MEDICAL SCIENCE MONITOR

MOLECULAR BIOLOGY

e-ISSN 1643-3750
© Med Sci Monit, 2019; 25: 4285-4304
DOI: 10.12659/MSM.916828

Received: 2019.04.08
Accepted: 2019.05.19
Published: 2019.06.09

Weighted Gene Coexpression Network Analysis Identifies Key Genes and Pathways Associated with Idiopathic Pulmonary Fibrosis

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

BCDEF Zheng Wang*
BCEF Jie Zhu*
ADG Fengzhe Chen
ADG Lixian Ma

Department of Infectious Diseases, Qilu Hospital, Shandong University, Jinan, Shandong, P.R. China

* Zheng Wang and Jie Zhu are co-first authors

Corresponding Authors: Lixian Ma, e-mail: mlx_sdu@163.com, Fengzhe Chen, e-mail: 18560082133@163.com
Source of support: Departmental sources

Background: Idiopathic pulmonary fibrosis (IPF) is a life-threatening disease with an unknown etiology. Gene expression microarray data have provided some insights into the molecular mechanisms of IPF. This study aimed to identify key genes and significant signaling pathways involved in IPF using bioinformatics analysis.

Material/Methods: Differentially expressed genes (DEGs) were identified using integrated analysis of gene expression data with a robust rank aggregation (RRA) method. The Connectivity Map (CMAP) was used to identify gene-expression signatures associated with IPF. Weighted gene coexpression network analysis (WGCNA) was used to explore the functional modules involved in the pathogenesis of IPF.

Results: A total of 191 patients with IPF and 101 normal controls from six genome-wide expression datasets were included. CMAP predicted several small molecular agents as potential gene targets in IPF. Several functional modules were detected that showed the highest correlation with IPF, including an extracellular matrix (ECM) component, and a myeloid leukocyte migration and activation component involved in the immune response. Hub genes were identified in the key functional modules that might have a role in the progression of IPF.

Conclusions: WGCNA was used to identify functional modules and hub genes involved in the pathogenesis of IPF.

MeSH Keywords: Biomarkers, Pharmacological • Gene Regulatory Networks • Idiopathic Pulmonary Fibrosis

Full-text PDF: https://www.medscimonit.com/abstract/index/idArt/916828

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

Background

Idiopathic pulmonary fibrosis (IPF) is a devastating illness characterized by irreversible lung fibrosis [1]. Although the overall prevalence of IPF is not high, the incidence of the disease has recently increased. In Europe and North America, the annual incidence of IPF is estimated to be between 2.8 and 18 cases per 100,000 individuals [2]. The median age of IPF is about 65 years, and men have a higher incidence [3,4]. Wound healing results in fibrosis and is believed to be the basis for the pathogenesis of IPF and includes the stages of homeostasis, inflammation, cell migration, cell proliferation, and extracellular matrix (ECM) remodeling. IPF may be due to chronic injury of the alveolar epithelium that results in pulmonary fibrosis and structural lung remodeling [5]. However, the etiology of IPF remains unknown.

Currently, IPF is considered to be a result of the interaction between genetic and environmental risk factors [6], and aging might influence the susceptibility to lung fibrosis, as the incidence of IPF increases with age [7]. Also, genome-wide association studies have shown that some genes related to host defense and epithelial barrier function may also be involved in the pathogenesis of IPF [8]. Among these genes, a variant of the MUC5B promoter region was shown to be involved in the development of IPF [9]. In terms of environmental factors, cigarette smoking has been proposed to be the most common association with IPF [10]. Other inhaled environmental agents include exposure to metal dust or wood dust, sand, and spores from soil [4,11]. IPF remains a challenging disease to treat, and further studies are needed to improve the understanding of the underlying molecular mechanisms to identify gene targets for the development of novel therapies.

There are a large number of microarray gene expression datasets that are publicly available from the Gene Expression Omnibus (GEO) database. There is an increasing demand to integrate gene expression datasets to obtain more accurate results [12]. There have been no previous studies that have undertaken comprehensive bioinformatics analysis in IPF. Therefore, this study aimed to perform a comprehensive analysis approach by using a robust rank aggregation (RRA) method to identify the differentially expressed genes (DEGs) from several datasets [13]. The connectivity map (CMAP) (*https://portals.broadinstitute.org/cmap/*) transcriptional expression database was chosen to identify gene-expression signatures associated with IPF, as CMAP represents a valuable tool for establishing the connections between genes, drugs, and diseases and contains over 7000 gene expression profiles reflecting 1309 bioactive compounds [14]. CMAP can be used to identify mechanisms of action of small molecules and identify novel therapeutic targets [15]. The signature of differentially expressed genes (DEGs) can be used to input into the CMAP. Also, weighted gene coexpression network analysis (WGCNA) is a powerful approach for exploring the complex relationships between gene expression profiles and phenotype [16]. Therefore, in the present study, WGCNA was used to build a gene coexpression network and screen important modules in the network, and to filter the hub genes in the essential modules.

Therefore, this study aimed to identify key genes and significant signaling pathways involved in IPF using bioinformatics analysis. DEGs from lung tissue microarray data were identified from patients with IPF and normal controls, and potential gene targets for the treatment of IPF were detected using the CMAP database. WGCNA was used to construct a coexpression network associated with IPF to identify significant modules and hub genes, that may be related to the pathogenesis of IPF.

Material and Methods

Datasets used

The gene expression datasets from idiopathic pulmonary fibrosis (IPF) were downloaded from the Gene Expression Omnibus (GEO) repository (*http://www.ncbi.nlm.nih.gov/geo/*). The GEO represents the largest resource of public microarray data and is widely used to identify key genes in disease. In this study, there were several selection criteria for data selection that included: (a) gene expression datasets which contained gene chip microarrays; (b) studies comparing gene expressions between patients with IPF and normal control lung tissue; (c) sample size in each chip dataset contained at least ten samples; (d) raw data or processed data were available in these databases; (e) hypersensitive pneumonitis, cryptogenic organizing pneumonia (COP), or respiratory bronchiolitis-interstitial lung disease (RBILD) were not included in this study. Searches were excluded if they did not meet the inclusion criteria.

All patients with IPF included in this study met the diagnostic criteria for IPF based on the American Thoracic Society (ATS) and European Respiratory Society (ERS) consensus statement. Lung tissue samples from the normal control groups were obtained from patients without IPF and included lung cancer patients and lung transplant patients. Ethical approval for this study was not required because the data were downloaded directly from public datasets.

Integrated analysis of gene expression datasets

Microarray data preprocessing were executed using R Project software and Bioconductor packages [17]. The raw microarray data were converted to expression data using the robust multi-array average (RMA) algorithm based on R language [18]. Also, two preprocessed series matrix files were directly downloaded

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

Table 1. Results of CMap analysis.

cmap name	Mean	n	p
Pregnenolone	–0.524	4	0.00022
Lycorine	–0.425	5	0.00555
Chloropyrazine	–0.347	4	0.00573
Securinine	–0.661	4	0.01215
Rotenone	–0.477	4	0.02043
Indoprofen	–0.326	4	0.02449
Megestrol	–0.348	4	0.03463
Terazosin	–0.328	4	0.03766

* The compounds tested in at least four experiments were ranked according to p value.

from the GEO repository. Probes were mapped to gene symbols. The average gene expression levels were calculated for genes represented by more than one probe. The differentially expressed genes (DEGs) between patients with IPF and normal controls in each dataset were screened by the Limma package in R [19]. Integrated analysis for the DEGs obtained from the six eligible datasets was performed using the R package (Robust Rank Aggregation), based on a robust rank aggregation (RRA) method. Also, |logFC| ≥1 and P<0.05 were used as the cutoff criteria for screening the DEGs.

Screening of small molecules

To identify candidate small molecules, we compared the DEGs with the 1,309 different compounds in the connectivity map (CMAP) used to identify gene-expression signatures associated with IPF. The DEGs were input to CMAP software for analysis, according to the website instructions. Small molecules with negative connectivity scores, representing the compounds that might reverse the input signature, were identified as therapeutic agents. The small molecules with average scores <–0.3 and P<0.05 were selected (Table 1).

Weighted gene coexpression network analysis (WGCNA)

The WGCNA package was used to construct the coexpression network. GSE32537 which including 119 patients with IPF and 50 normal controls was analyzed by WGCNA. We calculated the coefficient of variation for each gene. The top 5,000 genes with the highest coefficient of variation were selected for further analysis by constructing a weighted gene coexpression network. The appropriate power value was determined when the independence degree was up to 0.9. The merge cut height of 0.3 and the minimal module size of 100 were explored to identify coexpression modules. The modules were randomly color-labeled. Genes that could not be clustered into any given module were assigned to the grey module. The coexpression networks were visualized by Cytoscape software version: 3.6.1 (*http://www.cytoscape.org/*) [20]. We identified modules that were significantly associated with IPF by calculated the correlation between module eigengenes and clinical features. Finally, the modules that were highly correlated with clinical function were chosen for further analysis.

Identification and validation of hub genes in key modules

Module membership (MM) represented the correlation of gene expression profile with the module eigengene. The top 30 genes with the highest connectivity in key modules were regarded as hub genes and were visualized with Cytoscape. The expression statuses of hub genes in key modules were identified for validation. Venn diagram was plotted using the online JVenn tool to overlap the hub genes in significant modules and DEGs obtained from the RRA analysis [21].

Functional enrichment analysis

Genes in interest modules were uploaded to the online bioinformatics database Metascape (*http://metascape.org/*) and underwent Gene Ontology (GO) biological process, cellular component, and molecular function analysis as well as Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis [22]. The number of enriched genes ≥3 and P<0.01 were regarded as the cutoff criteria. Only the top ten enriched pathways were extracted.

Results

Microarray datasets for idiopathic pulmonary fibrosis (IPF)

After selection according to the eligibility criteria, six datasets were incorporated into the study: GSE10667 [23], GSE15197 [24], GSE21369 [25], GSE24206 [26], GSE32537 [27], GSE110147 [28] (Table 2). The main characteristics of these publicly available databases including gender ratio, and age distribution, origin, number of samples, GSE number, platforms, and source types are listed in Table 2. The number of patients with IPF ranged from 8 to 119, and the number of normal controls ranged from 6 to 50. In total, 191 patients with IPF and 101 normal controls were enrolled in the study.

Robust rank aggregation (RRA) analysis

By using the RRA approach to integrate six microarray datasets, a total of 368 differentially expressed genes (DEGs) comprising 248 upregulated and 120 downregulated genes were obtained (Supplementary Table 1). The top 25 upregulated and

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

Table 2. Summary of those 6 genome-wide gene expression datasets involving IPF patients.

Dataset ID	GSE number	Samples	IPF age(years	IPF Sex (M/F)	Source types	Platform	Authors
1	GSE10667	23 IPF samples and 15 controls	61.71±5.51	19/4	Lung tissues	GPL4133	Konishi K, Richards TJ, Kaminski N
2	GSE15197	8 IPF samples and 13 controls	60±5	5/3	Lung tissues	GPL6480	Rajkumar R, Konishi K, Richards TJ et al.
3	GSE21369	11 IPF samples and 6 controls	58.09±9.48	8/3	Lung tissue	GPL570	Cho JH, Gelinas R, Wang K et al.
4	GSE24206	8 IPF samples and 6 controls	61.25±6.32	7/1	Lung tissue	GPL570	Meltzer EB, Barry WT, D'Amico TA et al.
5	GSE32537	119 IPF samples and 50 controls	62.55±8.75	77/42	Lung tissue	GPL6244	Yang IV, Coldren CD, Leach SM et al.
6	GSE110147	22 IPF samples and 11 controls	62±6	17/5	Lung tissue	GPL6244	Cecchini MJ, Hosein K, Howlett CJ et al.

downregulated genes are shown in Figure 1. Among these DEGs, the roles of some genes in IPF have been validated in previous studies, including KRT5, BPIFB1, and AGER.

Screening for small molecules

The connectivity map (CMAP) database was used to search for small molecules with therapeutic potential in IPF. The small molecules with a high negative connectivity score are presented in Table 1. Among these potential therapeutic agents, pregnenolone and lycorine showed a smaller P-value and were potential therapeutic targets for IPF.

Weighted gene coexpression network analysis (WGCNA)

According to the order of the coefficient of variation, the 5,000 most variable genes were chosen for WGCNA. Hierarchical clustering analysis was performed, and the results are shown in Supplementary Figure 1. There were three outlier samples in the 169 samples that included GSM806290, GSM806408, GSM806411 when the threshold was set as 55. Outlier samples were removed from the cohort before further analysis. As shown in Figure 2A, power=5 was chosen as the soft-thresholding to ensure a scale-free network. We set the minimum module size to 100 genes and the minimum cut height for merging of modules at 0.3. The weighted gene coexpression network analysis (WGCNA) identified eight modules ranging in size from 228 to 1808.

There were 45 genes that did not belong to any module were labeled in grey, which were excluded from further analysis. These coexpression modules were constructed and were presented in different colors (Figure 2B). Interactions between the eight coexpression modules were then analyzed (Figure 2C). To explore the clinical significance of the module, correlations between module eigengenes and clinic traits were analyzed. As shown in Figure 3, black, blue, magenta, and pink modules were positively correlated with two clinical traits, namely disease status and the St George's score for the severity of IPF. By contrast, yellow, brown, and red modules were found to be negatively associated with disease status and St George's score for severity of IPF.

Also, we found some positive correlations between the black module and smoking pack years. Combined with Figure 4, we observed that these seven modules yielded two main clusters; one included four modules (black, blue, magenta and pink module) while the other included three modules (yellow, brown and red). Furthermore, two pairs of modules had higher adjacencies, and they were the black and magenta module, yellow and brown module respectively. Also, the module eigengene of the black and yellow module showed a higher correlation with disease status (Figure 3). The black module had the strongest positive correlation with IPF (r=0.79; P=4e-37) while the yellow module had the strongest negative correlation with IPF (r=–0.81; P=3e-40). We plot a scatterplot for the correlation between module membership and gene significance in the two key modules, respectively (Figure 5). The coexpression networks were visualized with Cytoscape and are shown in Supplementary Figure 2.

Identification and validation of hub genes in the key modules

Hub genes may determine the characteristics of a module and play significant roles in biological processes. Consequently, the top 30 genes with the highest degree of connectivity in the black and yellow module were taken as hub genes, including COL14A1, TSHZ2, IL1R2, and SLCO4A1. Figure 6A and 6B showed the top 30 hub genes in the black and yellow module.

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

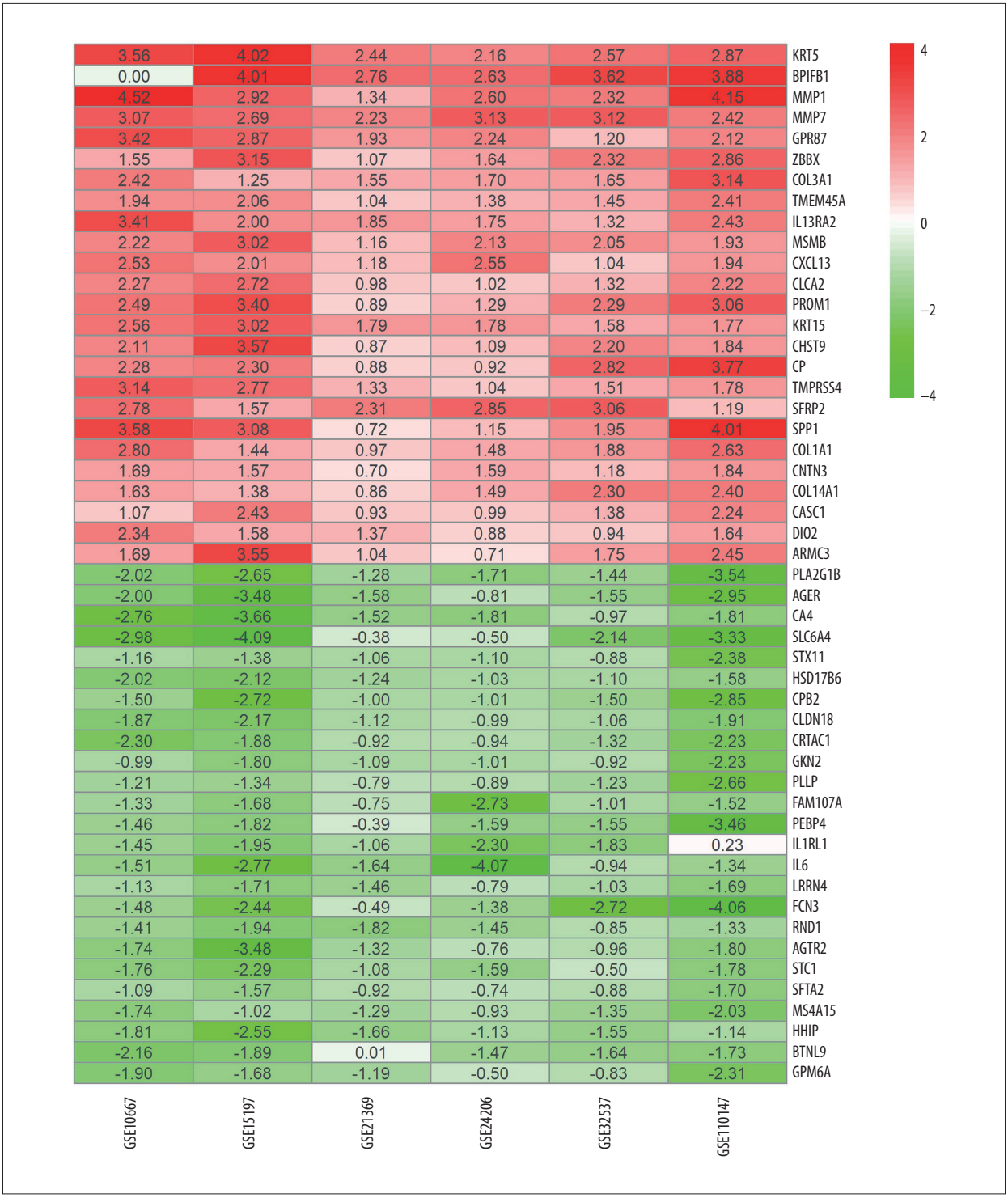


Figure 1. The top 25 upregulated genes and top 25 down-regulated genes in idiopathic pulmonary fibrosis (IPF). Each column represents one dataset, and each row represents one gene. The numbers in each rectangle show the logarithmic fold-change of genes in each dataset. Red indicates increased gene expression, and green indicates decreased gene expression.

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

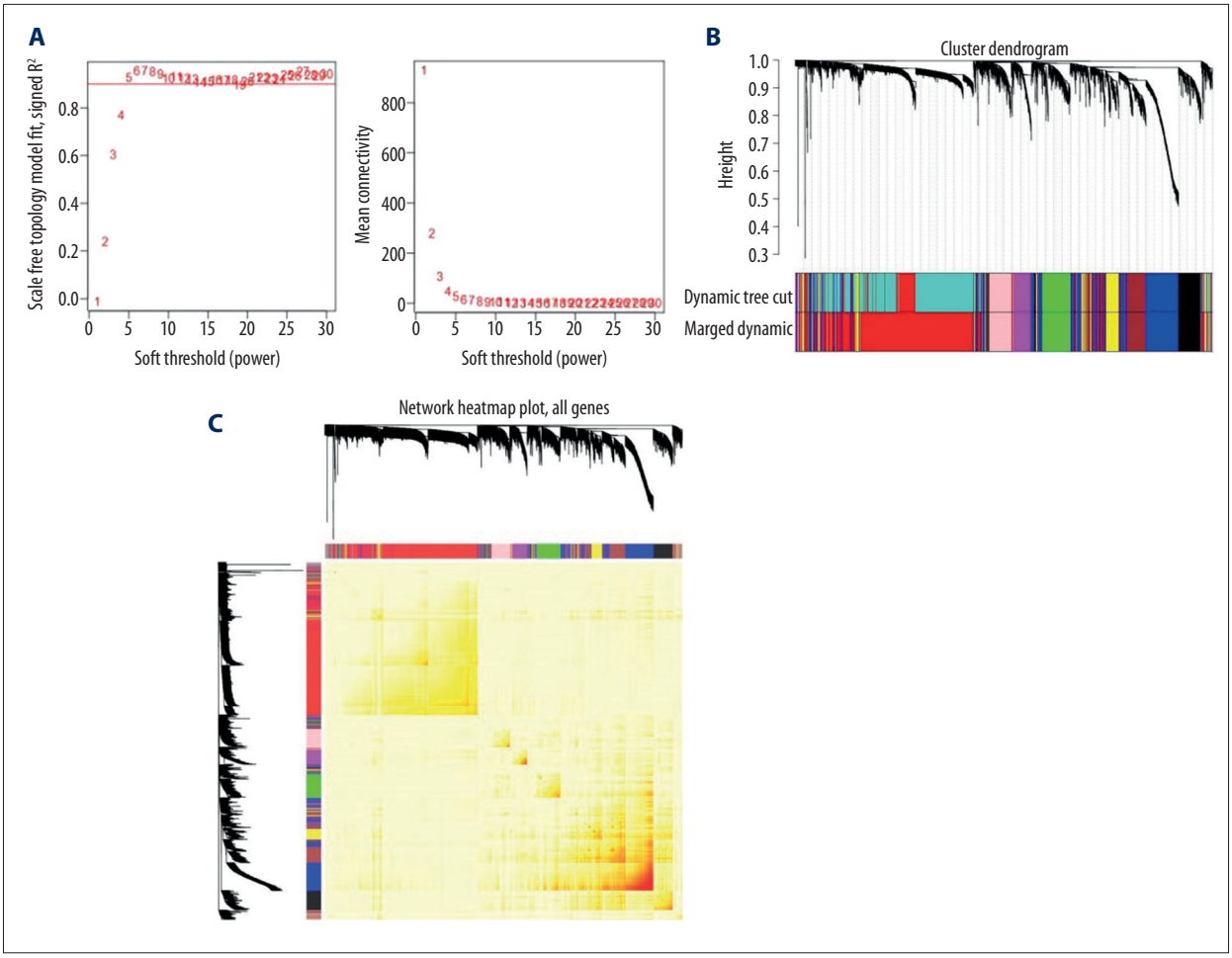


Figure 2. Plots in the weighted gene coexpression network analysis (WGCNA) using gene expression data from 119 patients with idiopathic pulmonary fibrosis (IPF) and 50 controls from GSE32537 datasets. (**A**) Network topology of different soft-thresholding powers. The left panel shows the influence of the soft-threshold power on the scale-free topology fit index. The right panel shows the influence of the soft-threshold power on the mean connectivity. (**B**) Cluster dendrogram of coexpression genes and functional modules in IPF. Eight coexpression modules were constructed and are shown in different colors. (**C**) A heatmap plot shows the gene network. Different colors of the horizontal axis and the vertical axis represent different modules. The light color indicates lower overlap, and the dark red indicates higher overlap.

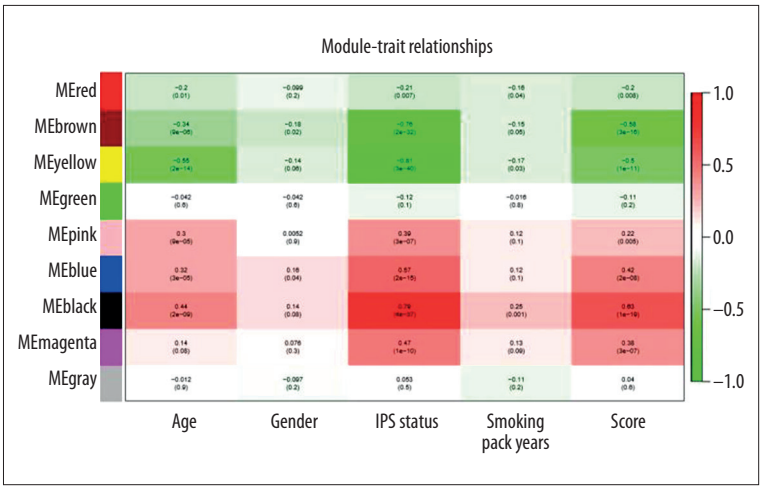


Figure 3. Heatmap of the correlation between module eigengenes and clinical traits of idiopathic pulmonary fibrosis (IPF). The table is color-coded by correlation according to the color legend. Each cell contains the corresponding correlation and p-value.

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

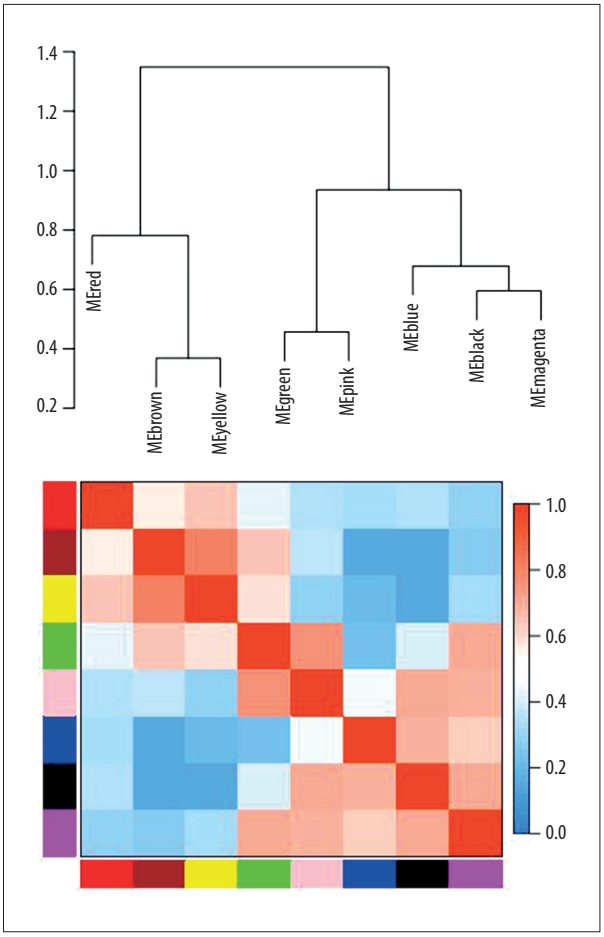


Figure 4. Heatmap plot of the adjacencies of modules. The change of color from blue (0) to red (1) in the heatmap represents the degree of connectivity of different modules from weak to strong.

We used DEGs generated from RRA analysis to validate the expression status of the top two hub genes in the black and yellow module. We overlapped the DEGs and genes in the black and yellow module by plotting a Venn diagram. These four hub genes were present in DEGs and significant modules, indicating their value as potential biomarkers for IPF. The results are presented in Figure 7.

Functional enrichment analysis

Functional enrichment analysis was performed for the genes in the constructed seven modules. The genes in the black module were mainly enriched in extracellular matrix (ECM) organization, skeletal development, and vasculature development (Figure 8A). Genes in the yellow module were enriched in myeloid leukocyte migration, leukocyte activation, and were involved in the immune response and the inflammatory response (Figure 8B). Genes in the brown module were mainly enriched in blood vessel development, cell junction organization, and sterol biosynthetic. Genes in the blue module were mainly involved in cilia, motile cilia, and O-glycan processing. Multiple signaling pathways were found to be involved in other modules, including nuclear division in the magenta module, T cell activation in the pink module, and olfactory transduction in the red module. The main results in Gene Ontology (GO)

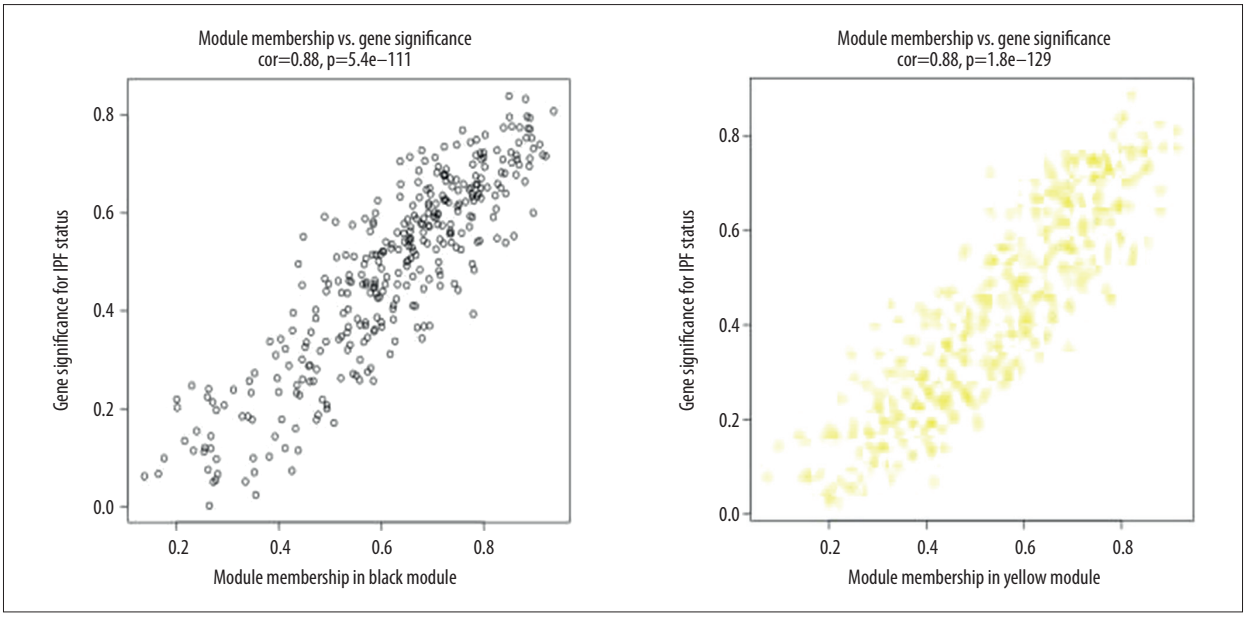


Figure 5. A scatterplot of gene significance for idiopathic pulmonary fibrosis (IPF) status versus module membership in the black module and yellow module. The correlation coefficient and the corresponding P-values are shown at the top of the scatterplot.

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

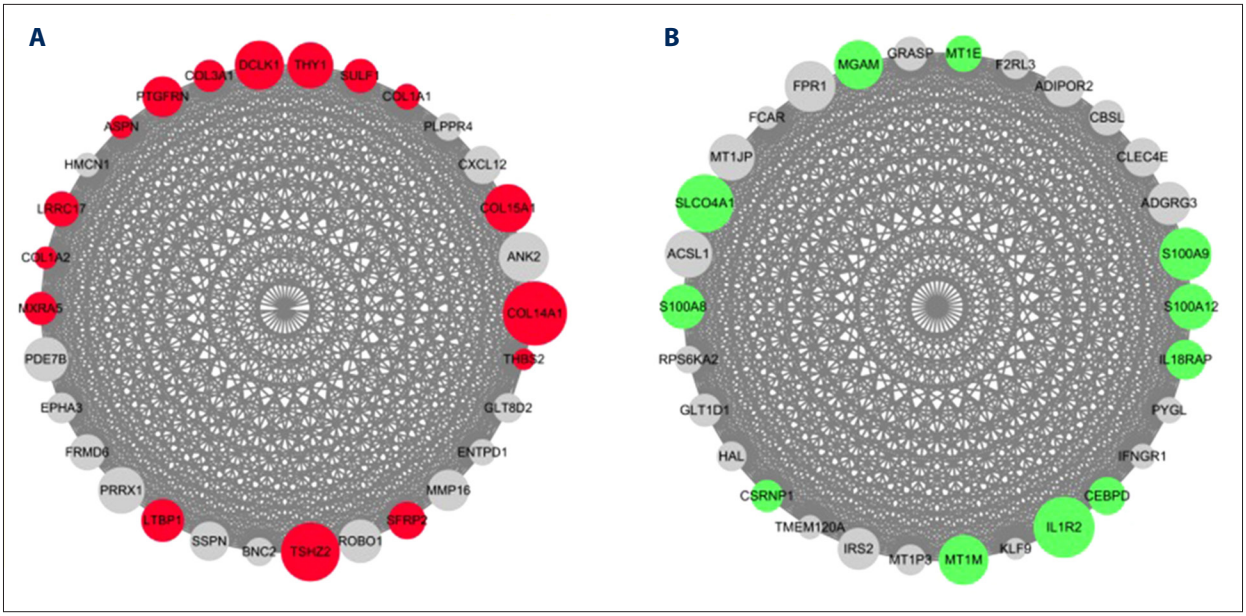


Figure 6. Network diagram of the top 30 genes in idiopathic pulmonary fibrosis (IPF). (**A**) The network of top 30 genes in the dark magenta module. (**B**) The network of top 30 genes in the yellow module. Node size: larger size indicates a higher degree of connectivity, and a smaller size indicates a lower degree of connectivity. Node color: Red indicates an upregulated gene. Green indicates a down-regulated gene.

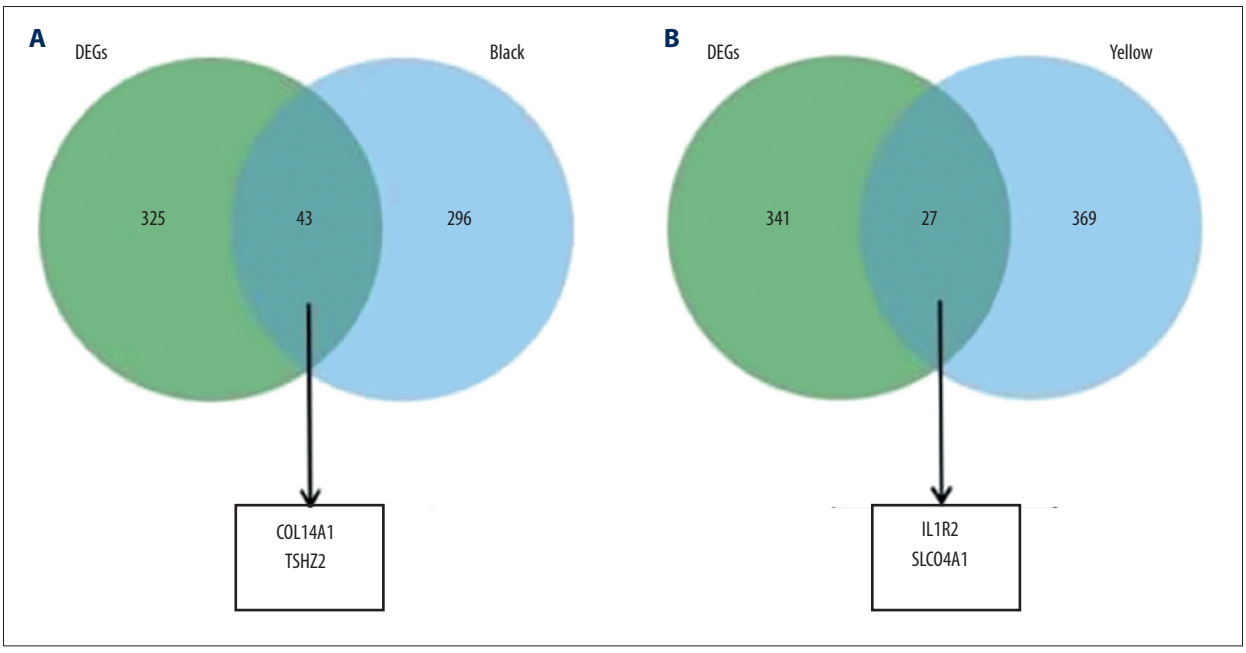


Figure 7. The overlap of differentially expressed genes (DEGs) and hub genes shown as a Venn diagram. (**A**) Identification of common genes between differentially expressed genes (DEGs) and the black module overlapping them. The two hub genes in the black module were also DEGs obtained from the robust rank aggregation (RRA) analysis. (**B**) Identification of common genes between DEGs and the yellow module overlapping them. The two hub genes in the yellow module were also DEGs obtained from the RRA analysis.

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

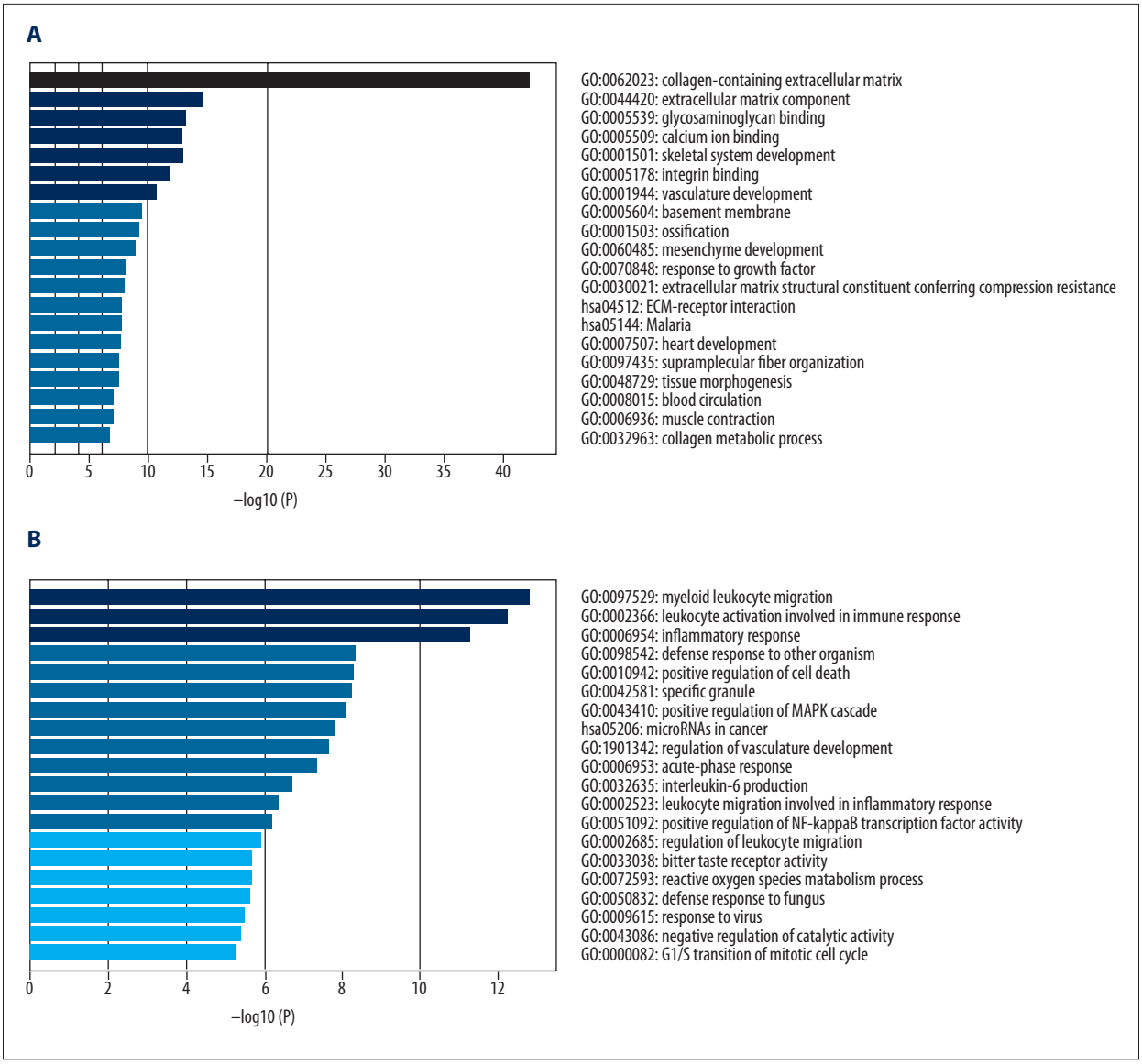


Figure 8. Functional and pathway enrichment analysis. (**A**) Functional and pathway enrichment analysis of the black module. (**B**) Functional and pathway enrichment analysis of the yellow module. Heatmap of top 20 clusters, colored by P-values. Each bar represents a cluster. The darker the color of the bar is, the smaller the P-value.

biological process, cellular component, and molecular function analysis and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of these seven modules are shown in Supplementary Table 2.

Discussion

Idiopathic pulmonary fibrosis (IPF) is a progressive and devastating disease that usually leads to death within five years after diagnosis [29]. The pathogenesis of the disease remains poorly understood. In the current study, we conducted an integrated analysis of gene expression data to explore the molecular pathogenesis of IPF. In this study, the robust rank aggregation (RRA) method, was used to screen significant differentially expressed genes (DEGs) from six independent gene expression datasets. A large number of DEGs were identified from which 248 DEGs were upregulated and 120 DEGs were down-regulated in IPF patient samples compared with control samples.

IPF is the result of chronic inflammation and healing in the lung. This study identified some DEGs that were closely associated with inflammation and healing. It was previously reported that KRT5 was highly expressed in the alveolar regions of the lung in IPF [30]. Also, airway stem cells have been shown to express KRT5 in lung injury [31]. The findings from

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

these previous studies support that KRT5 is associated with the process of healing in the alveolar epithelium. BPIFB1 has also been previously shown to be upregulated in respiratory disease, including in chronic obstructive pulmonary disease (COPD) where it has been associated with disease severity [32]. Also, an ulcer-associated cell lineage plays an important role in the healing process in inflammatory bowel disease, and BPIFB1 is one of the ulcer-associated cell lineage genes [33]. Therefore, it is possible that BPIFB1 plays a significant role in the healing process in IPF.

PLA2G1B, a secreted phospholipase, was among the most downregulated genes in IPF in the present study. There have been few previous studies on PLA2G1B in IPF, and so the aberrant expression of this gene should be validated by future studies. AGER, or advanced glycosylation end product-specific receptor, is one of the DEGs that were significantly down-regulated in the present study. Advanced glycosylation end product (AGE) has been previously shown to inhibit the wound-healing in diabetes by reducing the activity of epidermal stem cells [34]. However, blocking the expression of AGER may facilitate the healing process, which may explain the possible role for AEGR in IPF.

Some candidate molecules with therapeutic potential in IPF were explored based on the use of the connectivity map (CMAP) database to identify gene expression signatures. Pregnenolone is an inactive precursor of steroid hormones, and its potential functional effects have been previously studied [35]. Pregnenolone therapy has shown beneficial effects in schizophrenia [36], bipolar depression [37], and drug dependence [38]. Prospective studies are needed to investigate pregnenolone as a candidate therapeutic compound in IPF. Lycorine is an alkaloid derived from *Hymenocallis littoralis*, which has antibacterial, antiviral, and wound-healing properties [39]. The findings from a previous study suggested that lycorine might have therapeutic potential in patients with IPF [40]. This previous finding was consistent with the findings from the present study, and lycorine is a molecule that requires further study.

The method of WGCNA that was the basis of this study is an efficient approach to construct coexpressed modules and hub genes in several diseases [41]. In this study, we investigated the gene expression profile of GSE32537, including 119 patients with IPF and 50 normal controls to explore the molecular mechanism of IPF. Using WGCNA, the weighted coexpression network was constructed, and eight coexpression modules were identified. Seven of the modules showed a significant correlation with disease status, which were divided into two clusters, providing new insights into the pathogenesis of IPF. Among the identified modules, the black, blue, magenta, and pink modules were positively associated with disease status, whereas the yellow, brown, and red modules were negatively correlated with disease status. To determine the importance of these coexpression modules in the pathogenesis of IPF, enrichment analysis was performed using the Metascape online bioinformatics gene annotation and analysis database. The black module was found to be mainly enriched in collagen-containing extracellular matrix (ECM) and the ECM component. Increased accumulation of ECM is an important phase of the wound-healing process. ECM has been identified as an active contributor to the progression of fibrosis [42]. A therapeutic approach that targets the ECM may reverse fibrosis and restore lung function. The results of this study showed that the yellow module was mainly enriched in myeloid leukocyte migration and leukocyte activation involved in the immune response, and in cell migration. These results were consistent with previous studies that immune signaling pathways may be protective in IPF [43].

The black and yellow modules were found to be most significantly correlated with the disease status of IPF. Therefore, we screened the hub genes and validated the top two hub genes in the two modules. These four hub genes were all present in DEGs and significant modules, indicating their important roles in biological processes. COL14A1 is a member of the fibril-associated collagens with interrupted triple helices (FACIT) collagen family, with the main function of collagen binding and ECM components [44]. COL14A1 expression has been reported to be significantly upregulated in the pulmonary artery intima and media of patients with pulmonary arterial hypertension [45]. It is likely that COL14A1 plays a key role in the process of pulmonary interstitial fibrosis.

In this study, TSHZ2, a zinc-finger homeobox gene, was identified as a hub gene. Studies on the role of TSHZ2 are limited, and future studies are required to confirm its function in IPF. IL1R2 is a member of the interleukin-1 receptor family and acts as a negative regulator of the IL-1 system. The anti-inflammatory effect of IL1R2 has been confirmed by *in vivo* studies, and include chronic skin inflammation [46], and arthritis [47]. The combination of acute inflammation and healing due to persistent injury results in fibrosis [48]. In the present study, IL1R2 was both a hub gene and a significantly down-regulated gene in the IPF group, suggesting that the overactive wound healing was most likely associated with the functional deletion of IL1R2. SLCO4A1 is a member of solute carrier organic anion transporter family, which is involved in the transport of a variety of compounds [49]. There have been few studies on SLCO4A. However, SLCO4A1 has been shown to have a role in the proliferation of colorectal cancer cell and cell migration *in vitro* and to be a negative prognostic marker *in vivo* [50]. Further studies are required to determine the role of SLCO4A1 in IPF.

The key genes identified in this study were associated with the pathogenesis of fibrosis, including in IPF, and the findings

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

were supported by those from previously published studies. However, this study had several limitations. Only six datasets of microarray data for IPF were used. Few clinical and regional characteristics of the patients were available in each dataset, and there may have been relevant clinical differences that affected the findings. Also, the results obtained by bioinformatics analysis alone should be verified by *in vitro* and clinical studies, and so further studies are needed to validate these preliminary findings.

Conclusions

Through the comprehensive bioinformatics analysis of several gene expression datasets, we identified differentially expressed genes (DEGs) that might be related to the pathogenesis, diagnosis, and prognosis of idiopathic pulmonary fibrosis (IPF). Also, relevant small molecules were predicted using the connectivity map (CMAP) database to identify gene expression signatures associated with IPF, which might be potential candidate therapeutic compounds for IPF. We used weighted gene coexpression network analysis (WGCNA) to build a coexpression network associated with IPF and found two modules and four hub genes, which might identify novel insights into the molecular mechanisms underlying IPF. The findings of this study should be confirmed by future *in vitro* and clinical studies.

Conflict of interest

None.

Supplementary Files

Supplementary Table 1. Genes statistically differentially expressed between IPF and normal lung tissue.

Name	Pvalue	adjPvalue	logFC
KRT5	8.16E-19	2.28E-14	2.935163698
BPIFB1	1.66E-17	4.63E-13	2.816100513
MMP1	1.71E-14	4.77E-10	2.975972266
MMP7	1.34E-13	3.74E-09	2.778138388
GPR87	7.23E-13	2.02E-08	2.297678786
ZBBX	2.63E-12	7.32E-08	2.096953319
COL3A1	3.10E-12	8.63E-08	1.951786774
TMEM45A	3.76E-12	1.05E-07	1.713292282
IL13RA2	5.28E-12	1.47E-07	2.127651882
MSMB	5.52E-12	1.54E-07	2.086340891
CXCL13	5.89E-12	1.64E-07	1.873998487
CLCA2	7.28E-12	2.03E-07	1.753165017
PROM1	9.11E-12	2.54E-07	2.236427095
KRT15	9.44E-12	2.63E-07	2.084821277
CHST9	2.38E-11	6.65E-07	1.946570653
CP	3.27E-11	9.12E-07	2.161658822
TMPRSS4	5.71E-11	1.59E-06	1.929477651
SFRP2	1.05E-10	2.93E-06	2.29394319
SPP1	1.97E-10	5.49E-06	2.416271247
COL1A1	2.32E-10	6.48E-06	1.865800024
CNTN3	2.83E-10	7.88E-06	1.42949464
COL14A1	3.89E-10	1.09E-05	1.676170261

Name	Pvalue	adjPvalue	logFC
CASC1	5.95E-10	1.66E-05	1.506540304
DIO2	6.13E-10	1.71E-05	1.458989173
ARMC3	6.62E-10	1.85E-05	1.864747862
SULF1	8.52E-10	2.38E-05	1.407676411
SERPINB3	8.76E-10	2.44E-05	1.822592533
KLHL13	8.85E-10	2.47E-05	1.415638864
CXCL14	9.30E-10	2.59E-05	1.831177737
ASPN	9.62E-10	2.68E-05	1.840973138
SERPINB5	9.62E-10	2.68E-05	1.120680021
C20orf85	9.97E-10	2.78E-05	1.801814831
SPAG17	1.04E-09	2.91E-05	1.646375125
LRRC17	1.08E-09	3.02E-05	1.396901134
SLC27A2	1.14E-09	3.19E-05	1.556826086
POSTN	1.17E-09	3.27E-05	1.85302783
SNTN	1.42E-09	3.97E-05	1.875956956
KRT17	1.46E-09	4.07E-05	1.959920244
CDH3	1.68E-09	4.68E-05	1.930510747
FANK1	1.95E-09	5.43E-05	1.503991449
MMP13	2.09E-09	5.84E-05	1.757375715
CCDC146	2.34E-09	6.53E-05	1.492462573
RSPH1	3.52E-09	9.82E-05	1.724507872
FAM81B	3.76E-09	0.000104985	1.776467268

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

Name	Pvalue	adjPvalue	logFC
VTCN1	4.32E-09	0.000120533	1.683907215
COMP	5.67E-09	0.000158194	2.319702189
RGS22	5.86E-09	0.000163421	1.446072149
MUC16	6.35E-09	0.000177097	1.571239322
SPATA18	6.42E-09	0.000179133	1.394998539
TSHZ2	7.05E-09	0.000196678	1.118911148
CAPSL	7.48E-09	0.000208696	1.638856212
TTC25	7.63E-09	0.000212835	1.30957343
COL1A2	8.27E-09	0.000230588	1.502829774
SFRP4	9.37E-09	0.000261414	1.230284362
DCLK1	9.88E-09	0.000275575	1.272063729
CYP2F1	1.02E-08	0.000283711	1.223127058
WDR78	1.22E-08	0.000338888	1.267174919
PSD3	1.32E-08	0.000368616	1.140913494
WDR63	1.49E-08	0.000414831	1.401986713
LTBP1	1.58E-08	0.000439688	1.310736879
CFH	1.60E-08	0.000447383	1.387921751
COL6A3	1.67E-08	0.000466154	1.120968356
NELL2	2.06E-08	0.000573796	1.111579211
CCDC113	2.21E-08	0.000616382	1.272100841
C6	2.50E-08	0.000696175	1.739230141
ST6GALNAC1	3.03E-08	0.000843723	1.390920015
RPS4Y1	3.22E-08	0.000898467	1.966604231
SPAG6	3.51E-08	0.000979708	1.755246732
TTC29	3.73E-08	0.001040815	1.400611802
SCGB1A1	3.96E-08	0.001105576	1.472963895
MMP10	4.06E-08	0.001132187	1.850435932
SPATA17	4.39E-08	0.001225456	1.386846091
SLN	4.48E-08	0.001248606	1.539423446
S100A2	4.48E-08	0.001250099	2.029971633
DSC3	4.53E-08	0.00126211	1.633315606
TSPAN1	5.09E-08	0.001418504	1.501762841
AGBL2	5.14E-08	0.001433455	1.265354573
DYNLRB2	6.25E-08	0.001742309	1.426611109
LTF	6.70E-08	0.00186864	1.362389488
MYH11	6.96E-08	0.001941834	1.020881259

Name	Pvalue	adjPvalue	logFC
SLITRK6	7.19E-08	0.00200411	1.524418276
SLC4A11	7.32E-08	0.002041705	1.021480991
LRRC46	7.46E-08	0.002081847	1.39939705
LCN2	7.55E-08	0.00210533	1.519553185
MDH1B	8.81E-08	0.002456185	1.238493664
PTPRZ1	9.04E-08	0.00252096	1.027109023
THBS2	9.08E-08	0.002531893	1.217132935
KRT6A	9.20E-08	0.002564929	1.843710106
FAT2	9.48E-08	0.002643413	1.436657195
CCDC80	1.17E-07	0.003264358	1.287638765
COL15A1	1.23E-07	0.003431577	1.558457712
PRUNE2	1.24E-07	0.003458424	1.189978667
AQP5	1.52E-07	0.004250384	1.059448724
IGF1	1.56E-07	0.004358779	1.334563461
LGALS7B	1.58E-07	0.004406939	1.303091197
PTGFRN	1.67E-07	0.004670182	1.086813811
GPX8	1.73E-07	0.004814652	1.169574353
MNS1	2.05E-07	0.005724399	1.273932547
KRT14	2.08E-07	0.005811711	1.243073844
HSPA4L	2.14E-07	0.005980132	1.321865881
FNDC1	2.18E-07	0.006074683	1.48019782
GREM1	2.37E-07	0.006620838	1.465380281
C9orf135	2.75E-07	0.007658837	1.41427499
FBXO15	2.91E-07	0.008110255	1.068188235
CAPS	3.02E-07	0.008414294	1.197249034
THY1	3.02E-07	0.008436056	1.488425651
SYNPO2	3.22E-07	0.008985261	1.113568786
MUC5B	3.32E-07	0.009247558	1.197014038
CDH2	3.98E-07	0.011111537	1.260162496
STOML3	4.27E-07	0.011895241	1.499556832
KRT6C	4.60E-07	0.012826127	1.589275459
PLN	4.76E-07	0.013269168	1.064513896
SERPINF1	5.56E-07	0.01550185	1.162722912
DNAH12	5.83E-07	0.016251891	1.553317339
LRRN1	5.89E-07	0.016415964	1.328297804
TEKT1	6.01E-07	0.01675329	1.546529799

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

Name	Pvalue	adjPvalue	logFC
C11orf88	6.13E-07	0.017103725	1.484628508
COL5A2	6.29E-07	0.017538773	1.024345467
EYA2	6.29E-07	0.017543232	1.183016083
SERPINI2	6.29E-07	0.017543232	1.126653922
MEOX1	6.39E-07	0.017834057	1.228643451
CTHRC1	6.78E-07	0.018915173	1.529655688
MUC4	6.92E-07	0.019312725	1.084574554
MORN5	7.01E-07	0.01955195	1.494365253
TNFRSF17	7.30E-07	0.020371412	1.013826148
SCG5	7.31E-07	0.020381583	1.604114371
SERPIND1	8.50E-07	0.023696112	1.380147985
STK33	8.98E-07	0.025033914	1.255678765
ENKUR	9.00E-07	0.025094148	1.441136955
BCHE	9.00E-07	0.025094148	1.010033179
NEK10	9.52E-07	0.026552979	1.447262547
EFCAB1	9.74E-07	0.027176271	1.549417004
LRRIQ1	9.93E-07	0.027688035	1.502569257
ALDH3A1	1.03E-06	0.028723921	1.057312477
UBXN10	1.05E-06	0.029384698	1.17634531
GSTA1	1.08E-06	0.030063269	1.608489358
CCL18	1.09E-06	0.03047741	1.027242614
CTSE	1.25E-06	0.034820248	1.051893038
COL10A1	1.37E-06	0.038142045	1.30970437
CTSK	1.38E-06	0.038348612	1.094051078
TRIM29	1.46E-06	0.040680309	1.110754893
EFHC2	1.62E-06	0.04516481	1.129952453
GOLM1	1.64E-06	0.045682808	1.068976108
TEKT2	1.65E-06	0.045921326	1.358996882
C11orf70	1.77E-06	0.049452252	1.162790469
SOX2	1.89E-06	0.052746565	1.502511435
AMPD1	2.03E-06	0.056590851	1.005313806
CYP24A1	2.07E-06	0.057621316	1.284583408
CD24	2.21E-06	0.061555849	1.261804161
WDR66	2.23E-06	0.062226218	1.15117896
ANKFN1	2.28E-06	0.063660615	1.001663679
HS6ST2	2.46E-06	0.068471393	1.092390297

Name	Pvalue	adjPvalue	logFC
EIF1AY	2.47E-06	0.068819733	1.084622032
WDR49	2.80E-06	0.078202171	1.32678646
C9orf24	2.82E-06	0.078602712	1.508651167
DNAH9	2.95E-06	0.08233687	1.467446672
CXCL6	3.02E-06	0.084225814	1.319295673
CHST6	3.14E-06	0.087627154	1.358514328
PIP	3.23E-06	0.090079114	1.372153545
ABCA13	3.23E-06	0.090079114	1.326351733
C6orf118	3.37E-06	0.093849633	1.41969099
ARMC4	3.54E-06	0.09872715	1.222717911
RPGRIP1L	3.83E-06	0.10687075	1.099001594
KCNJ16	4.09E-06	0.113975886	1.083243034
FAP	4.18E-06	0.116579023	1.019870168
CCDC17	4.23E-06	0.117867727	1.156041743
PIH1D2	4.25E-06	0.118602386	1.005323245
C7orf57	4.56E-06	0.127210058	1.09585796
FAM83D	4.87E-06	0.13594777	1.237173638
EFHC1	4.93E-06	0.137528392	1.063614057
EFHB	5.01E-06	0.139621819	1.053752603
DNAI2	5.34E-06	0.148888408	1.179025366
TMEM190	5.44E-06	0.151608905	1.351028578
BPIFA1	5.49E-06	0.153040902	1.29842003
DYDC2	5.56E-06	0.154968104	1.290632913
SAA1	5.82E-06	0.162381965	1.167439591
SYT8	5.84E-06	0.162863691	1.054481629
C4orf22	5.89E-06	0.164313163	1.248512247
COL17A1	6.60E-06	0.184046213	1.676913349
BAAT	7.02E-06	0.19588472	1.125812636
PRSS12	7.08E-06	0.19749345	1.041357198
ACTG2	7.19E-06	0.200389326	1.098429643
FHOD3	7.24E-06	0.202037794	1.007947849
AKAP14	7.70E-06	0.214805786	1.155126263
TSGA10	8.25E-06	0.23012123	1.084478327
MXRA5	9.00E-06	0.251085283	1.072081569
CHIT1	9.20E-06	0.256665502	1.116001213
HHLA2	9.93E-06	0.276806417	1.243490407

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

Name	Pvalue	adjPvalue	logFC
VWA3B	1.09E-05	0.304184254	1.013600302
FAM216B	1.22E-05	0.339444516	1.003245966
ZMYND10	1.25E-05	0.349552078	1.06922834
ATP12A	1.26E-05	0.352113661	1.241793927
CDHR3	1.31E-05	0.366149094	1.003213287
C1orf87	1.33E-05	0.370438911	1.240688082
FCRL5	1.46E-05	0.407936307	1.074522645
IQUB	1.55E-05	0.432685114	1.273083757
DLEC1	1.56E-05	0.435688123	1.127617749
TP63	1.79E-05	0.498074147	1.022168715
DNER	1.87E-05	0.521842029	1.006616734
TDO2	1.92E-05	0.536740731	1.017985205
CCDC81	1.96E-05	0.547864087	1.111704424
DTHD1	2.21E-05	0.616533406	1.220070607
CCNA1	2.22E-05	0.618380293	1.109596078
DNAH6	2.50E-05	0.696039249	1.205496793
CCDC39	2.73E-05	0.762305285	1.064688193
SIX1	2.80E-05	0.781215901	1.166844918
TMEM232	2.92E-05	0.815239855	1.27748821
DSG3	2.98E-05	0.830395002	1.004793382
UGT1A6	3.08E-05	0.857680858	1.080278111
DNAH3	3.37E-05	0.940459718	1.020961433
AK7	3.47E-05	0.967454427	1.025283586
CLIC6	3.98E-05	1	1.050256203
SERPINB4	4.48E-05	1	1.286082732
SPATA4	5.09E-05	1	1.054065177
DNAH7	5.29E-05	1	1.07821031
DNAH5	5.93E-05	1	1.025356953
CCDC78	6.17E-05	1	1.001758647
CAPN13	6.26E-05	1	1.075410711
DNAI1	7.23E-05	1	1.169928958
RIBC2	7.97E-05	1	1.006702152
KLK12	8.41E-05	1	1.242575651
CRLF1	8.88E-05	1	1.150088178
SRD5A2	9.56E-05	1	1.075894376
CLDN8	9.74E-05	1	1.30318712

Name	Pvalue	adjPvalue	logFC
C10orf107	0.000109055	1	1.060098777
STOX1	0.000109344	1	1.109895823
C10orf81	0.000118127	1	1.119302483
DDX3Y	0.000140361	1	1.049553425
FAM183A	0.000153129	1	1.040469983
CILP	0.000162266	1	1.173032082
RSPH4A	0.000198839	1	1.091928915
USP9Y	0.000203297	1	1.007151508
EYA1	0.000317943	1	1.10128004
MS4A8	0.00032339	1	1.005164192
PTPRT	0.000360859	1	1.169300297
GSTA5	0.000373902	1	1.000889231
FHAD1	0.000393343	1	1.047998957
IGFL2	0.000465987	1	1.312853186
RHOV	0.000475109	1	1.033831468
LDLRAD1	0.00055986	1	1.029189153
PLA2G2A	0.000583188	1	1.096318802
DNAAF1	0.000908779	1	1.062363628
FAM92B	0.001034847	1	1.00072388
GSTA2	0.001066812	1	1.005385962
SPATS1	0.001099321	1	1.031573291
MMP11	0.00110875	1	1.016851831
CWH43	0.001450876	1	1.036477759
TNS4	0.001582539	1	1.171200138
PLA2G1B	1.44E-14	4.00E-10	−2.106784443
AGER	8.34E-14	2.33E-09	−2.062573006
CA4	4.34E-13	1.21E-08	−2.088124716
SLC6A4	3.08E-12	8.60E-08	−2.237860051
STX11	6.55E-12	1.83E-07	−1.325682912
HSD17B6	8.40E-12	2.34E-07	−1.514855943
CPB2	1.25E-11	3.47E-07	−1.76408969
CLDN18	1.65E-11	4.59E-07	−1.520279583
CRTAC1	3.07E-11	8.55E-07	−1.598044156
GKN2	4.35E-11	1.21E-06	−1.339491525
PLLP	5.22E-11	1.46E-06	−1.354622717
FAM107A	5.54E-11	1.55E-06	−1.504206175

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

Name	Pvalue	adjPvalue	logFC
PEBP4	6.23E-11	1.74E-06	−1.712436843
IL1RL1	7.30E-11	2.04E-06	−1.393300726
IL6	1.78E-10	4.97E-06	−2.045946662
LRRN4	2.55E-10	7.11E-06	−1.300700327
FCN3	2.76E-10	7.69E-06	−2.097486722
RND1	2.79E-10	7.79E-06	−1.466163154
AGTR2	4.24E-10	1.18E-05	−1.676540353
STC1	5.83E-10	1.63E-05	−1.497051803
SFTA2	5.83E-10	1.63E-05	−1.149823392
MS4A15	6.77E-10	1.89E-05	−1.392910619
HHIP	8.55E-10	2.39E-05	−1.641952122
BTNL9	9.30E-10	2.59E-05	−1.479401795
GPM6A	1.25E-09	3.48E-05	−1.401440693
SLC26A9	1.31E-09	3.64E-05	−1.213384003
HYAL1	1.34E-09	3.74E-05	−1.049135629
NECAB1	1.65E-09	4.61E-05	−1.653062787
SPRY4	2.34E-09	6.53E-05	−1.063958117
RTKN2	3.17E-09	8.84E-05	−1.844810146
CACNA2D2	3.17E-09	8.84E-05	−1.325524239
HBEGF	3.37E-09	9.39E-05	−1.013418905
CSF3R	3.44E-09	9.61E-05	−1.248564827
ANKRD1	3.84E-09	0.000107106	−1.41285721
VIPR1	4.21E-09	0.000117331	−1.559310652
TMEM100	4.69E-09	0.000130713	−1.694707846
CA2	5.96E-09	0.000166087	−1.028516653
RXFP1	6.38E-09	0.000177931	−1.311761392
MT1M	6.77E-09	0.000188683	−1.747649733
SLC6A14	7.90E-09	0.000220316	−1.089465157
CXCL2	1.19E-08	0.000330772	−1.004769994
ADRB1	1.35E-08	0.000375261	−1.139606331
PIGA	1.35E-08	0.000376122	−1.217675912
LRRC32	1.78E-08	0.000497564	−1.128069969
TNNC1	2.31E-08	0.000643874	−1.27225718
ANXA3	2.37E-08	0.00066157	−1.058524507
ZNF385B	3.25E-08	0.000907556	−1.229033975
PROK2	3.61E-08	0.00100569	−1.514684378

Name	Pvalue	adjPvalue	logFC
CHI3L2	4.18E-08	0.001166833	−1.427095228
SLC19A3	4.57E-08	0.001274104	−1.10459671
S100A12	4.62E-08	0.00128834	−2.252364042
GPIHBP1	5.50E-08	0.001534937	−1.187373977
LAMP3	5.76E-08	0.001606384	−1.159421533
DAPK2	7.35E-08	0.002049604	−1.02316703
ACADL	7.61E-08	0.002122618	−1.119836013
MME	7.67E-08	0.002138363	−1.092020465
STXBP6	8.62E-08	0.002403444	−1.092919132
NDRG4	9.12E-08	0.002542679	−1.016978682
EPB41L5	1.00E-07	0.002788995	−1.009261675
CSF3	1.02E-07	0.002831121	−1.251734223
CDH13	1.33E-07	0.003708031	−1.007819606
GRIA1	1.34E-07	0.003740217	−1.377236278
S1PR1	1.39E-07	0.0038712	−1.029547489
MGAM	1.41E-07	0.003940929	−1.466668873
ABCA3	1.46E-07	0.004074518	−1.199031193
EDNRB	1.84E-07	0.005127183	−1.128719906
MAFF	2.20E-07	0.006122403	−1.097428593
SLC39A8	2.56E-07	0.007141459	−1.216623582
KLF4	3.11E-07	0.008664018	−1.041025187
SDR16C5	3.82E-07	0.010656914	−1.275861012
CCK	4.33E-07	0.012065446	−1.159781479
ITLN2	4.63E-07	0.012905847	−1.722536404
IL1R2	4.63E-07	0.012905847	−1.225516832
SERTM1	4.85E-07	0.013532256	−1.235269195
MATN3	5.39E-07	0.015037563	−1.013765474
SLCO4A1	6.21E-07	0.017322509	−1.218439326
FGG	6.48E-07	0.018079474	−1.260330186
HECW2	7.80E-07	0.021754525	−1.113597001
APOH	8.86E-07	0.024696133	−1.081456999
SUSD2	8.96E-07	0.02498895	−1.008091337
NAMPT	1.15E-06	0.032162761	−1.047173312
SLC46A2	1.73E-06	0.048361585	−1.026945366
SOSTDC1	1.84E-06	0.051255152	−1.228238809
SLC5A9	2.34E-06	0.065287501	−1.228911687

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

Name	Pvalue	adjPvalue	logFC
CCDC85A	2.39E-06	0.066612171	–1.067734881
IL18RAP	2.44E-06	0.068162213	–1.202066709
ERRFI1	2.88E-06	0.08029286	–1.09242787
ORM1	3.26E-06	0.090825371	–1.085120865
CXCL3	3.26E-06	0.091010933	–1.001088364
SPOCK2	3.43E-06	0.095778504	–1.023342385
S100A3	3.49E-06	0.097452831	–1.061209678
CEBPD	3.87E-06	0.107980901	–1.093847526
S100A8	4.39E-06	0.122472366	–1.339030199
SLCO1A2	5.14E-06	0.143355387	–1.086207968
ESM1	6.05E-06	0.16871941	–1.078977525
CSRNP1	6.49E-06	0.180900922	–1.136678398
FGFBP2	8.96E-06	0.249934931	–1.148678049
BDNF	8.98E-06	0.250476093	–1.076553749
ZFP36	1.01E-05	0.281124207	–1.027441113
DLL4	1.05E-05	0.292138522	–1.063256088
DEFA1B	1.16E-05	0.322399438	–1.011654806
VNN2	1.17E-05	0.325560059	–1.015560522

Name	Pvalue	adjPvalue	logFC
PTX3	1.34E-05	0.374984707	–1.338994441
F11	1.52E-05	0.423768653	–1.184438918
HTR3C	1.66E-05	0.464166403	–1.178516229
HMGCS2	1.70E-05	0.47451476	–1.14093025
ARC	2.03E-05	0.567374491	–1.211855264
ORM2	2.07E-05	0.577788023	–1.019830926
FPR2	2.11E-05	0.587693475	–1.029517016
GPX3	2.24E-05	0.623819328	–1.009177645
XIST	2.28E-05	0.635537729	–1.103246381
MT1E	2.36E-05	0.658399738	–1.006714478
NR4A2	2.54E-05	0.706999207	–1.027905066
RS1	2.64E-05	0.734963593	–1.020756087
FOSB	3.08E-05	0.857680858	–1.206609882
RBP2	3.38E-05	0.941426241	–1.007947503
KLRF1	3.44E-05	0.96012864	–1.045789521
IL13	4.17E-05	1	–1.023145887
S100A9	8.75E-05	1	–1.045992124
ADM	0.00020169	1	–1.022596555

Supplementary Table 2. Pathway and process enrichment analysis of those functional coexpression modules in IPF.

Modules	GO	Category	Description	Count	%	Log10(P)	Log10(q)
Black module	GO: 0062023	GO Cellular Components	Collagen-containing extracellular matrix	55	16.57	–42.21	–37.86
	GO: 0044420	GO Cellular Components	Extracellular matrix component	14	4.22	–14.54	–10.96
	GO: 0005539	GO Molecular Functions	Glycosaminoglycan binding	23	6.93	–13.09	–9.58
	GO: 0005509	GO Molecular Functions	Calcium ion binding	39	11.75	–12.87	–9.47
	GO: 0001501	GO Biological Processes	Skeletal system development	33	9.94	–12.8	–9.45
	GO: 0005178	GO Molecular Functions	Integrin binding	17	5.12	–11.75	–8.43
	GO: 0001944	GO Biological Processes	Vasculature development	38	11.45	–10.59	–7.37
	GO: 0005604	GO Cellular Components	Basement membrane	13	3.92	–9.38	–6.27
	GO: 0001503	GO Biological Processes	Ossification	24	7.23	–9.13	–6.1
	GO: 0060485	GO Biological Processes	Mesenchyme development	20	6.02	–8.9	–5.9

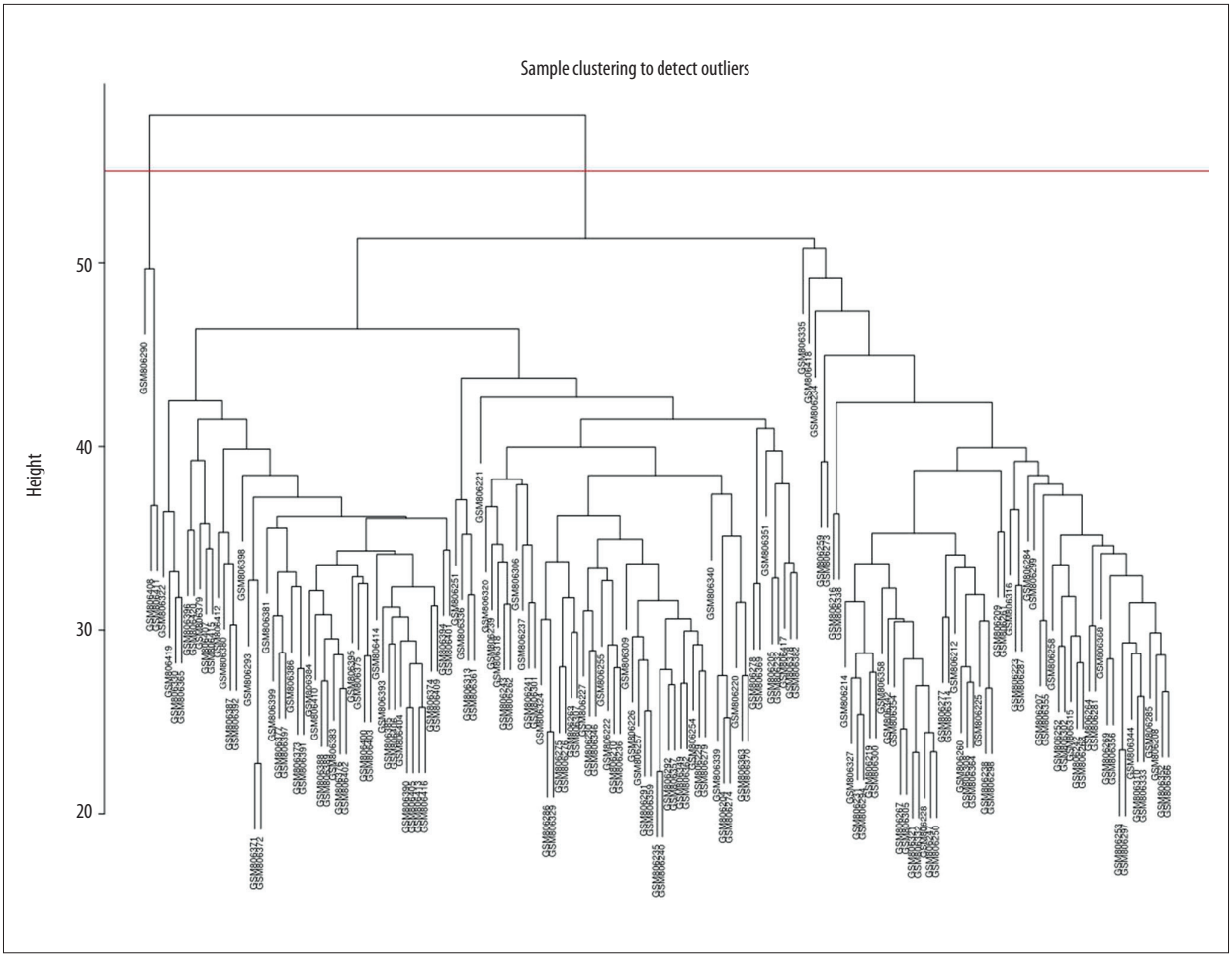
Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Modules	GO	Category	Description	Count	%	Log10(P)	Log10(q)
Yellow module	GO: 0097529	GO Biological Processes	Myeloid leukocyte migration	23	6.04	–12.85	–8.61
	GO: 0002366	GO Biological Processes	Leukocyte activation involved in immune response	41	10.76	–12.25	–8.42
	GO: 0006954	GO Biological Processes	Inflammatory response	42	11.02	–11.26	–7.86
	GO: 0098542	GO Biological Processes	Defense response to other organism	30	7.87	–8.34	–5.44
	GO: 0010942	GO Biological Processes	Positive regulation of cell death	34	8.92	–8.28	–5.41
	GO: 0042581	GO Cellular Components	Specific granule	16	4.2	–8.25	–5.41
	GO: 0043410	GO Biological Processes	Positive regulation of MAPK cascade	30	7.87	–8.09	–5.26
	hsa05206	KEGG Pathway	MicroRNAs in cancer	21	5.51	–7.84	–5.04
	GO: 1901342	GO Biological Processes	Regulation of vasculature development	25	6.56	–7.67	–4.89
	GO: 0006953	GO Biological Processes	Acute-phase response	9	2.36	–7.35	–4.62
Blue module	GO: 0005929	GO Cellular Components	Cilium	67	8.08	–16.89	–12.53
	GO: 0031514	GO Cellular Components	Motile cilium	26	3.14	–9.31	–5.55
	GO: 0016266	GO Biological Processes	O-glycan processing	11	1.33	–5.19	–1.98
	GO: 0045503	GO Molecular Functions	Dynein light chain binding	7	0.84	–4.81	–1.76
	GO: 0004497	GO Molecular Functions	Monooxygenase activity	13	1.57	–4.47	–1.47
	GO: 1904158	GO Biological Processes	Axonemal central apparatus assembly	3	0.36	–4.39	–1.43
	GO: 0023024	GO Molecular Functions	MHC class I protein complex binding	3	0.36	–4.39	–1.43
	GO: 0002223	GO Biological Processes	Stimulatory C-type lectin receptor signaling pathway	10	1.21	–4.15	–1.26
	GO: 0001889	GO Biological Processes	Liver development	15	1.81	–4.12	–1.24
	GO: 0045177	GO Cellular Components	Apical part of cell	28	3.38	–3.95	–1.14
Mengeta module	GO: 0000280	GO Biological Processes	Nuclear division	54	24	–45.01	–40.65
	GO: 0071103	GO Biological Processes	DNA conformation change	46	20.44	–42.92	–39.27
	GO: 0044770	GO Biological Processes	Cell cycle phase transition	47	20.89	–30.44	–27.28
	GO: 0006260	GO Biological Processes	DNA replication	36	16	–30.31	–27.19
	GO: 0005819	GO Cellular Components	Spindle	35	15.56	–25.8	–22.84
	GO: 1903046	GO Biological Processes	Meiotic cell cycle process	26	11.56	–22.27	–19.43
	GO: 0006281	GO Biological Processes	DNA repair	36	16	–19.92	–17.18
	GO: 0034508	GO Biological Processes	Centromere complex assembly	15	6.67	–17.6	–14.98
	GO: 0005815	GO Cellular Components	Microtubule organizing center	36	16	–15.54	–13.01
	GO: 0030496	GO Cellular Components	Midbody	18	8	–13.36	–10.96

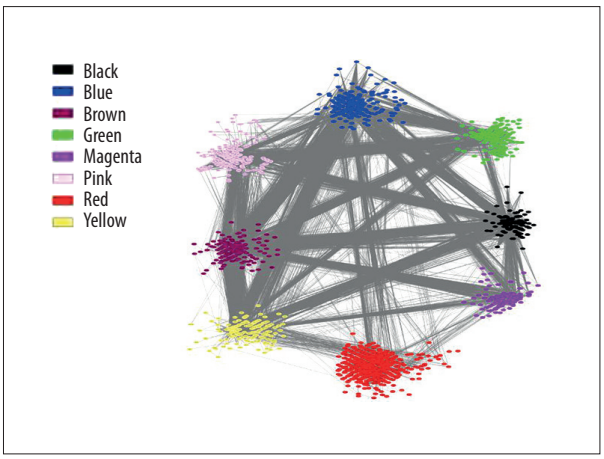
Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Modules	GO	Category	Description	Count	%	Log10(P)	Log10(q)
Pink module	GO: 0042110	GO Biological Processes	T cell activation	56	19.24	–39.21	–34.85
	GO: 0098552	GO Cellular Components	Side of membrane	49	16.84	–27.35	–23.77
	GO: 0050852	GO Biological Processes	T cell receptor signaling pathway	27	9.28	–23.27	–20.03
	GO: 0001816	GO Biological Processes	Cytokine production	50	17.18	–23.19	–19.98
	GO: 0045058	GO Biological Processes	T cell selection	16	5.5	–18.68	–15.81
	GO: 0019221	GO Biological Processes	Cytokine-mediated signaling pathway	43	14.78	–17.21	–14.39
	hsa04660	KEGG Pathway	T cell receptor signaling pathway	18	6.19	–15.43	–12.71
	GO: 0042101	GO Cellular Components	T cell receptor complex	10	3.44	–15	–12.29
	GO: 0046631	GO Biological Processes	alpha-beta T cell activation	19	6.53	–14.5	–11.84
	GO: 0031349	GO Biological Processes	Positive regulation of defense response	29	9.97	–13.11	–10.51
	GO: 0001568	GO Biological Processes	Blood vessel development	66	10.48	–17.01	–12.65
Brown module	GO: 0034330	GO Biological Processes	Cell junction organization	34	5.4	–12.64	–8.98
	GO: 0016126	GO Biological Processes	Sterol biosynthetic process	18	2.86	–11.61	–8.1
	GO: 0070848	GO Biological Processes	Response to growth factor	53	8.41	–10.72	–7.26
	GO: 0030155	GO Biological Processes	Regulation of cell adhesion	48	7.62	–9.5	–6.25
	GO: 0003018	GO Biological Processes	Vascular process in circulatory system	21	3.33	–8.83	–5.7
	GO: 0005911	GO Cellular Components	Cell-cell junction	34	5.4	–7.63	–4.72
	GO: 0070372	GO Biological Processes	regulation of ERK1 and ERK2 cascade	28	4.44	–7.38	–4.5
	GO: 0008285	GO Biological Processes	Negative regulation of cell proliferation	46	7.3	–6.92	–4.18
	GO: 0007610	GO Biological Processes	Behavior	38	6.03	–6.59	–3.93
	GO: 0001568	GO Biological Processes	Blood vessel development	66	10.48	–17.01	–12.65
Red module	hsa04740	KEGG Pathway	Olfactory transduction	108	6.76	–34.96	–30.6
	GO: 0031424	GO Biological Processes	Keratinization	58	3.63	–19.08	–15.72
	GO: 0030594	GO Molecular Functions	Neurotransmitter receptor activity	26	1.63	–7.43	–4.33
	GO: 0005179	GO Molecular Functions	Hormone activity	26	1.63	–7.12	–4.04
	GO: 0005261	GO Molecular Functions	Cation channel activity	48	3	–7.04	–3.98
	hsa04080	KEGG Pathway	Neuroactive ligand-receptor interaction	41	2.57	–5.92	–2.96
	GO: 0007188	GO Biological Processes	Adenylate cyclase-modulating G protein-coupled receptor signaling pathway	34	2.13	–5.38	–2.5
	GO: 0009566	GO Biological Processes	Fertilization	29	1.81	–5.22	–2.35
	GO: 0005549	GO Molecular Functions	Odorant binding	19	1.19	–4.83	–2.03
	GO: 0007210	GO Biological Processes	Serotonin receptor signaling pathway	11	0.69	–4.51	–1.82

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



Supplementary Figure 1. Sample clustering to detect outliers. The red line is used to distinguish the outlier samples. The threshold is set as 55, and three outlier samples (GSM806290, GSM806408, GSM806411) were removed from the sample cohort.



Supplementary Figure 2. The visualization of the coexpression networks. Nodes represent genes, and node color is the same as the module color node.

References:

1. Richeldi L, Collard HR, Jones MG: Idiopathic pulmonary fibrosis. Lancet, 2017; 389: 1941–52

2. Hutchinson J, Fogarty A, Hubbard R, McKeever T: Global incidence and mortality of idiopathic pulmonary fibrosis: A systematic review. Eur Resp J, 2015; 46: 795–806

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

3. Raghu G, Weycker D, Edelsberg J et al: Incidence and prevalence of idiopathic pulmonary fibrosis. Am J Respir Crit Care Med, 2006; 174: 810–16
4. Raghu G, Collard HR, Egan JJ et al: An official ATS/ERS/JRS/ALAT statement: Idiopathic pulmonary fibrosis: Evidence-based guidelines for diagnosis and management. Am J Respir Crit Care Med, 2011; 183: 788–824
5. Betensley A, Sharif R, Karamichos D: A systematic review of the role of dysfunctional wound healing in the pathogenesis and treatment of idiopathic pulmonary fibrosis. J Clin Med, 2016; 6(1): pii: E2
6. Kropski JA, Lawson WE, Young LR, Blackwell TS: Genetic studies provide clues on the pathogenesis of idiopathic pulmonary fibrosis. Dis Model Mech, 2013; 6: 9–17
7. Gulati S, Thannickal VJ: The aging lung and idiopathic pulmonary fibrosis. Am J Med Sci, 2019; 357: 384–89
8. Noth I, Zhang Y, Ma SF et al: Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: A genome-wide association study. Lancet Respir Med, 2013; 1: 309–17
9. Seibold MA, Wise AL, Speer MC et al: A common MUC5B promoter polymorphism and pulmonary fibrosis. N Engl J Med, 2011; 364: 1503–12
10. Baumgartner KB, Samet JM, Stidley CA et al: Cigarette smoking: A risk factor for idiopathic pulmonary fibrosis. Am J Respir Crit Care Med, 1997; 155: 242–48
11. Taskar VS, Coultas DB: Is idiopathic pulmonary fibrosis an environmental disease? Proc Am Thorac Soc, 2006; 3: 293–98
12. Rung J, Brazma A: Reuse of public genome-wide gene expression data. Nat Rev Genet, 2013; 14: 89–99
13. Kolde R, Laur S, Adler P, Vilo J: Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics, 2012; 28: 573–80
14. Lamb J, Crawford ED, Peck D et al: The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. Science, 2006; 313: 1929–35
15. Song Y, Liu J, Huang S, Zhang L: Analysis of differentially expressed genes in placental tissues of preeclampsia patients using microarray combined with the Connectivity Map database. Placenta, 2013; 34: 1190–95
16. Langfelder P, Horvath S: WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics, 2008; 9: 559
17. Gentleman RC, Carey VJ, Bates DM et al: Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol, 2004; 5(10): R80
18. Irizarry RA, Hobbs B, Collin F et al: Exploration, normalization, and summaries of high-density oligonucleotide array probe level data. Biostatistics, 2003; 4: 249–64
19. Smyth GK: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol, 2004; 3: Article3
20. Shannon P, Markiel A, Ozier O et al: Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res, 2003; 13: 2498–504
21. Bardou P, Mariette J, Escudié F et al: jvenn: An interactive Venn diagram viewer. BMC Bioinformatics, 2014; 15: 293
22. Tripathi S, Pohl MO, Zhou Y et al: Meta- and orthogonal integration of influenza "OMICs" data defines a role for UBR4 in virus budding. Cell Host Microbe, 2015; 18(6): 723–35
23. Konishi K, Gibson KF, Lindell KO et al: Gene expression profiles of acute exacerbations of idiopathic pulmonary fibrosis. Am J Respir Crit Care Med, 2009; 180: 167–75
24. Rajkumar R, Konishi K, Richards TJ et al: Genomewide RNA expression profiling in lung identifies distinct signatures in idiopathic pulmonary arterial hypertension and secondary pulmonary hypertension. Am J Physiol Heart Circ Physiol, 2010; 298(4): H1235–48
25. Cho JH, Gelinas R, Wang K et al: Systems biology of interstitial lung diseases: Integration of mRNA and microRNA expression changes. BMC Med Genomics, 2011; 4: 8
26. Meltzer EB, Barry WT, D'Amico TA et al: Bayesian probit regression model for the diagnosis of pulmonary fibrosis: Proof-of-principle. BMC Med Genomics, 2011; 4: 70
27. Yang IV, Coldren CD, Leach SM et al: Expression of cilium-associated genes defines novel molecular subtypes of idiopathic pulmonary fibrosis. Thorax, 2013; 68: 1114–21
28. Cecchini MJ, Hosein K, Howlett CJ et al: Comprehensive gene expression profiling identifies distinct and overlapping transcriptional profiles in non-specific interstitial pneumonia and idiopathic pulmonary fibrosis. Respiratory Res, 2018; 19: 153
29. Ley B, Collard HR, King TE Jr.: Clinical course and prediction of survival in idiopathic pulmonary fibrosis. Am J Respir Crit Care Med, 2011; 183: 431–40
30. Smirnova NF, Schamberger AC, Nayakanti S et al: Detection and quantification of epithelial progenitor cell populations in human healthy and IPF lungs. Respir Res, 2016; 17: 83
31. Zuo W, Zhang T, Wu DZ et al: p63(+)Krt5(+) distal airway stem cells are essential for lung regeneration. Nature, 2015; 517: 616–20
32. De Smet EG, Seys LJ, Verhamme FM et al: Association of innate defense proteins BPIFA1 and BPIFB1 with disease severity in COPD. Int J Chron Obstruct Pulmon Dis, 2018; 13: 11–27
33. Thorsvik S, van Beelen Granlund A, Svendsen TD et al: Ulcer-associated cell lineage expresses genes involved in regeneration and is hallmarked by high neutrophil gelatinase-associated lipocalin (NGAL) levels. J Pathol, 2019 [Epub ahead of print]
34. Zhu J, Wang P, Yu Z et al: Advanced glycosylation end product promotes forkhead box O1 and inhibits Wnt pathway to suppress capacities of epidermal stem cells. Am J Transl Res, 2016; 8: 5569–79
35. Vallee M: Neurosteroids and potential therapeutics: Focus on pregnenolone. J Steroid Biochem Mol Biol, 2016; 160: 78–87
36. Marx CE, Bradford DW, Hamer RM et al: Pregnenolone as a novel therapeutic candidate in schizophrenia: emerging preclinical and clinical evidence. Neuroscience, 2011; 191: 78–90
37. Brown ES, Park J, Marx CE et al: A randomized, double-blind, placebo-controlled trial of pregnenolone for bipolar depression. Neuropsychopharmacology, 2014; 39: 2867–73
38. Vallee M, Vitiello S, Bellocchio L et al: Pregnenolone can protect the brain from cannabis intoxication. Science, 2014; 343: 94–98
39. Subramaniam S, Sundarasekar J, Sahgal G, Murugaiyah V: Comparative analysis of lycorine in wild plant and callus culture samples of Hymenocallis littoralis by HPLC-UV method. ScientificWorldJournal, 2014; 2014: 408306
40. Karatzas E, Bourdakou MM, Kolios G, Spyrou GM: Drug repurposing in idiopathic pulmonary fibrosis filtered by a bioinformatics-derived composite score. Sci Rep, 2017; 7: 12569
41. Yan S, Wang W, Gao G et al: Key genes and functional coexpression modules involved in the pathogenesis of systemic lupus erythematosus. J Cell Physiol, 2018; 233(11) :8815–25
42. Haak AJ, Tan Q, Tschumperlin DJ: Matrix biomechanics and dynamics in pulmonary fibrosis. Matrix Biol, 2018; 73: 64–76
43. Huang Y, Ma SF, Espindola MS et al: Microbes are associated with host innate immune response in idiopathic pulmonary fibrosis. Am J Respir Crit Care Med, 2017; 196: 208–19
44. Bauer M, Dieterich W, Ehnis T, Schuppan D: Complete primary structure of human collagen type XIV (undulin). Biochim Biophys Acta, 1997; 1354: 183–88
45. Hoffmann J, Marsh LM, Pieper M et al: Compartment-specific expression of collagens and their processing enzymes in intrapulmonary arteries of IPAH patients. Am J Physiol Lung Cell Mol Physiol, 2015; 308: L1002–13
46. Rauschmayr T, Groves RW, Kupper TS: Keratinocyte expression of the type 2 interleukin 1 receptor mediates local and specific inhibition of interleukin 1-mediated inflammation. Proc Natl Acad Sci USA, 1997; 94: 5814–19
47. Bessis N, Guery L, Mantovani A et al: The type II decoy receptor of IL-1 inhibits murine collagen-induced arthritis. Eur J Immunol, 2000; 30: 867–75
48. Shah A, Amini-Nik S: The role of phytochemicals in the inflammatory phase of wound healing. Int J Mol Sci, 2017; 18(5): pii: E1068
49. Hagenbuch B, Meier PJ: Organic anion transporting polypeptides of the OATP/SLC21 family: Phylogenetic classification as OATP/SLCO superfamily, new nomenclature and molecular/functional properties. Pflugers Arch, 2004; 447(5): 653–65
50. Ban MJ, Ji SH, Lee CK et al: Solute carrier organic anion transporter family member 4A1 (SLCO4A1) as a prognosis marker of colorectal cancer. J Cancer Res Clin Oncol, 2017; 143: 1437–47

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