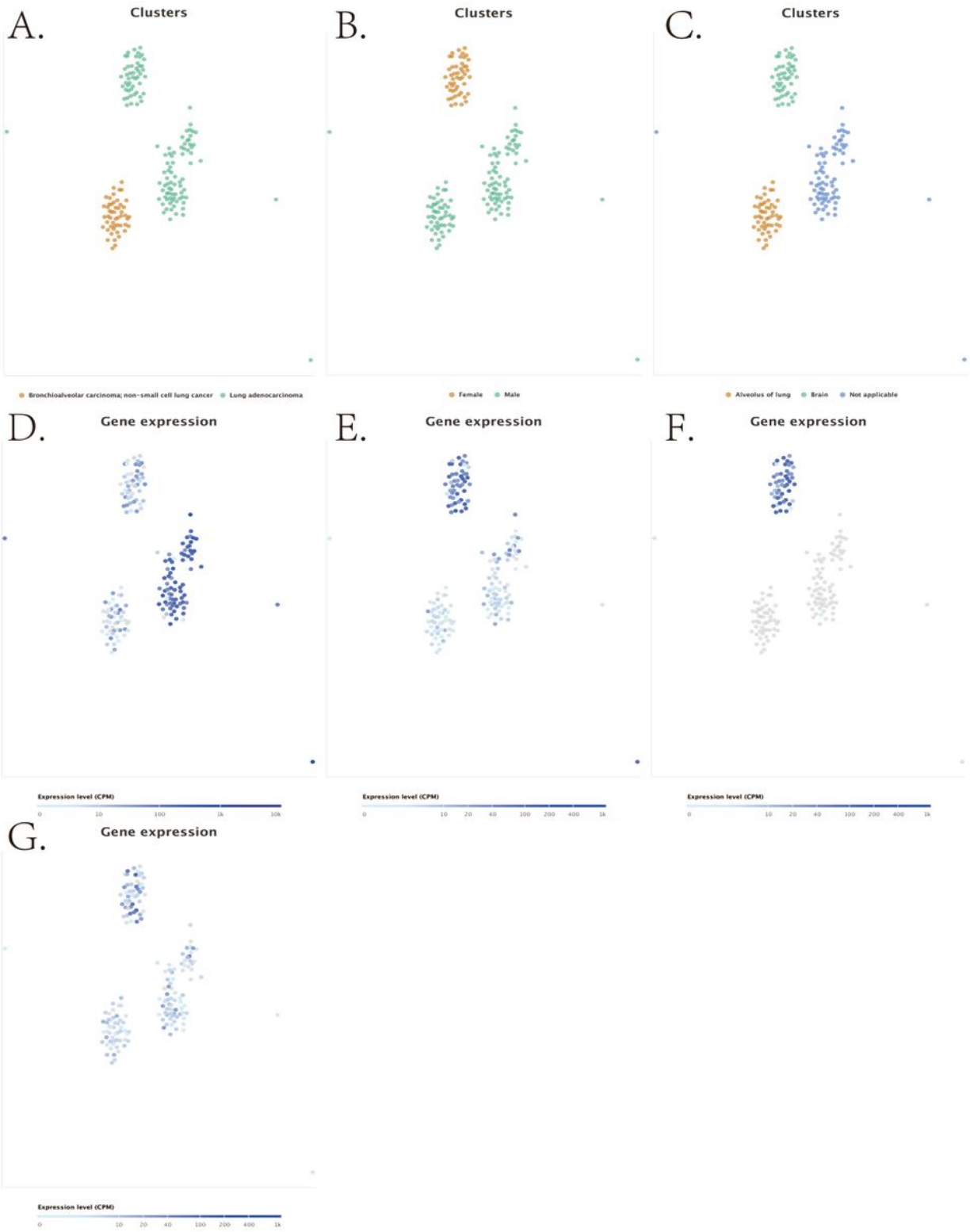

Figure 7

Expression distribution of 4 key genes in LA cells via single cell RNA sequencing of 176 lung adenocarcinoma patient-derived cells in EMBL-EBL (Each point represents a cell). A. The distribution of sample cells in diseases (Non-small cell lung cancer vs Lung adenocarcinoma). B. The distribution of sample cells in sex (Female vs Male)C. the distribution of sample cells in metastatic sites (Alveolus of lung vs Brain vs Not applicable). D. Expression distribution of COL1A1 gene in sample cells. E. Expression

distribution of COL1A2 gene in sample cells. F. Expression distribution of COL3A1 gene in sample cells. G. Expression distribution of DCN gene in sample cells.