Insulin-like growth factor binding protein 7 as a candidate biomarker for systemic sclerosis

Y.-M. Yan[1], J.-N. Zheng[1], Y. Li[2,3], Q.-R. Yang[1], W.-Q. Shao[4], Q. Wang[1]

[1]Department of Dermatology, Zhongshan Hospital, Fudan University, Shanghai;
[2]Department of Stomatology, Zhongshan Hospital, Fudan University, Shanghai;
[3]State Key Laboratory of Molecular Engineering of Polymers, Fudan University, Shanghai;
[4]Department of Laboratory Medicine, Zhongshan Hospital, Fudan University, Shanghai, China.

Yue-Mei Yan, MD*
Ji-Na Zheng, MD*
Yang Li, MD*
Qiao-Rong Yang, MD
Wen-Qi Shao, MD
Qiang Wang, MD

*These authors contributed equally.

Please address correspondence to:
Qiang Wang,
Department of Dermatology,
Zhongshan Hospital,
Fudan University,
180 Fenglin Road,
Xuhui District,
Shanghai 200032, China
E-mail: wangqiang7766@163.com

Received on April 15, 2020; accepted in revised form on August 31, 2020.

Clin Exp Rheumatol 2021; 39 (Suppl. 131): S66-S76.

Key words: systemic sclerosis, weighted correlation network analysis, insulin-like growth factor binding protein 7, enzyme-linked immunosorbent assay, gene expression omnibus

Funding: this study was supported by grants awarded to Q. Wang from the National Natural Science Foundation of China (81641087), and the Research Fund of Shanghai Municipal Commission of Health and Family Planning (201640071).

Data availability statement: the datasets[GSE58095 (9), GSE32413 (10), GSE125362 (11), GSE76885 (12), GSE45485 (13), GSE95065 (14)] for this study can be found in the public GEO (http://www.ncbi.nlm.nih.gov/geo/).

Competing interests: none declared.

ABSTRACT

Objectives. *Systemic sclerosis (SSc) is an autoimmune disease clinically characterised by skin and internal organs fibrosis with high mortality. However, the pathogenesis of SSc is still controversial and the effect of the current treatment is far from satisfactory. We aimed to find out novel candidate genes related to the pathological process in SSc.*

Methods. *In this study, the weighted correlation network analysis (WGCNA) was conducted to identify the key module and hub genes most related to SSc in GSE58095, a microarray dataset from the Gene Expression Omnibus (GEO) database. Also, the key module was analysed by Gene Ontology (GO) analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis. Then we validated hub genes in other datasets (GSE32413, GSE125362, GSE45485, GSE76885, GSE95065). The serum of 37 patients with SSc and 25 healthy control subjects (HCs) were recruited and detected by Enzyme-Linked Immunosorbent Assay (ELISA).*

Results. *Five interested genes (IGFBP7, LRRC32, STMN2, C1QTNF5, CPXM1) were up-regulated in SSc microarray datasets from the GEO. And the level of serum IGFBP7, which encodes a secreted protein, was up-regulated in SSc patients-also in dcSSc patients and SSc with ILD patients.*

Conclusions. *Among the five interested genes, the IGFBP7 was a novel candidate gene for SSc and may be served as potential target and early biomarker for accurate treatment, which also provides further insights into the pathogenesis of SSc at the molecular level.*

Introduction

Systemic sclerosis (SSc), or scleroderma, is a heterogeneous connective tissue disease characterised by multi-organ fibrosis (1, 2). It can be classified as diffuse cutaneous SSc (dcSSc) and limited cutaneous SSc (lcSSc). SSc has the highest cause-specific mortality among all the rheumatic diseases (3). The cardiac factor is the leading cause of mortality, followed by lung involvement, both pulmonary hypertension and/or pulmonary fibrosis (4). The basic pathogenesis includes vascular damage, inflammation and connective tissue repair. Among them, the development of progressive systemic fibroproliferative process characteristic is crucial. However, it remains misty, which poses a threat to the effects of drugs for disease remission and reversion. Recently, target treatment of fibrosis in systemic sclerosis shows its prospect inspired by extensive studies. For example, tocilizumab, a kind of monoclonal antibody, shows its effectiveness and safety in the treatment of SSc associated interstitial lung disease (5).

Thereby, confronted with SSc, such an intractable autoimmune disease, we may turn to large-scale gene expression analysis using systems biology for some clues. Weighted correlation network analysis (WGCNA) (6), an R package for weighted correlation network analysis, has been previously successfully applied in various biological contexts to reveal the relationship between modules and clinical features and identify candidate biomarkers or therapeutic targets in several diseases. The advantage of a weighted co-expression network over an unweighted network lies in avoiding information loss by setting artificial threshold parameters in WGCNA (6).

In our study, we constructed a co-expression network of the expression profile data GSE58095 downloaded from the Gene Expression Omnibus (GEO) database in the environment of R (v. 3.6.1). Genes share similar biological function and biological processes are divided into the same co-expression module by clustering techniques. We confirmed the most SSc-related co-expression module and identified po-

tential functions of the genes within it by Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses. We also identified 27 real hub genes that possibly play a central role in SSc and constructed a protein-protein interaction (PPI) network to find key genes that interact with many other genes. Among them, we validated five genes, namely insulin-like growth factor binding protein 7 (*IGFBP7*), leucine-rich repeat-containing 32 (*LRRC32*), stathmin 2 (*STMN2*), complement C1q tumor necrosis factor-related protein 5 (*C1QTNF5*), carboxypeptidase X, M14 family member 1 (*CPXM1*), that were barely studied. Their biological functions have both similarities and differences. Previous studies show that *IGFBP7*, *STMN2*, *C1QTNF5* are all related to cell adhesion which is presented in the result of functional analysis. And *IGFBP7*, *STMN2*, *CPXM1* all take part in the process of osteogenesis and osteoblast. Besides, *IGFBP7* works in the activation and proliferation of fibroblasts; *C1QTNF5* influences extracellular deposits and participates in immune-mediated damage; *CPXM1* acts on collagen and extracellular matrix; *LRRC32* activates Regulatory T cells (Treg cells) to induce immune response by protecting *FOXP3* expression.

By referring to related data, we found that protein *LRRC32*, *STMN2* cannot be secreted into the serum. As for *CPXM1*, *C1QTNF5*, they are missing in GSE95065, which indicates their weak or unstable expression in human body. However, *IGFBP7* is not only free of the above-mentioned genetic defects, but also proved to be vital in the protein protein interaction (PPI) network. So, *IGFBP7* was the only interested gene we chose for further research. Here, we attempted to investigated the *IGFBP7* levels with immunological and clinical traits in 37 Chinese patients with SSc and 25 healthy control subjects (HCs). Furthermore, integrative microarray datasets of skin samples from patients with SSc and HCs were utilised to explore the underlying mechanism by which *IGFBP7* exerts its function in the pathogenesis of SSc through bioinformatic analysis. *IGFBP7* was also proved to possess a relatively satisfying diagnosis value through analysis. Based on these findings, we identified *IGFBP7* a potential candidate biomarker for SSc. Our findings may point to the potential candidate genes for accurate therapy of SSc and provide powerful evidences for better understanding the pathogenesis of SSc.

Material and methods

Patients and controls

A total of 37 SSc patients diagnosed as SSc according to ACR/EULAR 2013 (7) were recruited at Zhongshan hospital of Fudan University (Shanghai, China) in our study. The patients' clinical data at the time of SSc diagnosis were obtained through medical record reviews. 25 persons with no history of pulmonary, autoimmune, cardiovascular, or other diseases were recruited as healthy control subjects (HCs). The study was carried out in accordance with the recommendations of the Zhongshan Hospital Research Ethics Committee. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

Data collection and preprocess

All microarray datasets were obtained from the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/), a public data repository providing functional genomic information (8). To eliminate the interference as far as possible, the screening criteria we adopted were as follows: 1) Both SSc and healthy control groups are included for each gene expression dataset; samples from drug trials were excluded, except for those baseline and healthy control samples; replication samples were ought to be abandoned as well; 2) Tissues originate from skin biopsy on Homo sapiens; 3) A minimum of 10 SSc and healthy control samples in each array; 4) Inclusion of >5,000 genes in the GEO platform. Based on the selection principle above, 6 datasets were eligible in this study: GSE58095 (9), GSE32413 (10), GSE125362 (11), GSE76885 (12), GSE45485 (13), GSE95065 (14). The dataset utilised for WGCNA analysis was GSE58095, which consists of in total of 43 healthy control and 59 SSc skin samples. Other datasets were used for candidate genes validation. Information about these 6 datasets was summarised in Supplementary Table S1.

The probe annotation for GSE58095 was conducted under the R environment using the R package "limma" and "impute" with the microarray platform file. The gene expression value of GSE58095 has already been log2 transformed.

Co-expression network construction

We constructed a co-expression network by using the R package "WGCNA" in the R environment. Firstly, through variance analysis, we obtained the top 25% most variant genes for subsequent analysis. Then, we constructed an adjacency matrix based on Pearson's correlation analysis of all pairs of genes. Here, we needed to set soft-thresholding parameter β (15) to construct a scale-free co-expression network, namely a topological overlap matrix (TOM) (16). Using a dynamic tree-cutting algorithm (6) and the merging threshold function at 0.40, we merged the close modules into 13 modules.

Identification of key module and hub genes

We considered "SSc" and "normal" as clinical traits and calculated their correlation with the modules. It is regarded to contribute to the pathogenesis of the disease if the correlation between modules and SSc trait is positive. Through principal component analysis, we obtained 13 module eigengenes (MEs), the core component of the corresponding gene module, as an index to evaluate the degree of correlation. We extracted the gene module of the highest correlation with SSc for subsequent studies.

Hub genes usually play a key role in biological processes and gene regulation (17). A gene can be considered as a hub gene if it has a unique character, such as high gene significance (GS), high module membership (MM), and high intramodular connectivity (IC) in the network (15). Here are the criteria for hub genes we defined: 1) genes in the key module were among the top 5% genes of gene significance (GS); 2) with module group members (MM)

greater than 0.700; 3) with log fold change (logFC) greater than 0.500 in GSE58095. To find out the key genes which interact with many other genes in the key module, we built the protein-protein interaction (PPI) network using Search Tool for the Retrieval of Interacting Genes (STRING) online database (https://string-db.org/) (18). The minimum required interaction score was highest confidence (0.900). In Cytoscape software (version 3.7.2), we visualised the molecular interaction network and used "DMNC" algorithm in an plug-in named "cytohubba" to find top 25 genes that closely connected regions in this network.

Functional enrichment analysis of key module

To find out the functional and molecular features of genes in the key module, we referred to the Database for Annotation, Visualization and Integrated Discovery (DAVID) (https://david.ncifcrf.gov/home.jsp/) to perform the Gene Ontology analysis (GO) (19) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis. GO analysis is mainly described from the following three aspects: biological process (BP), cellular component (CC), molecular functions (MF). It indicated statistically significant if p-value <0.05, as well as a Benjamini value <0.05. We utilised the Tableau software (the version of 2019.4) to visualise the results. Meanwhile, we further studied the interested terms using the R package "GOplot" and "ggplot2".

Identification of hub genes in the key module

We identified the interested genes in GEO datasets, namely GSE58095, GSE32413, GSE125362, GSE76885, GSE45485, GSE95065, according to their expression levels in SSc and normal groups respectively. The R package "ggstatsplot" was used in this step.

Enzyme-linked immunosorbent assay (ELISA)

The level of human IGFBP-rp1/IGFBP-7 was detected by ELISA kits (RayBio®) according to the manufacturer's protocols.

Table I. Subject characteristics.

Variable of SSc (n=37)	n (%) or mean ± SD	Binary logistic regression			Variable of HCs (n=25)
		OR	95% CI	p-value	
Male	5 (13.51%)	1.263	0.157,10.157	0.826	8 (32.00%)
Female	32 (86.49%)				17 (68.00%)
Age (years)	52.58 ± 12.18	NA	NA	NA	45.64 ± 8.29
Course (years)	8.68 ± 11.10	NA	NA	NA	
mRSS	15.58 ± 12.05	NA	NA	NA	
dcSSc	22 (59.46%)	1.444	0.121,1.759	0.396	
lcSSc	15 (40.54%)				
ILD	19 (51.35%)	0.371	0.098,1.403	0.144	
PAH	5 (13.51%)	5.143	0.516,51.292	0.163	
Raynaud's phenomenon	32 (86.49%)	1.000	0.220,10.218	1.000	
Treatments Corticosteroid monotherapy	21 (56.76%)	0.800	0.093,6.848	0.839	
Corticosteroid & immunosuppressant therapy	11 (29.73%)	1.600	0.369,6.946	0.530	
Corticosteroid or immunosuppres-sant therapy	32 (86.49%)	0.588	0.086,4.009	0.588	
Pulmonary function test					
FVC (% predicted)	70.73 ± 8.13	NA	NA	NA	
DL_{CO} (%predicted)	60.98 ± 11.18	NA	NA	NA	
FEV1/FVC,%	83.59 ± 9.02	NA	NA	NA	
Anitibody examination					
Anti-scl-70	15 (40.54%)	0.556	0.147,2.103	0.387	
ANA	36 (97.28%)	0.000	NA	1.000	
ACA	10 (27.03%)	3.394	0.717,16.073	0.124	
Anti-RNP	8 (21.62%)	2.051	0.411,10.238	0.381	
Blood routine examination					
ESR, mm/h	20.03 ± 15.85	NA	NA	NA	
PLT, 10^9/L	202.89 ± 62.38	NA	NA	NA	
MPV, fL	11.12 ± 1.40	NA	NA	NA	
PDW, fL	0.22 ± 0.07	NA	NA	NA	
PCT, %	26.22 ± 12.89	NA	NA	NA	
P-LCR, %	20.31 ± 11.31	NA	NA	NA	

OR: odds ratio; CI: confidence interval; dcSSc: diffuse cutaneous systemic sclerosis; lcSSc: limited cutaneous systemic sclerosis; ILD: interstitial lung disease; PAH: pulmonary arterial hypertension; mRSS: the modified Rodnan skin score; FVC: forced vital capacity; DL_{CO}: diffusing capacity of the lung for carbon monoxide; FEV1: forced the first second of expiratory volume; ANA: anti-nuclear antibody; ACA: anti-centromere antibody; anti-RNP: anti-ribonucleoprotein antibody; ESR: erythrocyte sedimentation rate; PLT: platelet count; MPV: mean platelet volume; PDW: platelet distribution width; PCT: platelet crit; P-LCR: platelet large cell ratio; NA: not available.

Statistical analysis

Statistical analysis was performed with the IBM SPSS Statistics 26.0 and R software. Binary logistic regression was utilised to determine the risk factor of gene expression level to certain clinical traits by the IBM SPSS Statistics 26.0. A box plot was utilised to reveal the gene serum levels in different groups by the R package "ggsignif". Receiver operating characteristics (ROC) was utilised to calculate the area under the curve (AUC) of *IGFBP7* to evaluate the ability for the diagnosis of SSc by the R package "pROC". A p-value of <0.05 was considered statistically significant.

Results

Information about the patients and controls for validation

Clinical characteristics of 37 SSc patients for validation were summarised in Table I. There was more female (86.49%), more Raynaud's phenomenon (86.49%), and without pulmonary arterial hypertension (PAH) (86.49%) in SSc patients. The binary logistic regression analysis indicated that there was no relation between the risk factor of gene expression level and the certain clinical traits. HCs consisted of 8 men (32.00%) and 17 women (68.00%). The average age of 25 HCs was 45.64±8.29 years.

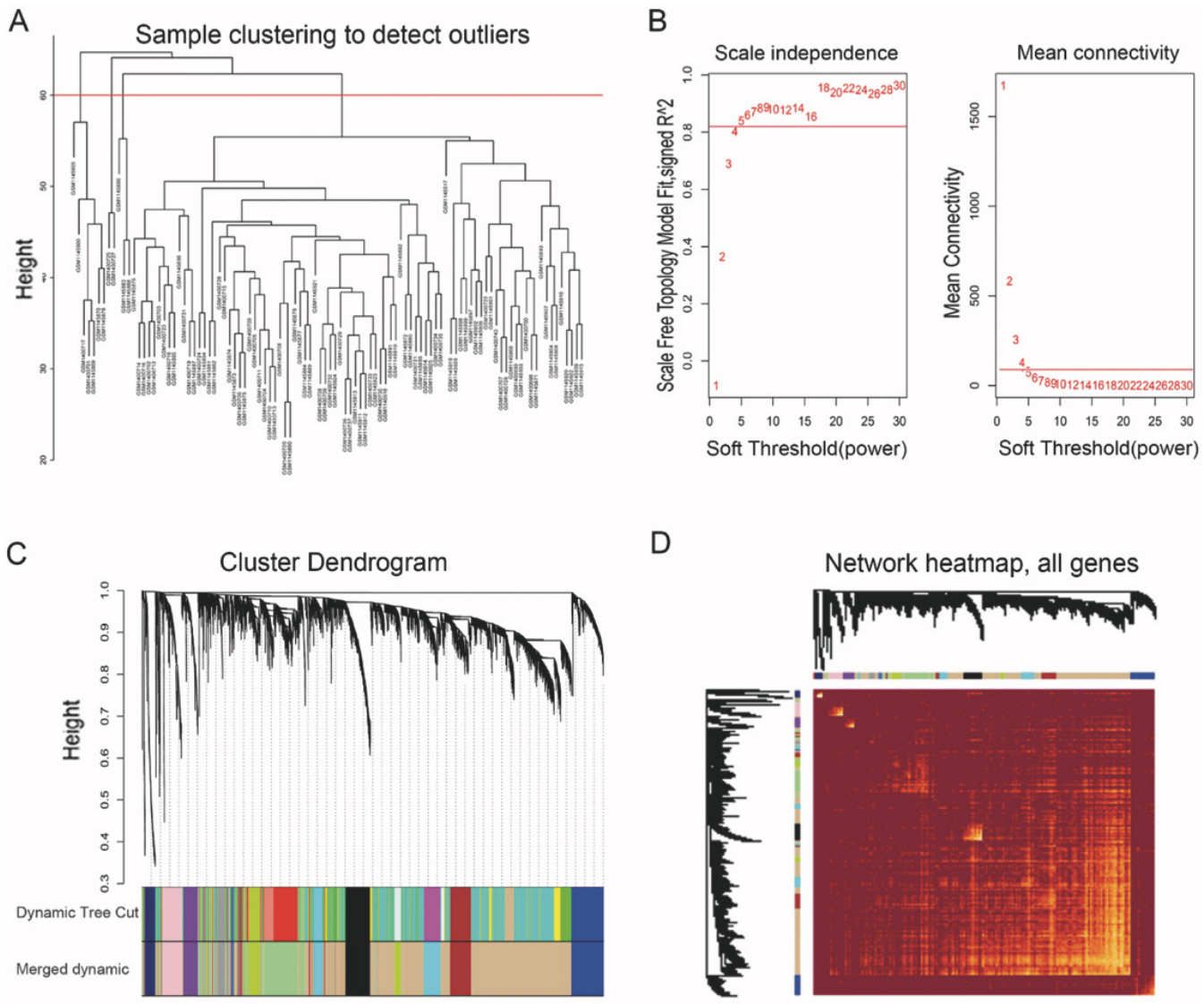


Fig. 1. Construction of weighted gene co-expression network.
(**A**) 12 samples (GSM1145905, GSM1145900, GSM1400717, GSM1400701, GSM1145869, GSM1145870, GSM1145878, GSM1400725, GSM1400727, GSM1145886, GSM1145883, GSM1145888) were excluded.
(**B**) Analysis of the scale-free topology model fit index for soft threshold powers (β) and the mean connectivity for soft threshold powers.
(C) The cluster dendrogram of genes in GSE58095. Each branch means one gene. Each colour means one co-expressed module.
(D) Interactive relationship analysis of co-expression genes. The light colour indicates topological overlap, while the darker colour indicates a high topological overlap.

Co-expression networks and key module

After filtering the genes by variance analysis, the expression profiles of 7,619 genes were left as the input data set of WGCNA. With a cut height of 60, we got rid of 12 outlier samples based on sample clustering with the hierarchical clustering method (20) (Fig. 1A). When 0.9 was used as the correlation coefficient threshold, the soft-thresholding power was selected as 5 (Fig. 1B). 17 co-expression modules were finally constructed after clustering with the TOM-based dissimilarity algorithm and merged into 13 modules according to their innate similarity (Fig. 1C). The eigengene adjacency heatmap (Fig. 1D) revealed the high independence between the co-expressed modules and gene expression in 13 modules.

Module-trait correlations analysis showed that multiple modules were related to SSc (Fig. 2A). Clearly, among them, the lightgreen (r=0.56, *p*=9e-09) module was the key module, with a scale of 770 genes. The dendrogram and adjacency heatmap of eigengenes (Fig. 3C) further indicated that the lightgreen module was closest to SSc traits. The correlation value between module membership in the lightgreen module and gene significance for SSc was 0.74 (Fig. 2B), suggesting reasonable relativity.

Functional annotation of the key co-expression module

The top 10 significant terms of GO and KEGG were exhibited in Figure 3. The complete results were given in Supplementary Tables S1 and 2.

For the most SSc-related module, enriched GO-BP terms were mainly

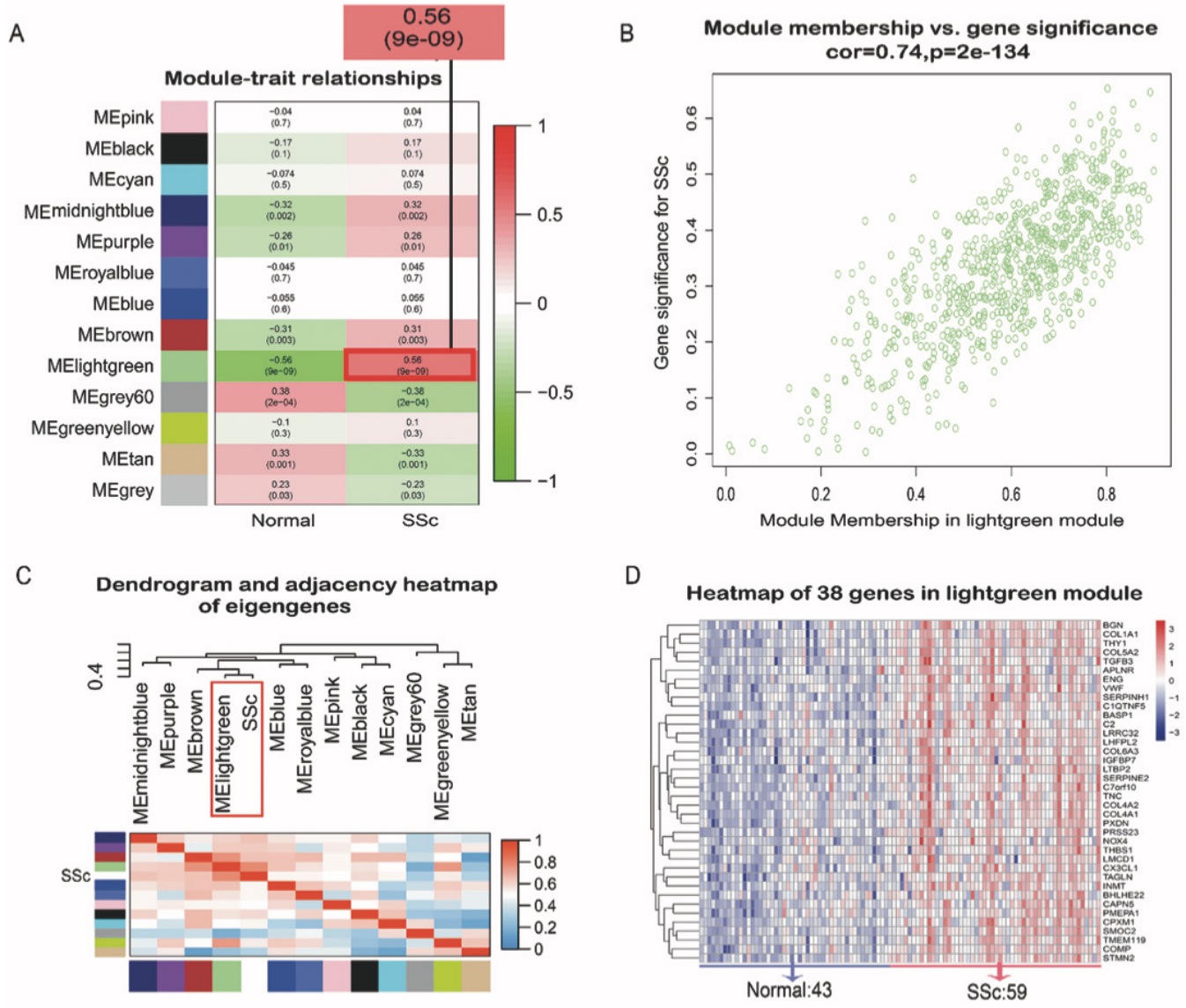


Fig. 2. Module-trait relationships.
(A) Heatmap of the correlation between module eigengenes and clinical traits. *P*-value is shown in each color cell coded by the correlation between modules and traits (red indicates positive correlation).
(B) Scatter plot of module eigengenes in the lightgreen module.
(C) Dendrogram and unsupervised hierarchical clustering heatmap of module eigengenes and SSc.
(D) Heatmap of 38 genes in the lightgreen module.
SSc: systemic sclerosis.

about "cell adhesion" (gene count=77, *p*=6.22E-29), "signal transduction" (gene count=73, *p*=1.08E-05), "immune response" (gene count=59, *p*=2.62E-18) (Fig. 3A). For GO-CC, enriched terms were generally involved in extracellular substance and membrane (Fig. 3B), such as "plasma membrane" (gene count=209, *p*=2.66E-08), "extracellular space" (gene count=129, *p*=3.91E-25). Enriched GO-MF terms were mainly about substance binding and extracellular matrix (Fig. 3C), such as "calcium ion binding" (gene count=54, *p*=1.07E-06), "extracellular matrix structural constituent" (gene count=20, *p*=1.98E-12), and so on. The results of KEGG enrichment were roughly about adhesion, signalling pathway and infection (Fig. 3D), such as "Focal adhesion" (gene count=37, *p*=2.24E-12), "PI3K-Akt signalling pathway" (gene count=31, *p*=5.66E-04).

Hub genes identification and dataset validation

Based on gene significance (GS) for SSc, there are 38 genes rank the top 5% of all genes in the lightgreen module (Table II). The heat map (Fig. 2D) revealed differences in expression in general between the two groups: down-regulated in the normal group while up-regulated in the disease group. According to the standards of hub genes we set, five genes (*APLNR*, *NOX4*, *BASP1*, *BHLHE22*, *PRSS23*) of them are with module membership (MM) less than 0.700; Besides, six genes (*LHFPL2*, *C2*, *CAPN5*, *LMCD1*, *PMEPA1*, *CX3CL1*) of the rest are with log fold change

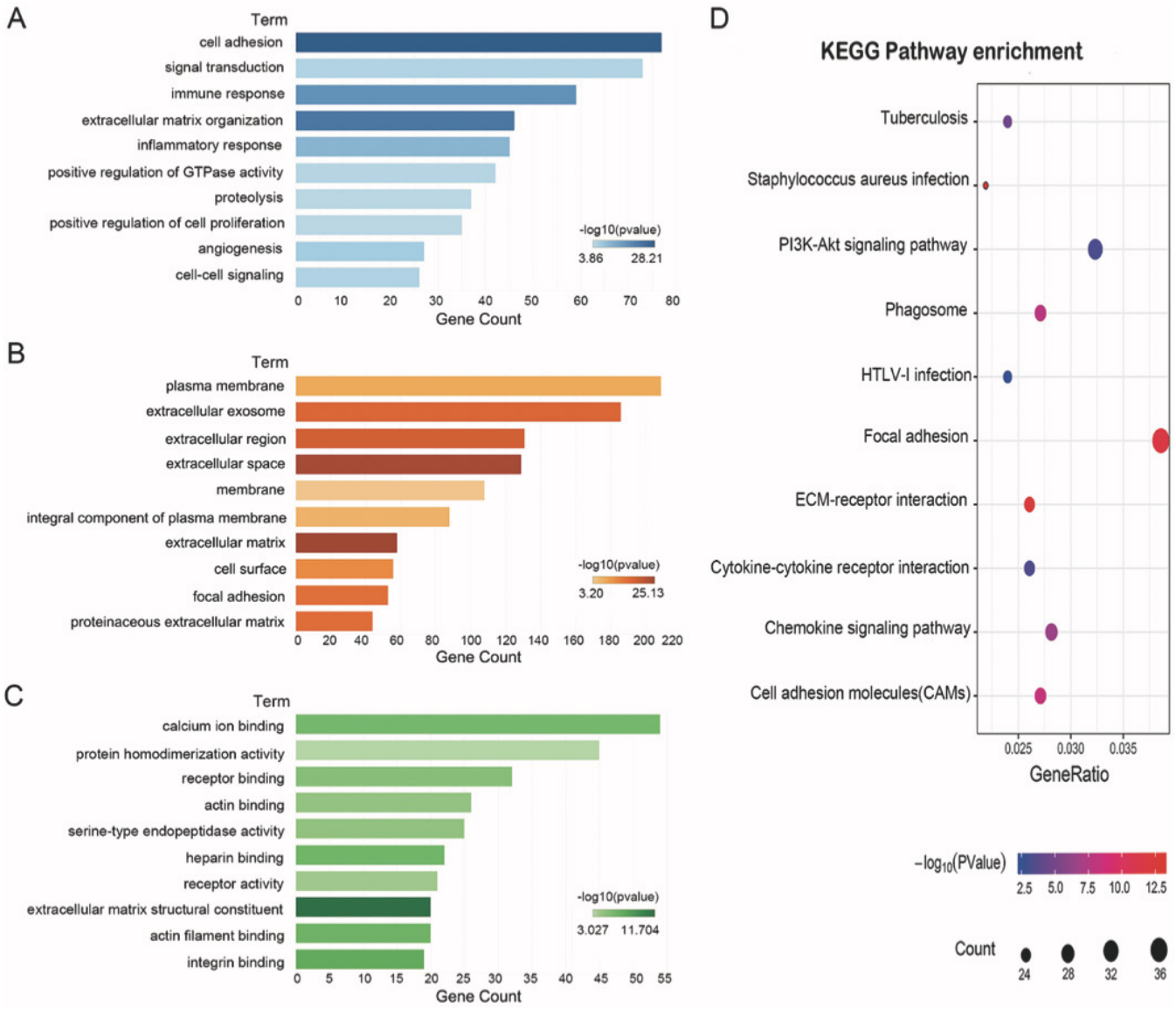


Fig. 3. GO and KEGG pathway enrichment analysis.
(A) Top 10 biological process (BP) terms in the lightgreen module. The length of each bar means the amounts of genes. The different color of each bar means -$\log_{10}$(*P*-value).
(B) Top 10 cellular component (CC) terms in the lightgreen module. The length of each bar means the amounts of genes. The different color of each bar means -$\log_{10}$(*P*-value).
(C) Top 10 molecular functions (MF) terms in the lightgreen module. The length of each bar means the amounts of genes. The different color of each bar means -$\log_{10}$(*P*-value).
(D) The top 10 KEGG enrichment pathways of genes in the lightgreen module. The size of each circle means the amounts of genes. The different color of each circle means -log10(*P*-value). GeneRatio means the ratio of genes in the key module that belong to this pathway divided by the number of genes in the background gene cluster that belong to this pathway.
GO: gene ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes.

(logFC) less than 0.500, which means moderate differential expression. Therefore, in the lightgreen module, there are 27 real hub genes in total. Most hub genes we found have already been confirmed their correction or significance with the disease, for example, *THY1* has been proposed as a potential marker of systemic sclerosis (21, 22). While 5 hub genes, namely *IGFBP7*, *LRRC32*, *STMN2*, *C1QTNF5*, *CPXM1*, have barely been studied before, indicating their potential influence in the progress of SSc and role for biomarkers.

384 genes whose *p*-value of GS or MM >0.0001 in the key module were removed. After excluding the isolated nodes, the PPI network was composed of 364 nodes and 497 edges (Fig. 4A). A significant densely-connected module was identified by "cytohubba" plug-in, which had 43 nodes and 290 edges (Fig. 4B). The top 25 nodes or genes were generally divided into 2 subnetworks: one including 14 genes (like *COL8A1*, *COL4A1*, *COL4A2*, *COL6A1*, *COL6A3*, *COL4A4*, *COL8A2*, *LEPRE1*, *LEPREL2*, *SERPINH1*, *PLOD1*, *COL1A1*, *COL5A2*, *COL1A2*) mainly encoding collage;

Table II. Top 5% genes in the lightgreen module of GSE58095 according to GS.

	Gene name	Value of GS	Value of MM	LogFC	Hub gene?
1	C1QTNF5	0.653	0.802	1.015	yes
2	THY1	0.647	0.892	1.387	yes
3	STMN2	0.623	0.730	1.269	yes
4	COL4A1	0.614	0.815	0.865	yes
5	CPXM1	0.604	0.786	0.897	yes
6	LHFPL2	0.594	0.813	0.445#	no
7	LTBP2	0.591	0.767	0.735	yes
8	COMP	0.591	0.730	1.475	yes
9	APLNR	0.584	0.615#	0.877	no
10	LRRC32	0.583	0.866	0.791	yes
11	VWF	0.578	0.795	0.673	yes
12	PXDN	0.575	0.814	0.813	yes
13	COL4A2	0.573	0.826	0.716	yes
14	C7orf10	0.573	0.794	0.709	yes
15	TMEM119	0.566	0.900	0.732	yes
16	C2	0.563	0.741	0.446#	no
17	TGFB3	0.559	0.794	0.626	yes
18	IGFBP7	0.556	0.810	0.531	yes
19	COL1A1	0.551	0.772	1.230	yes
20	PRSS23	0.545	0.694#	0.674	no
21	CAPN5	0.543	0.801	0.498#	no
22	NOX4	0.540	0.588#	0.372#	no
23	TAGLN	0.539	0.735	0.514	yes
24	INMT	0.538	0.794	0.633	yes
25	SMOC2	0.537	0.835	0.906	yes
26	COL5A2	0.534	0.748	0.828	yes
27	THBS1	0.531	0.713	0.832	yes
28	SERPINE2	0.530	0.731	0.774	yes
29	COL6A3	0.530	0.795	0.599	yes
30	BASP1	0.530	0.612#	0.470#	no
31	LMCD1	0.529	0.726	0.386#	no
32	PMEPA1	0.529	0.788	0.283#	no
33	CX3CL1	0.528	0.670	0.438#	no
34	TNC	0.527	0.758	0.599	yes
35	BHLHE22	0.526	0.628#	0.394#	no
36	BGN	0.526	0.714	0.574	yes
37	SERPINH1	0.526	0.771	0.607	yes
38	ENG	0.525	0.794	0.520	yes

GS: gene significance; MM: module membership; LogFC: log fold change.
The *p*-value of GS, MM, LogFC are all <0.01.
#the data is not up to the standard we set.

the other including 11 genes (such as *IGFBP7*, *PRSS23*, *TNC*, *STC2, SCG2, LGALS1, CYR61, CALU, TMEM132A, MXRA8, IGFBP4*).

These 5 genes above were chosen as interested genes for downstream validation. As we can see, *IGFBP7* (Fig. 5A), *LRRC32* (Fig. 5B), *STMN2* (Figure 5C) were significantly increased in the SSc patients in 6 datasets (GSE58095, GSE32413, GSE125362, GSE45485, GSE76885, GSE95065). Also, *C1QTNF5*, *CPXM1* were missing in GSE95065 for some reason. So *C1QTNF5* (Fig. 5D), *CPXM1* (Fig. 5E) were significantly increased in the SSc patients in 5 datasets (GSE58095, GSE32413, GSE125362, GSE45485, GSE76885).

Serum IGFBP7 level was increased in SSc patients

IGFBP7 is a gene encoding a secreted protein that can be detected in serum. Therefore, we conducted an ELISA test to detect the concentration of *IGFBP7* in the serum of 37 SSc patients and 25 HCs.

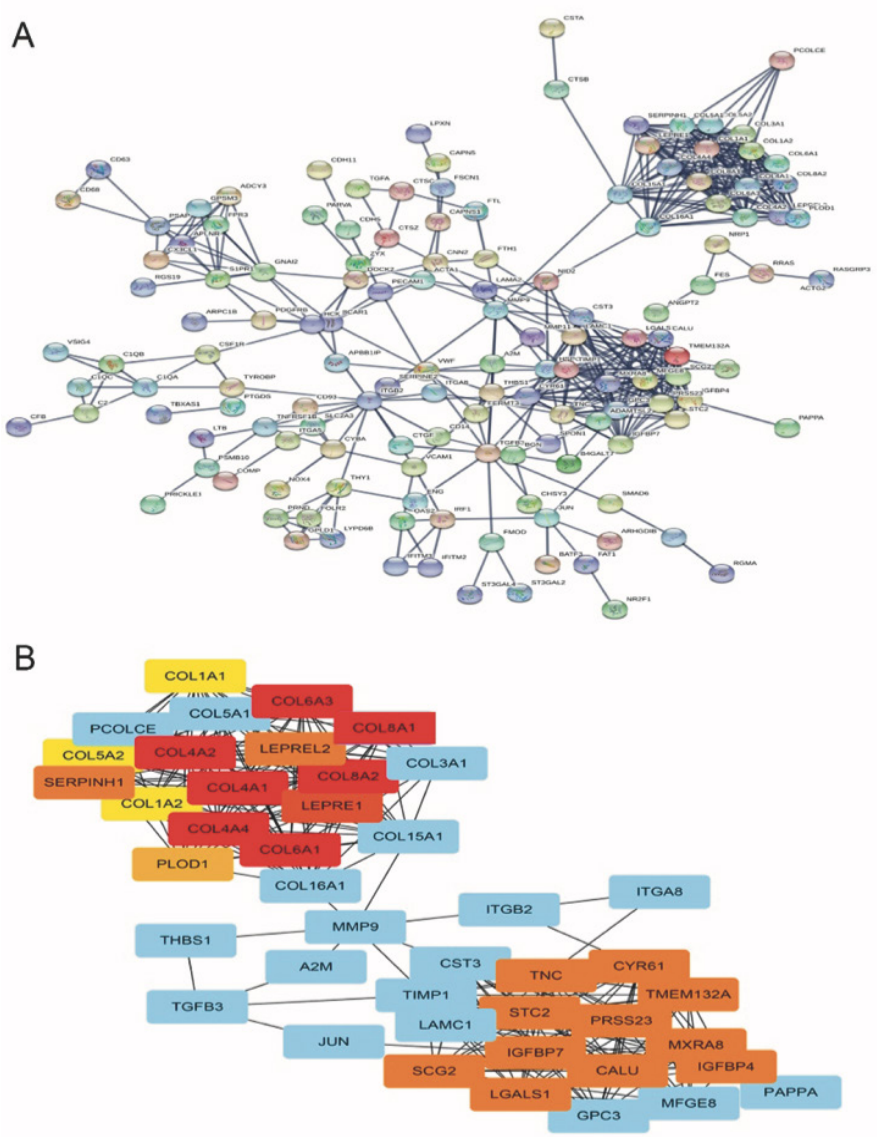
In Figure 6A, clearly, the expression levels of serum *IGFBP7* are statistically different between the SSc group and the control group. The mean value of serum *IGFBP7* in SSc patients was relatively higher than that in HCs (473.81±368.40 ng/mL *vs*. 281.65±183.76 ng/mL, *p*=0.0089). Also, the *IGFBP7* levels of dcSSc patients was significantly higher than that in HCs (532.63±371.91 ng/mL *vs*. 281.65 ±183.76 ng/mL, *p*=0.0074). Besides, the *IGFBP7* levels of SSc with ILD patients was significantly higher than that in HCs (549.68±350.19 ng/mL *vs*. 281.65 ±183.76 ng/mL, *p*=0.0055).

Based on the findings above, we further evaluated the ability of *IGFBP7* for the diagnosis of patients of SSc, dcSSc and SSc with ILD. In this step, we calculated the area under the curve (AUC) of ROC curves by using R software. The values of AUC in SSc, dcSSc, SSc with ILD, dcSSc combined with ILD patients were 0.649, 0.704, 0.762, 0.713, respectively (Fig. 6B). This result suggests that serum *IGFBP7* levels can offer a relatively satisfying diagnosis value, especially in SSc with ILD patients.

Discussion

The pathogenesis of SSc, a complex and heterogeneous disease, remains unclear. To discover novel biomarkers and therapeutic targets for SSc, numerous investigations utilising microarray and RNA-seq method were conducted. However, inconsistencies existed in the different expressed genes (DEGs) found in different studies. To the best of our knowledge, this is the first study to investigate the candidate genes of systemic sclerosis in skin biopsy using WGCNA analysis. WGCNA method has been successfully applied to explore the mechanisms of some diseases. For example, in a pan-cancer study, it was used to identify co-expression modules associated with cell cycle and thus providing therapeutic opportunities for cancer treatment (23). Through this approach, we identified the key gene co-expression module and validated five novel candidate genes such as *IGFBP7*, *LRRC32*, *STMN2*, *C1QTNF5*, and *CPXM1*.

Functional analysis revealed that genes in the key module were mainly enriched in cell adhesion, extracellular substance, and immune response. These three aspects all influence the development of progressive fibrosis, the prominent character of SSc. Recent reports showed that the excessive deposition of extracellular matrix components in connective tissues can contrib-



C

Rank	Name	Score
1	COL8A1	1.0685892
1	COL4A1	1.0685892
1	COL4A2	1.0685892
1	COL6A1	1.0685892
1	COL6A3	1.0685892
1	COL4A4	1.0685892
1	COL8A2	1.0685892
8	LEPRE1	1.0679305
9	IGFBP7	1.0515602
9	SCG2	1.0515602
9	STC2	1.0515602
9	TNC	1.0515602
9	LGALS1	1.0515602
9	IGFBP4	1.0515602
9	TMEM132A	1.0515602
9	CYR61	1.0515602
9	SERPINH1	1.0515602
9	MXRA8	1.0515602
9	PRSS23	1.0515602
9	LEPREL2	1.0515602
9	CALU	1.0515602
22	PLOD1	1.0247637
23	COL1A2	0.9990234
23	COL1A1	0.9990234
23	COL5A2	0.9990234

Fig. 4. PPI network and the subnetwork of the key module.
(A) Protein-protein interaction network of the key module. Each node represents a protein.
Different color of the nodes meant different types of protein. Each edge represents the interaction between proteins. The line thickness indicates the strength of data support.
(B) The significant densely-connected subnetwork identified by "cytohubba" plug-in. Different colours represent the degree of the genes.
(C) The rankings and scores of top 25 genes.

ute to fibrosis (24). Cell adhesion molecules can regulate the fibrotic process (25). The function of five candidate genes (*IGFBP7*, *LRRC32*, *STMN2*, *C1QTNF5*, *CPXM1*) we identified and validated were also found to be related to the pathology of SSc. *IGFBP7* (insulin-like growth factor binding protein 7) can interplay with extracellular matrix protein to induce cell adhesion and migration of endothelial cells (26, 27). And it might take part in the activation and proliferation of fibroblasts (28). *LRRC32* (leucine-rich repeat-containing 32) plays a critical role in immune regulation by safeguarding *FOXP3* expression in Treg cells (29), therefore it might be useful to treat autoimmunity and fibrotic diseases. *STMN2* (stathmin 2), as a crucial element of cytoskeletal regulation, functions in microtubule dynamics and cell migration through *RSK2* signals (30). *C1QTNF5* (complement C1q tumor necrosis factor-related protein 5) may participate in both cell-cell and cell-matrix adhesions (31) and influence extracellular deposits. The study of *CPXM1* (carboxypeptidase X, M14 family member 1) is rare. Its metallocarboxypeptidase family protein, *CPXM2*, is associated with extracellular matrix organisation, which may regulate early differentiation of connective tissues (32). However, the functions of *IGFBP7*, *STMN2* (33), *CPXM1* (34) are all related to bone metabolism, which was not presented in functional enrichment analysis results. For example, recombinant *IGFBP7* could induce a phenotypic switch from fibroblasts to osteoblasts (35). Interestingly, osteoclast function and osteogenic differentiation are regulated by osteopontin (36, 37), while osteopontin plays an important part in fibrosis

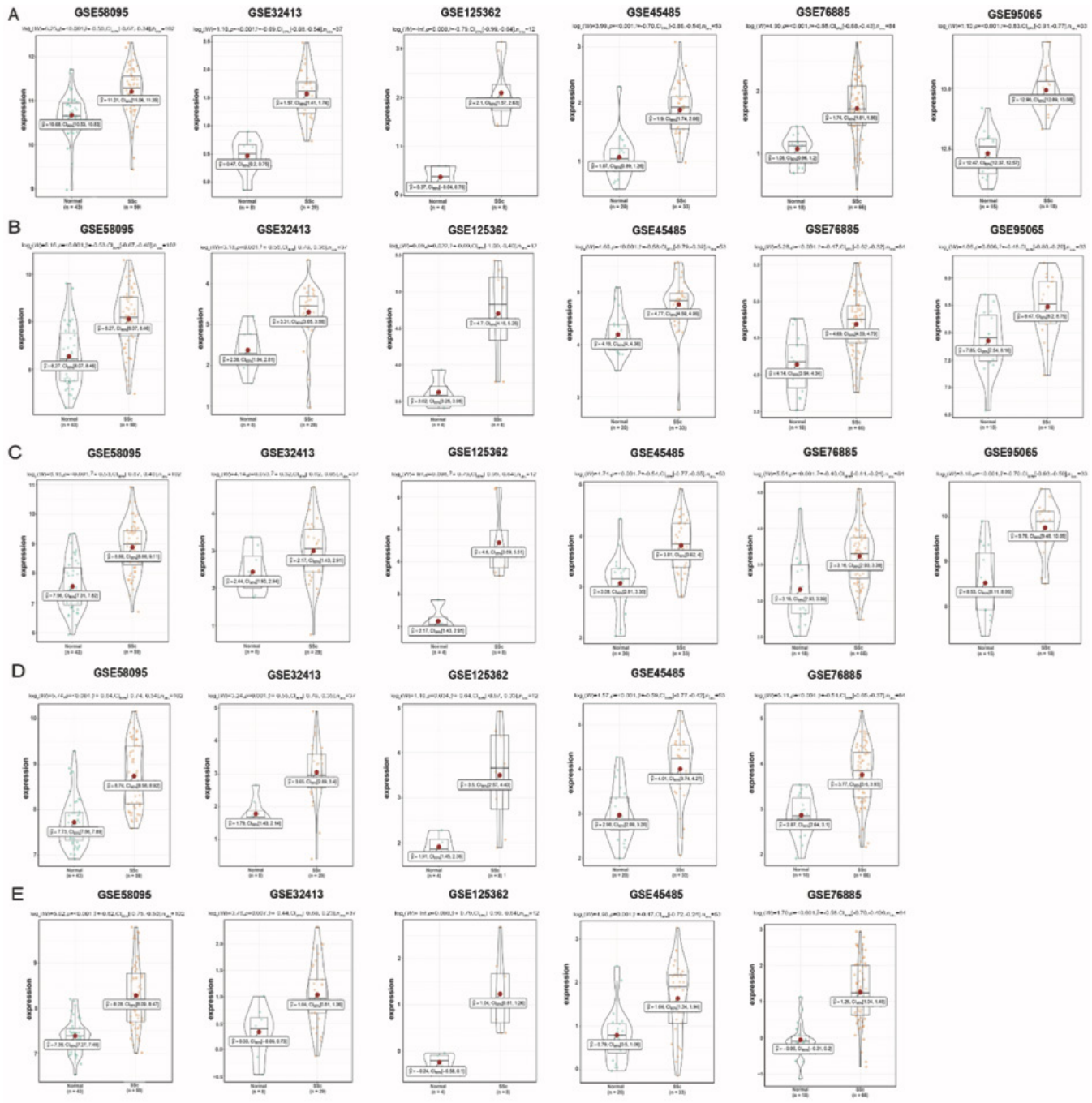


Fig. 5. Validation of 5 hub genes in datasets.
(A) *IGFBP7* expression level in GSE58095, GSE32413, GSE125362, GSE76885, GSE45485, GSE95065.
(B) *LRRC32* expression level in GSE58095, GSE32413, GSE125362, GSE76885, GSE45485, GSE95065.
(C) *STMN2* expression level in GSE58095, GSE32413, GSE125362, GSE76885, GSE45485, GSE95065.
(D) *C1QTNF5* expression level in GSE58095, GSE32413, GSE125362, GSE76885, GSE45485.
(E) *CPXM1* expression level in GSE58095, GSE32413, GSE125362, GSE76885, GSE45485.

through inducing fibroblast migration, proliferation and collagen production (38, 39). Therefore, it is likely that *IGFBP7* could influence the pathological process of fibrosis in systemic sclerosis via osteopontin.

Therefore, among the five candidate genes, we selected the *IGFBP7* key gene to be studied. It is a low-affinity insulin growth factor (IGF) binder, which belongs to the *IGFBP* superfamily. Firstly, like other 4 genes, it ranks top 5% (GS=0.56, MM=0.81) among 770 genes in the key module. Meanwhile, *IGFBP7* was one of the 25 key genes (score=1.05) in the PPI network, which suggested its central role in the key module. In other words, *IGFBP7* was proved to be a hub gene by both methods. Secondly, its probe data was not missing in any datasets, while *C1QTNF5*, *CPXM1* were missing in GSE95065. We found that the level of *IGFBP7* was significantly up-

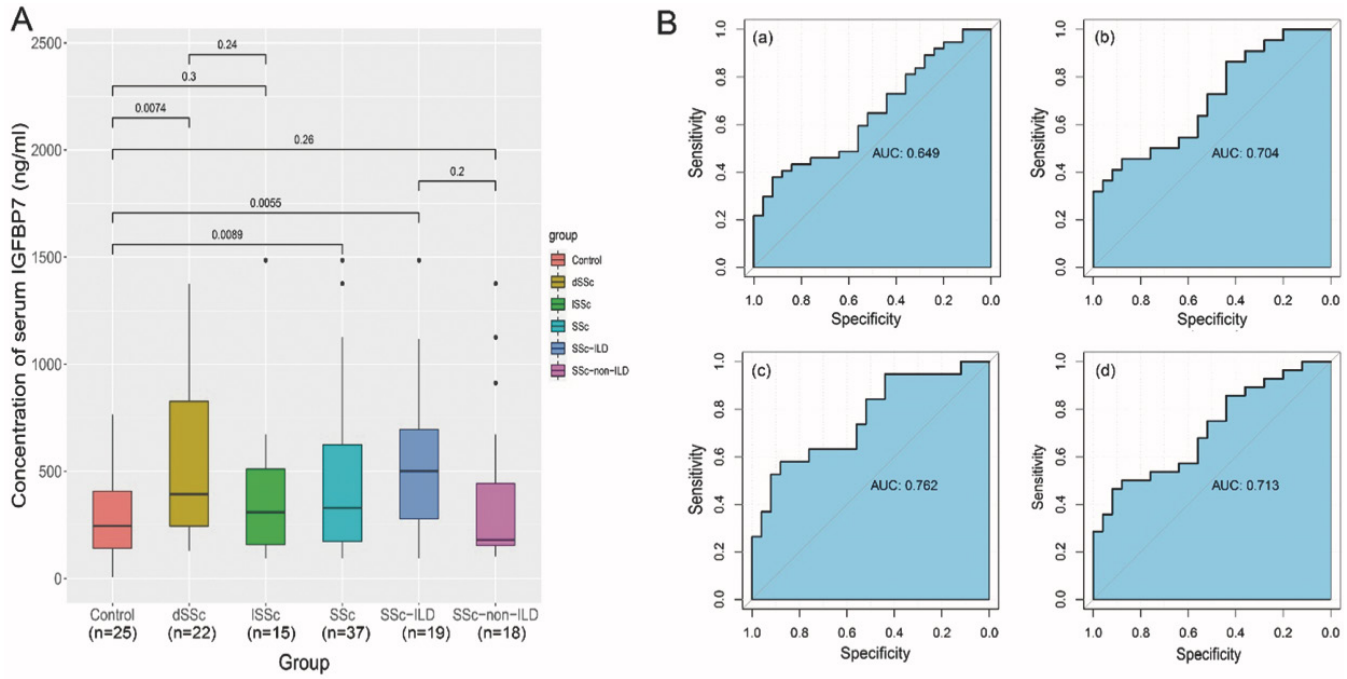


Fig. 6. Validation of *IGFBP7* with ELISA results.
(A) Box plot of serum *IGFBP7* levels in different groups. The different color of columns represents different groups. The height of the column means the concentration of *IGFBP7* protein in the serum. The numbers above the boxes represent *p*-value.
(B) Receiver operating characteristic curve (ROC). Serum *IGFBP7* level for the diagnosis of SSc patients (a), dSSc patients (b), SSc with ILD patients (c), and dSSc combined with ILD patients (d).

regulated in SSc patients in all datasets (GSE58095, GSE32413, GSE125362, GSE45485, GSE76885, GSE95065). Thirdly, *IGFBP7* is a gene encoding secreted protein that has been detected in serum in previous studies. Serum *IGFBP7* were increased and detected in patients with insulin resistance (IR) (23), esophageal squamous cell carcinoma (40) . Similarly, in our study, we found the *IGFBP7* protein level was elevated in the serum of SSc patients, dcSSc patients and SSc with ILD patients. It is likely that the higher the level of serum *IGFBP7*, the severer the degree of fibrosis process in organs. The detection in serum provides convenience for us in exploration and in the potential clinical appliance. Also, *IGFBP7* is a potential biomarker for the diagnosis of SSc patients, especially dcSSc patients and those with ILD. Compared with lcSSc patients and SSc without ILD patients, dcSSc patients and SSc with ILD patients tend to have a higher degree of fibrosis because there are more tissues involved. This can be explained as *IGFBP7* mainly functions in promoting fibrosis by inducing cell adhesion and activating fibroblasts.

Although the present study is the first to investigate SSc using WGCNA analysis, it has limitations. First of all, all of the datasets for exploration and validation were obtained from GEO online database. These studies used different platforms for gene expression analysis and were conducted on very distinct populations. And there is nothing we can do when *C1QTNF5*, *CPXM1* are missing in GSE95065. Secondly, we need to recruit more cases to alleviate statistical analysis error as far as possible and make our conclusion convincible. The five real core genes were demonstrated to be associated with the progression of SSc; nevertheless, we did not explore their pathways and signal transductions in detail and figure out the interactions between these genes and malignant features of SSc. Further studies should be performed to explore this issue.

In summary, our study found out the key gene co-expression module, which mainly enriched in cell adhesion, extracellular substance and immune response in the pathogenesis of SSc, and identified and validated five novel candidate genes (*IGFBP7*, *LRRC32*, *STMN2*, *C1QTNF5*, *CPXM1*). Among them, *IGFBP7* may serve as a promising prognostic predictor and therapeutic target for systemic sclerosis. The upregulation of *IGFBP7* may contribute to the fibrosis in the way of inducing cell adhesion and activating fibroblasts. These findings provide new insights into the development of SSc, although the exact molecular mechanism of candidate genes and functional pathways in SSc still need to be further explored.

References

1. DIDIER K, ROBBINS A, ANTONICELLI F, PHAM BN, GIUSTI D, SERVETTAZ A: Updates in systemic sclerosis pathogenesis: Toward new therapeutic opportunities. *Rev Med Interne* 2019; 40: 654-63.
2. SIERRA-SEPULVEDA A, ESQUINCA-GONZALEZ A, BENAVIDES-SUAREZ SA *et al*.: Systemic sclerosis pathogenesis and emerging therapies, beyond the fibroblast. *Biomed Res Int* 2019; 2019: 4569826.
3. TOLEDANO E, CANDELAS G, ROSALES Z *et al*.: A meta-analysis of mortality in rheumatic diseases. *Reumatol Clin* 2012; 8: 334-41.
4. ELHAI M, MEUNE C, BOUBAYA M *et al*.: Mapping and predicting mortality from systemic sclerosis. *Ann Rheum Dis* 2017; 76: 1897-905.
5. NARVAEZ J, L LUCH J, ALEGRE SANCHO JJ, MOLINA-MOLINA M, NOLLA JM, CASTELLVI

I: Effectiveness and safety of tocilizumab for the treatment of refractory systemic sclerosis associated interstitial lung disease: a case series. *Ann Rheum Dis* 2019; 78:123.
6. LANGFELDER P, HORVATH S: WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008; 9: 559.
7. VAN DEN HOOGEN F, KHANNA D, FRANSEN J *et al*.: 2013 classification criteria for systemic sclerosis: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Ann Rheum Dis* 2013; 72: 1747-55.
8. CLOUGH E, T BARRETT: The Gene Expression Omnibus Database. *Methods Mol Biol* 2016; 1418: 93-110.
9. ASSASSI S, SWINDELL WR, WU M *et al*.: Dissecting the heterogeneity of skin gene expression patterns in systemic sclerosis. *Arthritis Rheumatol* 2015; 67: 3016-26.
10. PENDERGRASS SA, LEMAIRE R, FRANCIS IP, MAHONEY JM, LAFYATIS R, WHITFIELD ML: Intrinsic gene expression subsets of diffuse cutaneous systemic sclerosis are stable in serial skin biopsies. *J Invest Dermatol* 2012; 5132: 1363-73.
11. FRANKS JM, MARTYANOV V, CAI G *et al*.: A machine learning classifier for assigning individual patients with systemic sclerosis to intrinsic molecular subsets. *Arthritis Rheumatol* 2019; 71: 1701-10.
12. HINCHCLIFF M, TOLEDO DM, TARONI JN *et al*.: Mycophenolate mofetil treatment of systemic sclerosis reduces myeloid cell numbers and attenuates the inflammatory gene signature in skin. *J Invest Dermatol* 2018; 138: 1301-10.
13. HINCHCLIFF M, HUANG CC, WOOD TA *et al*.: Molecular signatures in skin associated with clinical improvement during mycophenolate treatment in systemic sclerosis. *J Invest Dermatol* 2013; 133: 1979-89.
14. RICE LM, MANTERO JC, STIFANO G *et al*.: A proteome-derived longitudinal pharmacodynamic biomarker for diffuse systemic sclerosis skin. *J Invest Dermatol* 2017; 137: 62-70.
15. ZHANG B, HORVATH S: A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 2005; 4: Article17.
16. YIP A M, HORVATH S: Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics* 2007; 8: 22.
17. YU D, LIM J, WANG X, LIANG F, XIAO G: Enhanced construction of gene regulatory networks using hub gene information. *BMC Bioinformatics* 2017; 18: 186.
18. VON MERING C, HUYNEN M, JAEGGI D, SCHMIDT S, BORK P, SNEL B: STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* 2003; 31: 258-61.
19. HUANG DW, SHERMAN BT, ZHENG X *et al*.: Extracting biological meaning from large gene lists with DAVID. *Curr Protoc Bioinformatics* 2009; 13:13.11.
20. KIMES PK, LIU Y, NEIL HAYES D, MARRON JS: Statistical significance for hierarchical clustering. *Biometrics* 2017; 73: 811-21.
21. KOLLERT F, CHRISTOPH S, PROBST C *et al*.: Soluble CD90 as a potential marker of pulmonary involvement in systemic sclerosis. *Arthritis Care Res* (Hoboken) 2013; 65: 281-7.
22. NAZARI B, RICE LM, STIFANO G *et al*.: Altered dermal fibroblasts in systemic sclerosis display podoplanin and CD90. *Am J Pathol* 2016; 186: 2650-64.
23. LUO Z, WANG W, LI F *et al*.: Pan-cancer analysis identifies telomerase-associated signatures and cancer subtypes. *Mol Cancer* 2019; 18: 106.
24. SOBOLEWSKI P, MASLINSKA M, WIECZOREK M *et al*.: Systemic sclerosis - multidisciplinary disease: clinical features and treatment. *Reumatologia* 2019; 59: 221-233.
25. YOSHIZAKI A, YANABA K, IWATA Y *et al*.: Cell adhesion molecules regulate fibrotic process via Th1/Th2/Th17 cell balance in a bleomycin-induced scleroderma model. *J Immunol* 2010; 185: 2502-15.
26. KISHIBE J, YAMADA S, OKADA Y *et al*.: Structural requirements of heparan sulfate for the binding to the tumor-derived adhesion factor/angiomodulin that induces cord-like structures to ECV-304 human carcinoma cells. *J Biol Chem* 2000; 275: 15321-9.
27. SATO J, HASEGAWA S, AKAOGI K *et al*.: Identification of cell-binding site of angiomodulin (AGM/TAF/Mac25) that interacts with heparan sulfates on cell surface. *J Cell Biochem* 1999; 75: 187-95.
28. AKAOGI K, SATO J, OKABE Y, SAKAMOTO Y, YASUMITSU H, MIYAZAKI K: Synergistic growth stimulation of mouse fibroblasts by tumor-derived adhesion factor with insulin-like growth factors and insulin. *Cell Growth Differ* 1996; 7: 1671-7.
29. PROBST-KEPPER M, BALLING R, BUER J: FOXP3: required but not sufficient. the role of GARP (LRRC32) as a safeguard of the regulatory phenotype. *Curr Mol Med* 2010; 10: 533-9.
30. ALESI GN, JIN L, LI D *et al*.: RSK2 signals through stathmin to promote microtubule dynamics and tumor metastasis. *Oncogene* 2016; 35: 5412-5421.
31. STANTON CM, BOROOAH S, DRAKE C *et al*.: Novel pathogenic mutations in C1QTNF5 support a dominant negative disease mechanism in late-onset retinal degeneration. *Sci Rep* 2017; 7: 12147.
32. ZHAO X, LI R, WANG Q, WU M, WANG Y: Overexpression of carboxypeptidase X M14 family member 2 predicts an unfavorable prognosis and promotes proliferation and migration of osteosarcoma. *Diagn Pathol* 2019; 14: 118.
33. CHIELLINI C, GRENNINGLOH G, COCHET O *et al*.: Stathmin-like 2, a developmentally-associated neuronal marker, is expressed and modulated during osteogenesis of human mesenchymal stem cells. *Biochem Biophys Res Commun* 2008; 374: 64-8.
34. CHANG EJ, KWAK HB, KIM H, PARK JC, LEE ZH, KIM HH: Elucidation of CPX-1 involvement in RANKL-induced osteoclastogenesis by a proteomics approach. *FEBS Lett* 2004; 564: 166-70.
35. LU Z, CHIU J, LEE LR *et al*.: Reprogramming of human fibroblasts into osteoblasts by insulin-like growth factor-binding protein 7. *Stem Cells Transl Med* 2020; 9: 403-15.
36. ICER MA, GEZMEN-KARADAG M: The multiple functions and mechanisms of osteopontin. *Clin Biochem* 2018; 59: 17-24.
37. BAHRAMBEIGI V, SALEHI R, HASHEMIBENI B, ESFANDIARI E: Transcriptomic comparison of osteopontin, osteocalcin and core binding factor 1 genes between human adipose derived differentiated osteoblasts and native osteoblasts. *Adv Biomed Res* 2012; 1: 8.
38. PARDO A, GIBSON K, CISNEROS J *et al*.: Up-regulation and profibrotic role of osteopontin in human idiopathic pulmonary fibrosis. *PLoS Med* 2005; 2: 251.
39. TAKAHASHI F, TAKAHASHI K, OKAZAKI T *et al*.: Role of osteopontin in the pathogenesis of bleomycin-induced pulmonary fibrosis. *Am J Respir Cell Mol Biol* 2001; 24: 264-71.
40. HUANG X, HONG C, PENG Y *et al*.: The Diagnostic value of serum IGFBP7 in patients with esophageal squamous cell carcinoma. *J Cancer* 2019; 10: 2687-93.