Supplementary Information

Detecting the Molecular System Signatures of Idiopathic Pulmonary Fibrosis through Integrated Genomic Analysis

Indu Gangwar[1Ɨ3], Nitesh Kumar Sharma[1Ɨ3], Ganesh Panzade[13], Supriya Awasthi[1], Anurag Agrawal[2], Ravi Shankar[13]*

Affiliations:

[1] Studio of Computational Biology & Bioinformatics, CSIR-IHBT, Palampur (HP), India.

[2] Lung Diseases Translational Research Group, CSIR-IGIB, Mall Road, Delhi, India.

[3]Academy of Scientific and Innovative Research (AcSIR), CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT) Palampur (HP), India

Ɨ Contributed Equally

***Corresponding Author:** ravish@ihbt.res.in*

Supplementary Methods

High throughput data and processing

All initial data from microarray and RNA-seq based high throughput studies were collected from GEO using GEOquery 2.23.2[1] and SRA with accession IDs SRP010041 and SRP033095. Detailed list of data information has been made available in Supplementary Table S1. CUFFDIFF 2.2.0 was used for differential gene expression analysis with default parameters where FDR (<=0.05) was used for the selection of differentially expressed genes. Each microarray experiment's data was categorized into two classes, Normal and IPF individuals, while excluding all other specific condition. P-values were adjusted using Benjamini & Hochberg FDR. As in practice, genes with significant P-value (<0.05) and -1<LogFC >1 were considered as differentially expressed genes. Genes which displayed inconsistent and opposite expression status across the samples for any given condition were removed and only those genes were considered which exhibited single consistent trend in any given experiment for the given condition. This removed noise in the micro-array data. Z-score transformation of the expression data was done for all kind of platforms, in order to maintain a single comparable scale.

Two MA experiments, GSE31934 and GSE10921, were excluded due to extreme inconsistency and coverage of only a small set of human genes. This way, finally four microarray studies and two RNA-seq studies were taken forward for further study (GSE35147, GSE21411, GSE32537, GSE24206, SRP033095 and SRP010041). The four microarray studies were sorted on the basis of number of individuals in which GSE32537 and GSE24206 had maximum numbers. Both these studies have a consensus of 82 differentially expressed genes (Set S1) which were already reported in another RNA-seq study, SRP033095. Due to inconsistency in data as well as very small sample size, the RNA-seq experiment, SRP010041 was removed from the study. Considering the third microarray study, GSE21411 (11 IPF and 6 Control Individuals) and the RNA-seq based study,

SRP033095 (8 IPF and 7 Control), an overlap of 176 genes (Set S2) was found. To find out the consistency of these 176 genes, a consensus of these genes and previously reported 82 genes (Set S1) was made which resulted into a set of 39 overlapping genes, named as Set A in this study. In order to distinguish normal and IPF individual, k-means clustering was done using sets S1 and A. Clustering was performed on set S1 and A among five microarrays and two RNA-seq experiments using k-means function in R. Another set of 737 genes was prepared named as Set B which had common DE genes at least in two entirely separate studies.

Classification of IPF genes

For classification purpose, different datasets were created for training and testing using a support vector machine approach. A total of 159 IPF cases worked as positive samples while a total of 112 normal, sarcoidosis and NSIP cases were considered as negative and non-IPF cases. For each considered gene set, samples were randomly distributed into training and testing model 10 times, each time yielding different training and testing set pair and model. This was done to ensure that the performance of the classifier was not biased towards only certain combination of samples. LibSVM with RBF kernel and optimized cost and gamma factor was run. The models were checked for accuracy, sensitivity and specificity on different randomly generated test sets. Receiver Operator Characteristics plots with Area Under curve were made using ROCR function[2] in R. Matthews correlation coefficient (MCC) was calculated to measure perfection of SVM based classification.

Following equations were used to measure the classifier's performance:

$$Sensitivity = (TP / (TP + FN)) * 100$$

$$Specificity = (TN / (TN + FP)) * 100$$

$$Accuracy = (TN + TP) / (TP + TN + FP + FN)$$

$$MCC = (TP * TN) - (FP * FN) / \sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}$$

where TP= True Positive, TN= True Negative, FP=False Positive and FN= False Negative

Further system wide analysis was done for three datasets (Set A and Set B) as they were found to be most confident differentially expressed sets.

TRANSFAC and CHIP-Seq identified TF-target interactions

Besides the manual curation, the approach for the identification of transcription factors involved two more methods: Scanning through 1) TRANSFAC v8.3 and 2) CHIP-seq data. Identification was done for both sets (Set A, B). The first method includes the identification of potentially TF regulated differentially expressed genes, their 2kb upstream promoter regions up to first exon were identified and were searched for transcription factor binding sites using PROMO 3.0. This tool incorporates TRANSFAC database (v8.3) having largest available collection of eukaryotic factor specific weight matrices. TFBS and transcription factor associated with them were identified on the basis of sequence similarity to the matrix. Similarity threshold of 85% was applied to obtain high quality results. The Second method was based on the availability of TF peaks data from regulatory map of transcription factor binding sites (ReMap) repository as it contains validated transcription factor CHIP-seq peak data from well-established repositories ENCODE and GEO. Available transcription factor peak bed files for all 237 transcription factors were downloaded from this repository and their binding was checked across the target genes in both the data sets using bedtools. Details related to the TF-target interactions in the regulation of differentially expressed genes is covered in supplementary Table S6 and list of identified TF is given in supplementary Table S7.

miRNA data

So far three microarray studies have been published for miRNAs (GSE27430, GSE21394 and GSE32538). The study has been done using all these high-throughput sudies and a RNA-seq study data. Details are covered in the main text.

Statistical significance of Feed-Forward loops

Let suppose two edges E1 and E2 are connecting nodes N0 to N1 and N4 to N5, respectively. These were chosen randomly and rewired such that nodes were swapped constructing edges E1 =(N0-N5) and E2= (N4-N1). P-value of significance was calculated as proportion of randomized networks having particular FFL count greater than or equal to its number in real network to the total number of randomized networks (1000 in our study).

Network properties of gene regulatory network in IPF

Degree of a vertex/node in the network is the number of edges connecting to it. High degree nodes are characterized as hub nodes in the network. Centrality estimates the importance of node in a network. Betweenness centrality of a node is defined as the total occurrences of node in form of a bridge across shortest path of two other nodes. Farness of a node is a measure of sum of distances of the node with all other nodes and reciprocal of it is defined as the closeness centrality of the node. Tendency of a node in a graph to remain clustered together is referred as clustering coefficient whereas average shortest path length is described as mean value of steps through shortest paths of all potential node pairs of network. Indegree and outdegree terms are defined for directed networks in form of edges converging towards the node and diverging from the node respectively.

Pathway crosstalk network analysis in IPF

We initiated the crosstalk network analysis by filtering out the pathways containing less than six genes, considering sufficient number of genes to address biological relevance of analysis. In this way, we obtained an initial set of 67 KEGG pathways for crosstalk network analysis in IPF. Pathway pairs were built such that pairs with significant overlap of genes were discarded. Furthermore, genes common to both pathways were discarded to remove self-interactions between pathway pairs.

Protein interactions occurring between all pathway pairs were counted. Every pathway pair was randomized 1000 times and protein interaction counts in real network were compared with randomized ones. Only those randomized trials were taken into account having protein interactions greater than or equal to its number in the real network. To select the significant pathways crosstalk, probability of significance of pathway pairs was calculated as ratio of random permutations of pathway pairs having protein interaction count higher or equal to its number in real network with total randomized trials.

Table S1: Accuracy, sensitivity, specificity and MCC values for sets S1 and A in contrast to ten different random models

Set A								
	TP	**FN**	**FP**	**TN**	**Accuracy**	**Sensitivity**	**Specificity**	**MCC**
Shuffle 1	78	2	6	50	94.12	97.5	89.29	0.88
Shuffle 2	78	2	5	51	94.85	97.5	91.07	0.89
Shuffle 3	76	4	8	48	91.18	95	85.71	0.82
Shuffle 4	77	3	4	52	94.85	96.25	92.86	0.89
Shuffle 5	78	2	8	48	92.65	97.5	85.71	0.85
Shuffle 6	78	2	7	49	93.38	97.5	87.5	0.86
Shuffle 7	75	5	10	46	88.97	93.75	82.14	0.77
Shuffle 8	80	0	7	49	94.85	100	87.5	0.9
Shuffle 9	77	3	11	45	89.71	96.25	80.36	0.79
Shuffle 10	79	1	9	47	92.65	98.75	83.93	0.85
Set S1								
	TP	**FN**	**FP**	**TN**	**Accuracy**	**Sensitivity**	**Specificity**	**MCC**
Shuffle 1	76	4	11	45	88.97	95	80.36	0.77
Shuffle 2	72	8	9	47	87.5	90	83.93	0.74
Shuffle 3	73	7	7	49	89.71	91.25	87.5	0.79
Shuffle 4	77	3	8	48	91.91	96.25	85.71	0.83
Shuffle 5	79	1	9	47	92.65	98.75	83.93	0.85
Shuffle 6	76	4	12	44	88.24	95	78.57	0.76
Shuffle 7	78	2	5	51	94.85	97.5	91.07	0.89
Shuffle 8	76	4	7	49	91.91	95	87.5	0.83
Shuffle 9	80	0	11	45	91.91	100	80.36	0.84
Shuffle 10	76	4	6	50	92.65	95	89.29	0.85

Table S2: Functional role and status of Set A DEGs in IPF

Gene Name	Status in IPF	Role in IPF
Caveolin-2 (CAV1/2)	Down	It is a membrane protein which is present in lung fibroblasts and suppress TGFβ induced ECM production but in case of IPF, its decrease causes excessive TGFβ induced ECM production.
CBS	Down	It regulates homocysteine in plasma membrane and deficiency of CBS genes in mouse model already published as one of cause of fibrosis and inflammation
CDH3	Up	It works as a molecular target in various cancer like pancreatic, colorectal and gastric cancer[3]. Upregulation of this gene supports the proliferation and adhesion in IPF lungs
SFRP2	Up	Secreted frizzled related protein 2 works as inhibitor of Wnt signaling pathway and upregulation of this gene inhibits BMP2 gene which is a key factor in collagen maturation and biosynthesis during injury condition[4]
COL14A1	Up	It is a type of collagen, a member of the FACIT (fibril-associated collagen with interrupted triple helices) collagen family. This protein interacts with fibril surface and is directly involved in the regulation of fibrillogenesis.
COMP	Up	Cartilage oligomeric matrix protein (COMP) is a non-collagenous ECM protein, knockdown of this gene inhibits cell proliferation which shows direct impact of this gene under IPF condition[5]
CXCL14	Up	Production of CXCL14 is done by macrophages, fibrotic loci and type 2 alveolar epithelial cells, it's a mucosal chemokine which recruits many immune cell on mucosa to prevent a cell against foreign particle working as an anti-microbial gene[6]
DCLK1	Up	Doublecortin-like kinase 1 (DCLK1) is a member of kinase superfamily and worked as marker for cancer stem cell and directly involved in epithelial-mesenchymal transition to enhance it which is also a key phenotype of IPF[7]
DIO2	Up	Deiodinase iodothyronine type II (dio2) is a member of iodothyronine family which activates thyroid hormone by converting the prohormone thyroxin. From the study it was found that upregulation of this gene changes in signaling of thyroid hormone which may play a role in development and progression of IPF but direct involvement was not observed in IPF[8]
EPB41L5	Down	Down regulated Erythrocyte membrane protein band 4.1 like 5 (EPB41L5) is used as a biomarker for idiopathic pulmonary fibrosis[9]
FAM167A	Down	Downregulation of FAM167A is a novel finding discovery in this study and was

		found regulated by miR-27a and transcription factors (ATF3 and VDR). Previous studies reported that FAM167A is involved in sclerosis, a specific type of fibrosis, where fibrous deposition in an area occurs which has some sort of flexibility or movement like skin
FZD5	Down	Frizzled class receptor 5 (FZD5) encodes 7-transmembrane domain protein that are receptors for Wnt signaling believed to be the receptor for Wnt5a ligand and downregulation of this gene reduces the angiogenesis[10]
HMGCR	Down	It is a rate limiting factor for fatty acid biosysnthesis, and downregulation of this enzyme leads to downregualtion of fatty acid biosynthesis.
IGFBP4	Up	Elevated expression of Insulin-like growth-factor-binding proteins (IGFBP4) shows inhibitory effect on canonical Wnt signaling which plays a crucial role in IPF
KCNMA1	Up	Potassium calcium-activated channel subfamily M alpha 1 (KCNMA1) was found to act as muscle relaxant. In IPF condition its expression was elevated and altering lungs muscle cell tone.
KRT17	Up	Upregulated Keratin 17 involved in tissue repair and formation and maintenance of various skin appendages by regulating protein synthesis and epithelial cell growth through binding to the adapter protein SFN.
LTBP1	Up	Latent TGF β protein-1 (LTBP1) was found up-regulated in our study. A previously reported study suggested that alteration of TGF-β pathways alters the extracellular matrix composition in fibrotic lung[11]
MMP16	Up	Matrix metallopeptidase 16 (MMP16) was found upregulated and has an important involvement in the breakdown of ECM in normal physiological processes, like tissue remodeling, reproduction and embryonic development as well as in chronic lung diseases.
PAMR1	Up	Peptidase domain containing associated with muscle regeneration 1 (PAMR1) function is not very well defined but it was upregulated and it is believed that it may play an important role in formation of skeletal muscle.
PCDH7	Up	Protocadherin (PCDH7) belongs to cadherin super family and have role in cell-cell recognition and adhesion. Increased expression of this gene supports elevated cell adhesion process in IPF.
Thy-1	Up	Thy-1 was found upregulated and it is a glycoprotein encoded by cell surface antigens and elevates cell adhesion and cell communication.

Table S3: Differentially expressed TF-target interaction analysis for Set A and Set B genes.

TF- target interaction analysis for a Set B genes					
Differentially expressed genes	Correlation >=0.7	TF	Target gene	Pathways identified from target enrichment -Pvalue	GO-terms P-value
up-regulated (468)	2037 interactions	11	381	ECM-receptor interaction-(p=1.66E-11) integrin family cell surface interactions-(p=7.86E-16) Focal adhesion-(p=5.01E-10) IGF1 pathway-(p=8.02E-12)	Cell adhesion-(p=2.00E-12) ECM organization-(p=1.50E-12) Growth factor binding-(p=6.40E-05)
down-regulated (269)	1721 interactions	13	241	Metabolic pathway-(p=2.33E-06) Endocytosis- (p=7.42E-06)	Immune system process-p(1.17E-05) response to wounding (p= 1.08E-06) RAGE receptor binding-(p=0.0118)
TF- target interaction analysis for a Set A genes					
Differentially expressed genes	Correlation >=0.7	Transcription Factor	Target genes	Pathways identified from target enrichment -P value	GO terms P-value
up-regulated (22)	132 interactions	11	21	TGF- β signaling pathway (p=0.0008)	Cell adhesion(p=0.0317)
down-regulated (17)	110 interactions	13	15	Metabolic pathways (p=0.0005) Metabolism of lipids and lipoproteins (p= 0.0001)	Sterol biosynthetic process(p=7.68e-06)

Table S4: Differentially expressed up-regulated TF classification and role in IPF

Transcription factor	Class	Family	Expression in study	Function	Target Involvement in up-regulated pathways associated with IPF identified in study	Target genes GO enriched terms associated with IPF identified in study	role in IPF
SOX2- Sex Determining Region Y -Box 2	High mobility group domain (HMG) domain factors	SOX-related factors	up-regulated	proliferation and differentiation	beta1 integrin cell surface interactions, SIP1 pathway,ECM receptor interaction pathway	Cell adhesion, ECM organization.	Imbalanced activity of airway epithelial cells,results in formation of alveolar lesions in IPF[12]
SOX4- Sex Determining Region Y -Box 4	High mobility group domain (HMG) domain factors	SOX-related factors	up-regulated	mesenchymal cell expression	ECM-receptor interaction, Focal adhesion pathway, beta1 integrin cell surface interactions, SIP1 pathway	Cell adhesion, ECM organization, growth factor binding	Higher expression is responsible for mesenchymal formation from EMT[13]
EBF1- Early-B-cell related factor	Rel Homology Region factors	Early B-Cell Factor-related factors	up-regulated	differentiation and maturation of B-cells	ECM-receptor interaction, Focal adhesion pathway	Cell adhesion, ECM organization, growth factor binding	B-cell aggregation[14]
SIX1- Sine oculis homeobox-1	Homeo domain factors	HD-SINE factors	up-regulated	mesenchymal cell formation	beta1 integrin cell surface interactions, IGF1 pathway	Cell adhesion, ECM organization	Higher expression is responsible for EMT transitions[15]
SIX4- Sine oculis homeobox-4	Homeo domain factors	HD-SINE factors	up-regulated	cell proliferation	beta1 integrin cell surface interactions, signaling events mediated by focal adhesion kinase	Cell adhesion, ECM organization	Higher expression enhance the myofibroblast differentiation[16]
TFAP2A-Transcription factor AP-2-alpha	Basic helix-span-helix factors	AP-2 factors	up-regulated	cell growth and differentiation	ECM-receptor interaction, Focal adhesion, beta1 integrin cell surface interactions	Cell adhesion, ECM organization, growth factor binding	upregulation of TFAP2A promotes cell differentiation[17]
MYOCD-Myocardin	Basic-helix-loop-helix factors	ASC-related factors	up-regulated	Cell proliferation and differentiati	ECM receptor interaction,Focal adhesion, beta1 integrin cell	Cell adhesion, ECM organization,	Progressive and irreversible scarring of lung tissue[18]

	(Bhlh)			on	surface interactions	growth factor binding	
TP63-Tumor suppressor protein 63	P53 domain factors	P53 related factors	up-regulated	cell proliferation	ECM receptor interaction,Focal adhesion, beta1 integrin cell surface interactions	Cell adhesion, ECM organization, growth factor binding	Disordered proliferation of epithelial cells results in squamous metaplasia as one of the classified form of honeycomb structure in IPF[19]
TRIM29-Tripartite motif-containing protein29	C2H2 zinc finger factors	Trim-family	up-regulated	cell proliferation and transformation	ECM receptor interaction,Focal adhesion, P53 signaling pathway	ECM organization, ECM assembly, growth factor binding	Higher expression results in inhibition of apoptosis[20]
MYB-Transcription l activator Myb	Tryptophan cluster factors	Myb/SANT domain factors	up-regulated	cell proliferation	ECM receptor interaction,Focal adhesion, beta1 integrin cell surface interactions	Cell adhesion, ECM organization, growth factor binding	Enhance myofibroblast formation and differentiation from ECM transition[21]
BHLHE22-Class E basic helix-loop-helix protein 22	Basic-helix-loop-helix factors (Bhlh)	Tal- related factors	up-regulated	cell proliferation	ECM receptor interaction, Focal adhesion, beta1 integrin cell surface interactions	Cell adhesion, ECM organization, growth factor binding	Higher expression is responsible for enhance cell proliferation

Table S5: Differentially expressed down-regulated TF classification and role in IPF

Transcription factor	Description	Class	Family	Expression in study	Function	Involvement in down-regulated pathways associated with IPF identified in study	GO enriched terms associated with IPF identified in study	role in IPF
CEBPD	CCAAT/enhancer binding protein delta	Basic leucine zipper factors (bZIP)	C/EBP family	down-regulated	tumor suppressor function	Metabolic pathways, Endocytosis	Response to wounding, RAGE receptor binding	enhance fibroblast proliferation[22]
CSRNP1	Cysteine rich nuclear protein	AXUD/ CSRNP domain factors	CSRNP factors	down-regulated	tumor suppressor function	Metabolic pathways, Endocytosis	Response to stress, response to wounding	Down-regulation inhibits tumor suppression activity[23,24]
FOSL2	FOSL2	Basic leucine zipper factors (bZIP)	Fos-related factors	down-regulated	cellular proliferation	Metabolic pathways, Endocytosis	Response to stress, response to wounding	Down-regulation inhibits its anti-proliferation activity[25]
HIF3A	Hypoxia-inducible factor-3-alpha	Basic-helix-loop-helix factors (Bhlh)	PAS domain factors	down-regulated	mediates oxidative stress	Metabolic pathways, Endocytosis	Response to stress, response to wounding	Down-regulation leads to hypoxia induce damage[26]
HOPX	Homeodomain- only protein	Homeo domain factors	Paired-related HD factors	down-regulated	developing pulmonary airway	Endocytosis, Metabolic pathways	Response to chemical stimulus, RAGE receptor binding	Down-regulation lead to alveolar damage[27]
ID1	Inhibitors of Differentiation	Basic-helix-loop-helix factors (Bhlh)	ID family	down-regulated	inhibits differentiation	Endocytosis, Metabolic pathways	Response to wounding, RAGE receptor binding	Down-regulation induces impairment and fibrosis[28]
JUN	Jun	Basic leucine zipper factors (bZIP)	Jun-related factors	down-regulated	cell cycle progression and apoptosis	Endocytosis, Metabolic pathways	Response to stimulus, regulation of catalytic	Down-regulation induces reduction in apoptosis[25]

							activity	
KLF6	Kruppel - like factor 6	C2H2 zinc finger factors	Three-zinc finger Kruppel-related factors	down-regulated	differentiation and apoptosis	Endocytosis, Metabolic pathway	Response to stimulus, oxidoreductase activity	dysregulation of apoptosis in vasculature in IPF[29]
KLF9	Kruppel- like factor 9	C2H2 zinc finger factors	Three-zinc finger Kruppel-related factors	down-regulated	Regulates cell proliferation, B-cell proliferation, apoptosis	Endocytosis, Metabolic pathways	Response to stress, response to wounding	Down-regulation inhibits its apoptotic activity.
KLF15	Kruppel- like factor 15	C2H2 zinc finger factors	Three-zinc finger Kruppel-related factors	down-regulated	Cellular proliferation inhibition	Endocytosis, Metabolic pathway	Cellular response to stimulus, carbohydrate binding	Down-regulation inhibits its activity of regulating cellular proliferation[30]
NFE2	Nuclear factor erythroid 2	Basic leucine zipper factors (bZIP)	Jun-related factors	down-regulated	Regulation of megakaryocytic differentiation and regulation of platelet production	Endocytosis, Metabolic pathways	Immune system process, response to wounding, RAGE receptor binding	Down-regulation results in dysregulation of platelet production[31]
SMAD6/ MADH6	Mother against decapentaplegic homolg 6	SMAD/ NF-1 DNA-binding domain factors	SMAD factors	down-regulated	repressor activity	Endocytosis, Metabolic pathways	Response to stimulus, oxidoreductase activity	Down-regulation inhibits its antifibrotic activity[32]
AFF3	AF4	NA	AF4/FMR2 family member 3	down-regulated	inhibiting apoptosis	Endocytosis, Metabolic pathways	Response to external stimulus, oxidoreductase activity	Down-regulation inhibits its apoptosis activity[33]

Table S6: Transcription factor-target analysis using Promo (transfac v8.3) and ReMap

TF-target interaction analysis using PROMO for a Set A genes			
Differentially expressed genes	**TF-target interactions Correlation >=0.7**	**Transcription Factor**	**Target Genes**
up-regulated (22)	294 interactions	30	20
down-regulated (17)	152 interactions	24	15
TF-target interaction analysis using PROMO for a Set B genes			
up-regulated (468)	4103 interactions	34	375
down-regulated (269)	2976 interactions	34	230
TF-target interaction analysis using ReMap for a Set A genes			
Differentially expressed genes	**TF-target interactions Correlation >=0.7**	**Transcription Factor**	**Target Genes**
up-regulated (22)	325 interactions	78	21
down-regulated (17)	297 interactions	68	14
TF-target interaction analysis using ReMap for a Set B genes			
up-regulated (468)	4748 interactions	97	343
down-regulated (269)	4775 interactions	95	222

Table S7: List of TF regulating differentially expressed genes (A total 107 TFs were identified from both methods (TRANSFAC and CHIP-Seq). Enrichment analysis of their target genes showed the involvement of similar pathways as well as the involvement of same biological and molecular functions as identified in case of differentially expressed transcription factors.)

AR	EZH2	LEF1	RFX5	VDR
ATF1	FLI1	MAFF	RUNX1	WT1
ATF3	FOS	MAFK	RUNX1T1	XBP1
ATRX	FOSL1	MEF2C	RUNX2	ZEB1
BACH1	FOXA1	MEIS1	SMAD1	ZKSCAN1
BCL11A	FOXA2	MTA3	SMC1A	ZNF143
BCL3	FOXM1	MYC	SMC3	ZNF217
BCL6	FOXP2	NANOG	SMC4	
BHLHE40	GABPA	NCOR2	SNAPC1	
BRCA1	GATA2	NFATC1	SNAPC5	
CDK8	GATA3	NFATC2	SPI1	
CEBPA	GATA6	NFKB1	SRF	
CEBPB	GREB1	NKX2-1	STAT1	
CHD1	HDAC2	NOTCH1	STAT2	
CHD2	HMGN3	NR2F1	STAT4	
EGR1	HNF1B	PAX5	SUZ12	
EGR3	HNF4G	PML	TAF1	
ELF5	HOXD10	POU5F1	TAL1	
EOMES	IKZF1	PPARG	TCF4/TCF7L2	
ERG	IRF1	PPARGC1A	NR2F2	
ESR1	IRF4	PRDM1	TEAD4	
ESR2	JUNB	RAD21	TFAP2C	
ETS1	JUND	RARB	THAP1	
ETS2	KDM5B	RBPJ	TP53	
ETV1	KLF4	RCOR1	USF1	

Table S8: Novel, overlapping and contradictory set of DEmiRNAs during IPF

Upregulated	**Downregulated**
hsa-let-7c, hsa-miR-23b, hsa-miR-34a, hsa-miR-34c-3p, hsa-miR-34c-5p, hsa-miR-100, hsa-miR-133a, hsa-miR-135b, hsa-miR-302d, hsa-miR-328, hsa-miR-363, hsa-miR-429, hsa-miR-567, hsa-miR-634, hsa-miR-921, hsa-miR-1249	hsa-let-7d, hsa-let-7i, hsa-miR-28-3p, hsa-miR-33b, hsa-miR-139-3p, hsa-miR-146a, hsa-miR-342-3p, hsa-miR-346, hsa-miR-486-5p, hsa-miR-550, hsa-miR-580, hsa-miR-593, hsa-miR-601, hsa-miR-623, hsa-miR-627, hsa-miR-632, hsa-miR-636, hsa-miR-638, hsa-miR-641, hsa-miR-671-5p, hsa-miR-874, hsa-miR-877, hsa-miR-933, hsa-miR-939, hsa-miR-1228, hsa-miR-1287
Previously known miRNAs in our study	
Upregulated	**Downregulated**
hsa-miR-1, hsa-miR-10a*, hsa-miR-21, hsa-miR-31, hsa-miR-31*, hsa-miR-34b, hsa-miR-99a, hsa-miR-125b, hsa-miR-132, hsa-miR-133b, hsa-miR-143, hsa-miR-148a, hsa-miR-154, hsa-miR-155, hsa-miR-199a-5p, hsa-miR-199b-5p, hsa-miR-205, hsa-miR-214, hsa-miR-214*, hsa-miR-299-5p, hsa-miR-376a, hsa-miR-376b, hsa-miR-376c, hsa-miR-377, hsa-miR-379, hsa-miR-381, hsa-miR-382, hsa-miR-409-3p, hsa-miR-410, hsa-miR-411, hsa-miR-487b, hsa-miR-493*, hsa-miR-495, hsa-miR-509-5p, hsa-miR-539, hsa-miR-650, hsa-miR-654-3p	hsa-let-7b, hsa-let-7g, hsa-miR-17, hsa-miR-18a, hsa-miR-23a, hsa-miR-27a, hsa-miR-29c, hsa-miR-30a, hsa-miR-30b//hsa-miR-30b*, hsa-miR-30c-1*, hsa-miR-30c-2*, hsa-miR-30d, hsa-miR-30e, hsa-miR-33a, hsa-miR-92a, hsa-miR-93, hsa-miR-106b, hsa-miR-125a-3p, hsa-miR-126, hsa-miR-141, hsa-miR-151-3p, hsa-miR-181a, hsa-miR-181b, hsa-miR-181d, hsa-miR-184, hsa-miR-191, hsa-miR-203, hsa-miR-193-3p, hsa-miR-210, hsa-miR-221, hsa-miR-221*, hsa-miR-222, hsa-miR-223, hsa-miR-224, hsa-miR-320a, hsa-miR-326, hsa-miR-342-5p, hsa-miR-345, hsa-miR-362-5p, hsa-miR-375, hsa-miR-378, hsa-miR-422a, hsa-miR-423-5p, hsa-miR-425, hsa-miR-489, hsa-miR-500, hsa-miR-502-3p, hsa-miR-532-3p, hsa-miR-532-5p, hsa-miR-652, hsa-miR-663, hsa-miR-668, hsa-miR-744
Contradicted miRNAs with our study from previous studies	
Our study upregulated	**Previously known as downregulated**
hsa-miR-29a	hsa-miR-29a
Our study downregulated	**Previously known as upregulated**
hsa-miR-92b/hsa-miR-92b*, hsa-miR-324-3p, hsa-miR-765	hsa-miR-92b/hsa-miR-92b*, hsa-miR-324-3p, hsa-miR-765

Table S9: Differentially expressed top ten miRNAs targeting maximum number of genes and their functional enrichment.

Upregulates miRNA target pathway enrichment			
miRNA	**Enriched KEGG pathway**	**Target genes**	**P-value of enrichment**
hsa-miR-10a*	Steroid biosynthesis	16	0.000022
	Metabolic pathways		0.0078
hsa-miR-133a	Basal cell carcinoma	19	0.0003
	Endocytosis		0.0035
	Pathways in cancer		0.0089
hsa-miR-23b*	Terpenoid backbone biosynthesis	21	4.51E-008
	Metabolic pathways		0.0000124
	Steroid biosynthesis		0.0000384
	Melanogenesis		0.0011
hsa-miR-302d	Melanogenesis	17	0.0007
	Neuroactive ligand-receptor interaction		0.0051
hsa-miR-31	Terpenoid backbone biosynthesis	16	0.0000135
	Steroid biosynthesis		0.000022
	Metabolic pathways		0.000042
hsa-miR-34b*	Steroid biosynthesis	17	4.91E-008
	Metabolic pathways		0.00000308
	Biosynthesis of unsaturated fatty acids		0.0000306
	Glycine, serine and threonine metabolism		0.000072

	Basal cell carcinoma		0.0002
	Pathways in cancer		0.0072
hsa-miR-34c-3p	Terpenoid backbone biosynthesis	32	0.000000168
	Steroid biosynthesis		0.0000905
	Biosynthesis of unsaturated fatty acids		0.0001
	Metabolic pathways		0.0002
	Basal cell carcinoma		0.0008
hsa-miR-495	Neuroactive ligand-receptor interaction	30	0.0000373
	Pathways in cancer		0.0000752
	Focal adhesion		0.0004
	Endocytosis		0.0086
hsa-miR-539	Terpenoid backbone biosynthesis	36	0.000000241
	Steroid biosynthesis		0.000000513
	Metabolic pathways		0.00000342
	Biosynthesis of unsaturated fatty acids		0.0001
	Basal cell carcinoma		0.001
	Drug metabolism - cytochrome P450		0.0017
hsa-miR-654-3p	Neuroactive ligand-receptor interaction	18	0.0057
	Metabolic pathways		0.0109
Top 10 Downregulated miRNA target pathway enrichment			
hsa-miR-17*	Protein digestion and absorption	18	0.0005

	ECM-receptor interaction		0.0006
	Amoebiasis		0.0009
	Focal adhesion		0.0031
hsa-miR-210	Hypertrophic cardiomyopathy (HCM)	31	2.97e-05
	Focal adhesion		0.0004
	mTOR signaling pathway		0.0006
	p53 signaling pathway		0.0011
	Protein digestion and absorption		0.0016
	Dilated cardiomyopathy		0.0019
	Salivary secretion		0.0019
	Leukocyte transendothelial migration		0.0032
	Cell cycle		0.0036
hsa-miR-30a	Hypertrophic cardiomyopathy (HCM)	53	0.0044
hsa-miR-30d*	Hypertrophic cardiomyopathy (HCM)	26	1.73e-05
	mTOR signaling pathway		0.0005
	Arrhythmogenic right ventricular cardiomyopathy (ARVC)		0.0009
	Dilated cardiomyopathy		0.0014
	Cell adhesion molecules (CAMs)		0.0029
	Focal adhesion		0.0065
hsa-miR-30e	Cytokine-cytokine receptor interaction	37	0.0014
	Arrhythmogenic right ventricular cardiomyopathy (ARVC)		0.0018
	Hypertrophic cardiomyopathy (HCM)		0.0022
hsa-miR-33b*	Focal adhesion	21	0.0001
	Protein digestion and absorption		0.0006
	ECM-receptor interaction		0.0007

	Cell adhesion molecules (CAMs)		0.0017
	Cytokine-cytokine receptor interaction		0.0066
hsa-miR-345	Hypertrophic cardiomyopathy (HCM)	27	0.0012
	Cell adhesion molecules (CAMs)		0.0029
hsa-miR-532-5p	Leukocyte transendothelial migration	25	0.0019
	Chemokine signaling pathway		0.0049
	Cytokine-cytokine receptor interaction		0.0095
hsa-miR-627	Mucin type O-Glycan biosynthesis	59	0.0007

Table S10: Description of various Feed-Forward loops and interaction in IPF-mediated regulatory networks

	Motif	Number of FFLs	Genes	miRNAs	TFs	MiRNA-gene	MiRNA-TF	TF-gene	TF-miRNA	Gene-Gene
Set A genes network	3node TF-FFL	17	5	7	8	13	-	15	11	-
	3node MiRNA-FFL	246	33	39	35	137	86	161	-	-
	Total	263	33	43	42	143	86	176	11	-
Set A genes network	4node TF-FFL	-	-	-	-	-	-	-	-	-
	4node MiRNA-FFL	13	14	5	5	16	5	16	-	5
	Total	13	14	5	5	16	5	16	-	5
Set B genes network	3node TF-FFL	65	36	19	21	49	-	63	32	-
	3node MiRNA-FFL	1250	242	64	64	584	153	963	-	-
	Total	1315	243	73	77	607	153	1026	32	-
Set B genes network	4node TF-FFL	1	2	1	1	2	-	2	1	1
	4node MiRNA-FFL	93	43	20	19	77	32	112	-	42
	Total	94	43	21	20	79	32	114	1	42

Table S11: Hub components of regulatory network in terms of IPF genes, IPF miRNAs and TFs

Set B Genes regulatory network							
IPF Genes	**Indegree/ Outdegree**	**Function**	**IPF miRNA**	**Indegree/ Outdegree**	**TFs**	**Indegree/ Outdegree**	**Function**
PDE4D	105/1	Alveolar Epithelial Cell (AEC) proliferation	miR-627	1/61	XBP1	4/458	Regulate cell proliferation and angiogenesis
NEDD9	95/0	Cell adhesion	miR-30a	6/58	RARB	1/386	Regulate cell growth and differentiation
FLT1	94/6	Angiogenesis and vasculogenesis	miR-30e	0/45	CEBPB	1/368	Control of genes involved in immune responses
RASGEF1B	93/0	Ras guanyl-nucleotide exchange factor activity	miR-539	2/36	AR	3/363	Androgen receptor responsible for regulating gene transcription
TMEM2	88/0	Transmembrane protein	miR-495	0/34	FOXA1	0/363	Regulation of metabolism and differentiation
Set A Genes regulatory network							
FZD5	76/0	Receptor of Wnt signaling proteins	miR-30e	0/19	TCF4/TCF7L2	5/27	Key regulator of Wnt signaling
KCNMA1	76/0	Repolarization of membrane potential	miR-30a	6/17	VDR	0/26	Regulation of cell proliferation and calcium homeostasis
SYTL2	71/0	Vesicle trafficking	miR-30d*	0/16	ETS1	2/25	Regulation of genes belonged to development and apoptosis
SLC39A8	62/0	Transporter of Zn and toxic cadmium ion	miR-345	0/14	STAT1	0/25	Stimulation of interferons
DCLK1	60/0	Neurogenesis, neuronal migration and apoptosis	miR-539	2/13	XBP1	4/24	Regulate cell proliferation and angiogenesis

Table S12: Detailed information of high-throughput data considered for study from different Data sources

Information of IPF Samples					
Datatype	**Dataset ID**	**Tissue type**	**#No. of patient**	**#No. of Normal individuals**	**Data Sources**
RNA-seq	SRP010041	Lung	3	3	SRA
RNA-seq	SRP033095	Lung	8	7	SRA
Gene Microarray	GSE10921	Lung	4	3	GEO
Gene Microarray	GSE21411	Lung	11	6	GEO
Gene Microarray	GSE31934	Lung	3	3	GEO
Gene Microarray	GSE35147	Lung	4	4	GEO
Gene Microarray	GSE32537	Lung	119	50	GEO
Gene Microarray	GSE24206	Lung	17	6	GEO
miRNAMicroarray	GSE27430	Lung	13	12	GEO
Information of Sarcoidosis Samples					
Gene Micro-array	GSE16538	Lung	6	6	GEO
Gene Micro-array	GSE19976	Lung	15	-	GEO
Information of NSIP Samples					
Gene Micro-array	GSE5774	Lung	12	-	GEO

References:

1. Davis, S. & Meltzer, P. S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 23, 1846–1847 (2007).
2. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. Bioinformatics 21, 3940–3941 (2005).
3. Yoshioka, H. et al. In vivo therapeutic effect of CDH3/P-cadherin-targeting radioimmunotherapy. Cancer Immunol. Immunother. 61, 1211–1220 (2012).
4. He, W. et al. Exogenously administered secreted frizzled related protein 2 (Sfrp2) reduces fibrosis and improves cardiac function in a rat model of myocardial infarction. Proc. Natl. Acad. Sci. U.S.A. 107, 21110–21115 (2010).
5. Vuga, L. J. et al. WNT5A Is a Regulator of Fibroblast Proliferation and Resistance to Apoptosis. Am J Respir Cell Mol Biol 41, 583–589 (2009).
6. Burkhardt, A. M. et al. CXCL17 Is a Mucosal Chemokine Elevated in Idiopathic Pulmonary Fibrosis That Exhibits Broad Antimicrobial Activity. J Immunol 188, 6399–6406 (2012).
7. P, C. & J, P. Regulatory Roles of Dclk1 in Epithelial Mesenchymal Transition and Cancer Stem Cells. Journal of Carcinogenesis & Mutagenesis 7, (2016).
8. Guoying Yu et al. in C68. OLD AND NEW PATHWAYS TO PULMONARY FIBROSIS A4932–A4932 (American Thoracic Society, 2012).
9. Bauer, Y. et al. A novel genomic signature with translational significance for human idiopathic pulmonary fibrosis. Am. J. Respir. Cell Mol. Biol. 52, 217–231 (2015).
10. Arderiu, G., Espinosa, S., Peña, E., Aledo, R. & Badimon, L. Monocyte-secreted Wnt5a interacts with FZD5 in microvascular endothelial cells and induces angiogenesis through tissue factor signaling. J Mol Cell Biol 6, 380–393 (2014).

11. Leppäranta, O. et al. Regulation of TGF-β storage and activation in the human idiopathic pulmonary fibrosis lung. Cell Tissue Res. 348, 491–503 (2012)
12. Plantier, L. et al. Ectopic respiratory epithelial cell differentiation in bronchiolised distal airspaces in idiopathic pulmonary fibrosis. Thorax 66, 651–657 (2011).
13. Vervoort, S. J., Lourenço, A. R., van Boxtel, R. & Coffer, P. J. SOX4 mediates TGF-β-induced expression of mesenchymal markers during mammary cell epithelial to mesenchymal transition. PLoS ONE 8, e53238 (2013).
14. Györy, I. et al. Transcription factor Ebf1 regulates differentiation stage-specific signaling, proliferation, and survival of B cells. Genes Dev. 26, 668–682 (2012).
15. Micalizzi, D. S. et al. The Six1 homeoprotein induces human mammary carcinoma cells to undergo epithelial-mesenchymal transition and metastasis in mice through increasing TGF-beta signaling. J. Clin. Invest. 119, 2678–2690 (2009).
16. Liu, Y., Chu, A., Chakroun, I., Islam, U. & Blais, A. Cooperation between myogenic regulatory factors and SIX family transcription factors is important for myoblast differentiation. Nucleic Acids Res 38, 6857–6871 (2010).
17. Eckert, D., Buhl, S., Weber, S., Jäger, R. & Schorle, H. The AP-2 family of transcription factors. Genome Biol. 6, 246 (2005).
18. Luchsinger, L. L., Patenaude, C. A., Smith, B. D. & Layne, M. D. Myocardin-related transcription factor-A complexes activate type I collagen expression in lung fibroblasts. J. Biol. Chem. 286, 44116–44125 (2011).
19. Murata, K. et al. p63 - Key molecule in the early phase of epithelial abnormality in idiopathic pulmonary fibrosis. Exp. Mol. Pathol. 83, 367–376 (2007).
20. Sho, T. et al. TRIM29 negatively regulates p53 via inhibition of Tip60. Biochim. Biophys. Acta 1813, 1245–1253 (2011).

21. Lee, K. S., Buck, M., Houglum, K. & Chojkier, M. Activation of hepatic stellate cells by TGF alpha and collagen type I is mediated by oxidative stress through c-myb expression. J. Clin. Invest. 96, 2461–2468 (1995).
22. Yu, X., Si, J., Zhang, Y. & Dewille, J. W. CCAAT/Enhancer Binding Protein-delta (C/EBP-delta) regulates cell growth, migration and differentiation. Cancer Cell Int. 10, 48 (2010).
23. Nakamura, T. et al. Axin, an inhibitor of the Wnt signalling pathway, interacts with beta-catenin, GSK-3beta and APC and reduces the beta-catenin level. Genes Cells 3, 395–403 (1998).
24. Königshoff, M. et al. Functional Wnt signaling is increased in idiopathic pulmonary fibrosis. PLoS ONE 3, e2142 (2008).
25. Rajasekaran, S., Vaz, M. & Reddy, S. P. Fra-1/AP-1 transcription factor negatively regulates pulmonary fibrosis in vivo. PLoS ONE 7, e41611 (2012).
26. Li, Q. F., Wang, X. R., Yang, Y. W. & Lin, H. Hypoxia upregulates hypoxia inducible factor (HIF)-3alpha expression in lung epithelial cells: characterization and comparison with HIF-1alpha. Cell Res. 16, 548–558 (2006).
27. Yin, Z. et al. Hop functions downstream of Nkx2.1 and GATA6 to mediate HDAC-dependent negative regulation of pulmonary gene expression. Am. J. Physiol. Lung Cell Mol. Physiol. 291, L191-199 (2006).
28. Myllärniemi, M. et al. Gremlin-mediated decrease in bone morphogenetic protein signaling promotes pulmonary fibrosis. Am. J. Respir. Crit. Care Med. 177, 321–329 (2008).
29. Patel, N. M. et al. Pulmonary arteriole gene expression signature in idiopathic pulmonary fibrosis. Eur. Respir. J. 41, 1324–1330 (2013).
30. Lu, Y. et al. Kruppel-like factor 15 regulates smooth muscle response to vascular injury--

brief report. Arterioscler. Thromb. Vasc. Biol. 30, 1550–1552 (2010).

31. Crooks, M. G., Fahim, A., Naseem, K. M., Morice, A. H. & Hart, S. P. Increased platelet reactivity in idiopathic pulmonary fibrosis is mediated by a plasma factor. PLoS ONE 9, e111347 (2014).
32. Bai, S., Shi, X., Yang, X. & Cao, X. Smad6 as a transcriptional corepressor. J. Biol. Chem. 275, 8267–8270 (2000).
33. Bitoun, E., Finelli, M. J., Oliver, P. L., Lee, S. & Davies, K. E. AF4 is a critical regulator of the IGF-1 signaling pathway during Purkinje cell development. J. Neurosci. 29, 15366–15374 (2009).

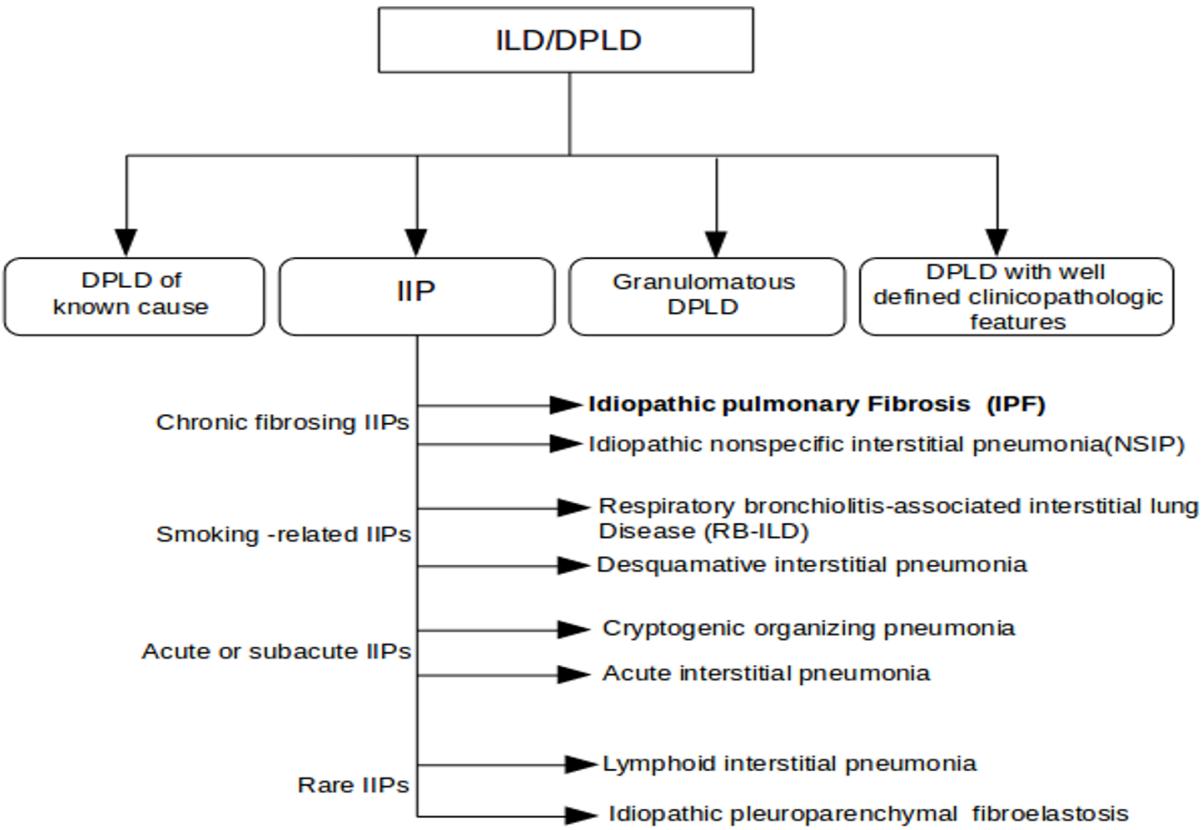

Figure S1: Classification of Interstitial lung diseases (ILD)

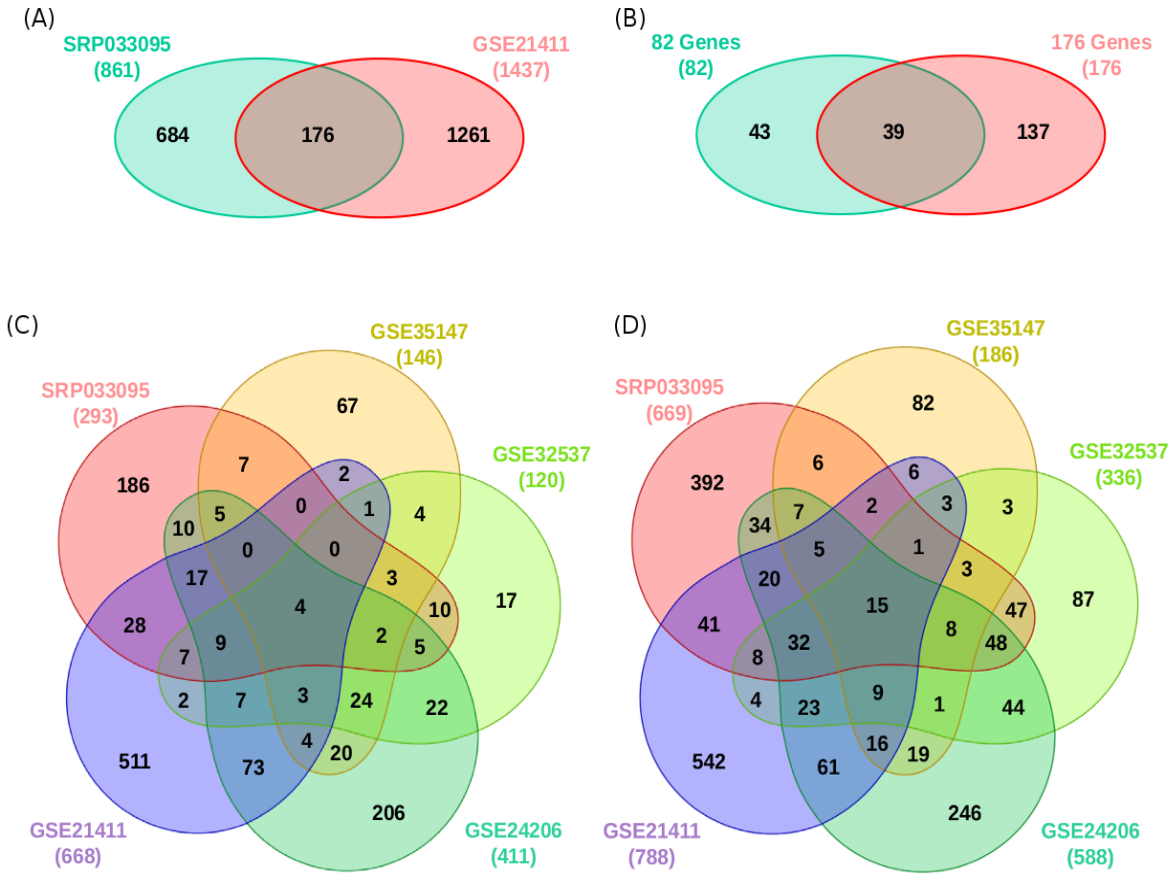

Figure S2: Overlap of DEGs among different experiments **(A)** SRP033095 and GSE21411, **(B)** previously reported 82 genes and 176 genes from our study, **(C)** Down regulated genes of GSE35147, GSE32537, GSE21411, GSE24206 and SRP033095, **(D)** Up regulated genes of GSE35147, GSE32537, GSE21411, GSE24206 and SRP033095

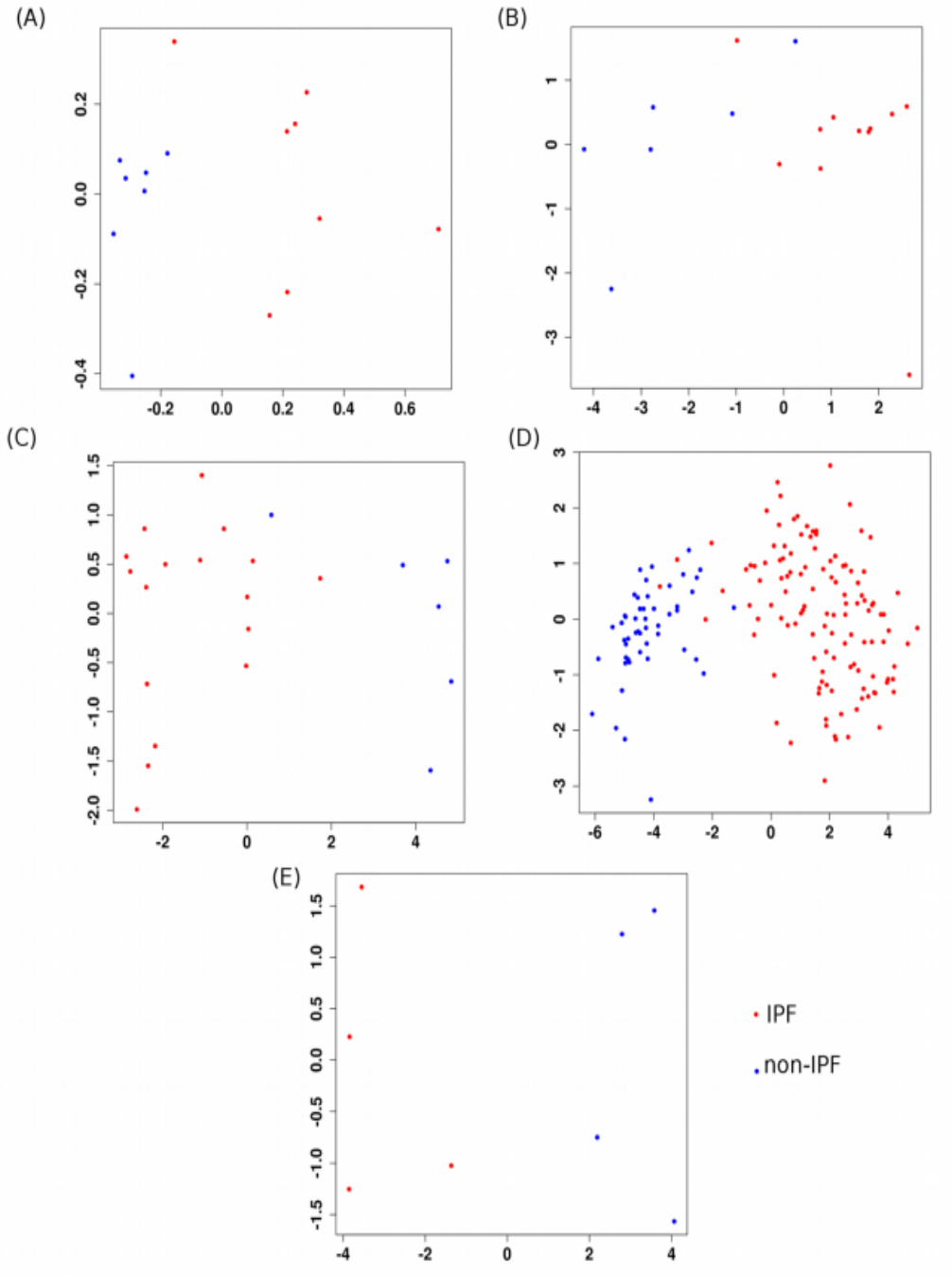

Figure S3: Clustering results of Set S1 for different experiments **(A)** SRP033095, **(B)** GSE21411, **(C)** GSE24206, **(D)** GSE32537, **(E)** GSE35147

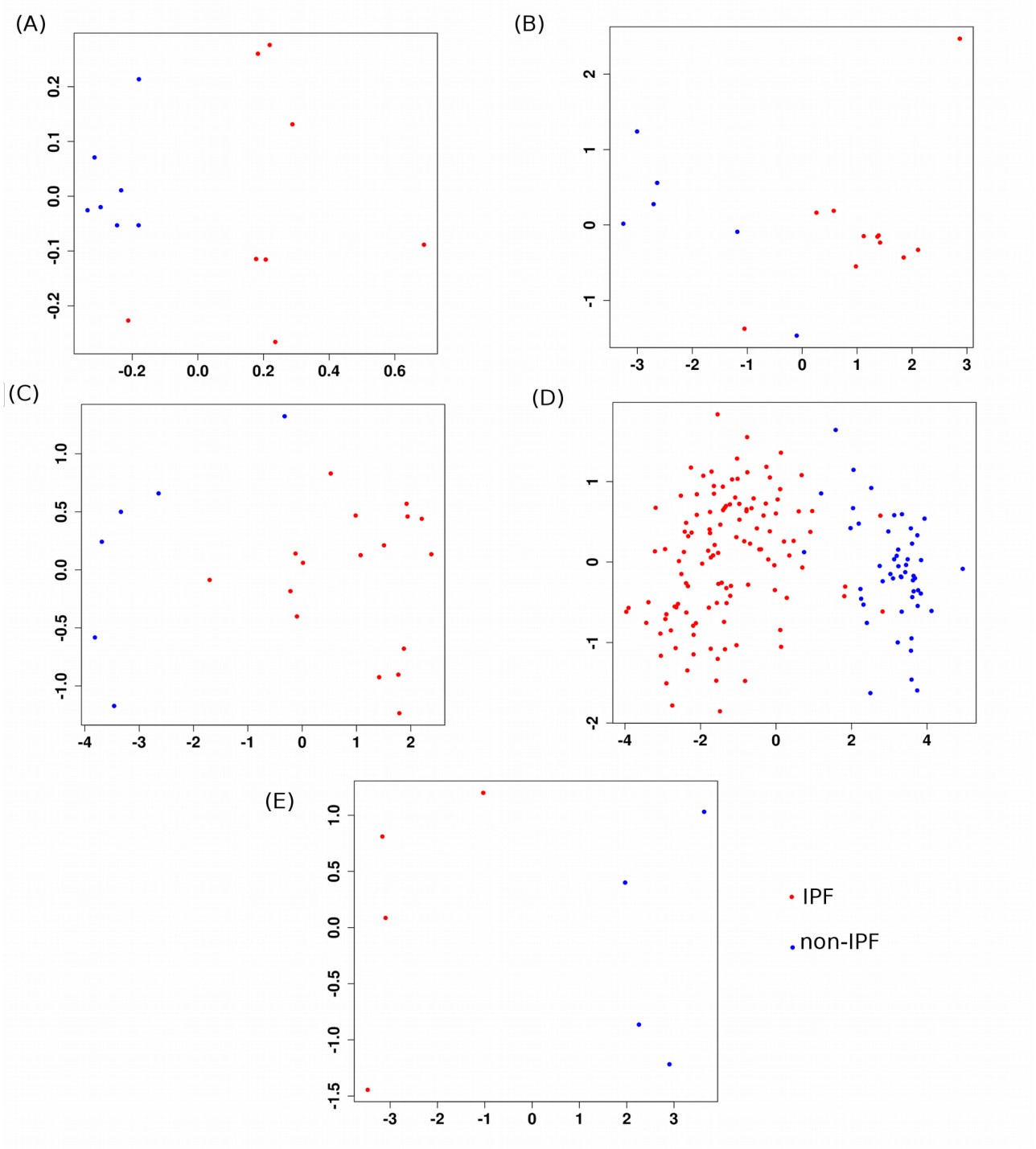

Figure S4: Clustering results of Set A for different experiments **(A)** SRP033095, **(B)** GSE21411, **(C)** GSE24206, **(D)** GSE32537, **(E)** GSE35147

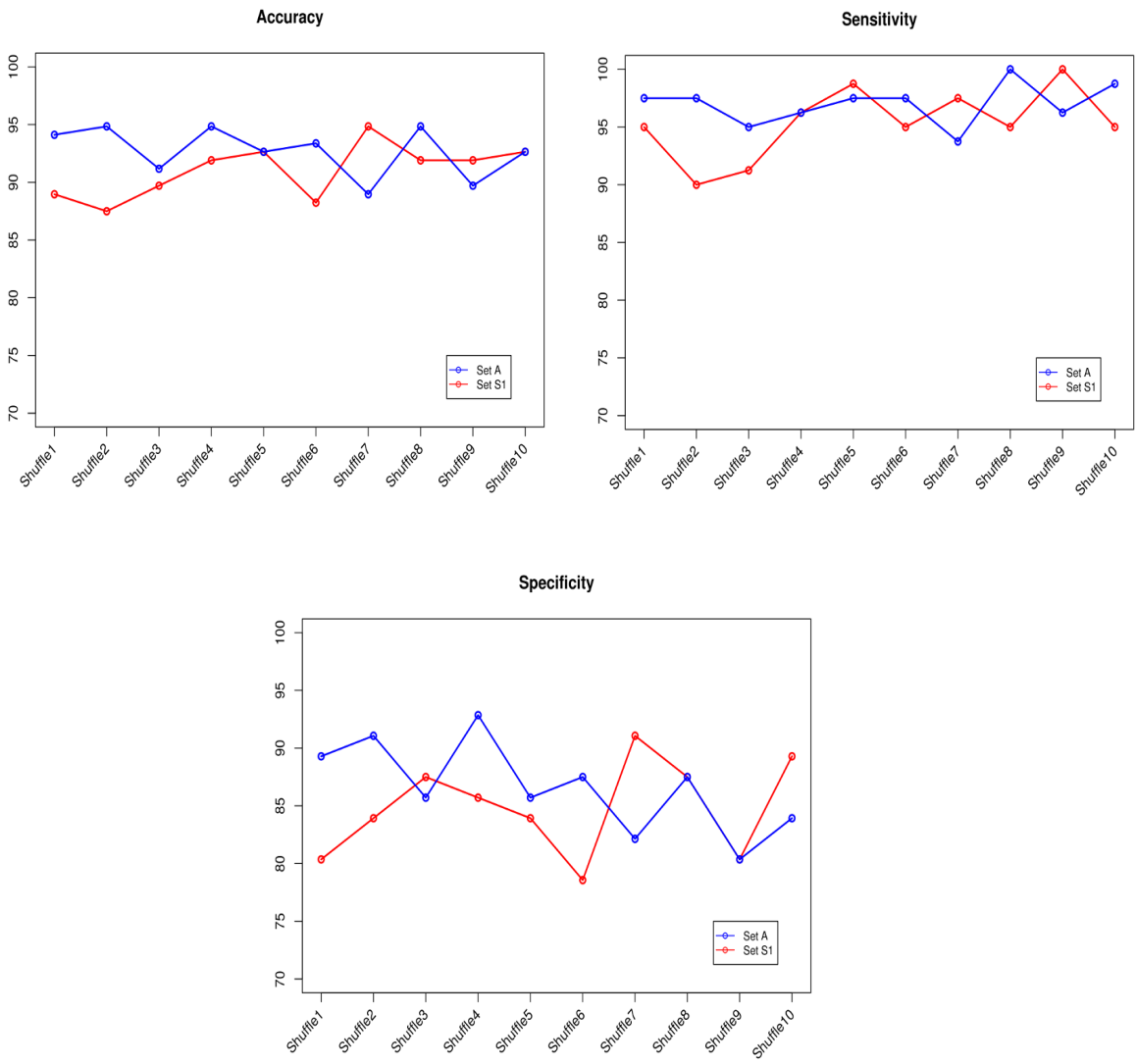

Figure S5: Accuracy, Sensitivity and Specificity plot for Set A and Set S1 on 10 randomized dataset models

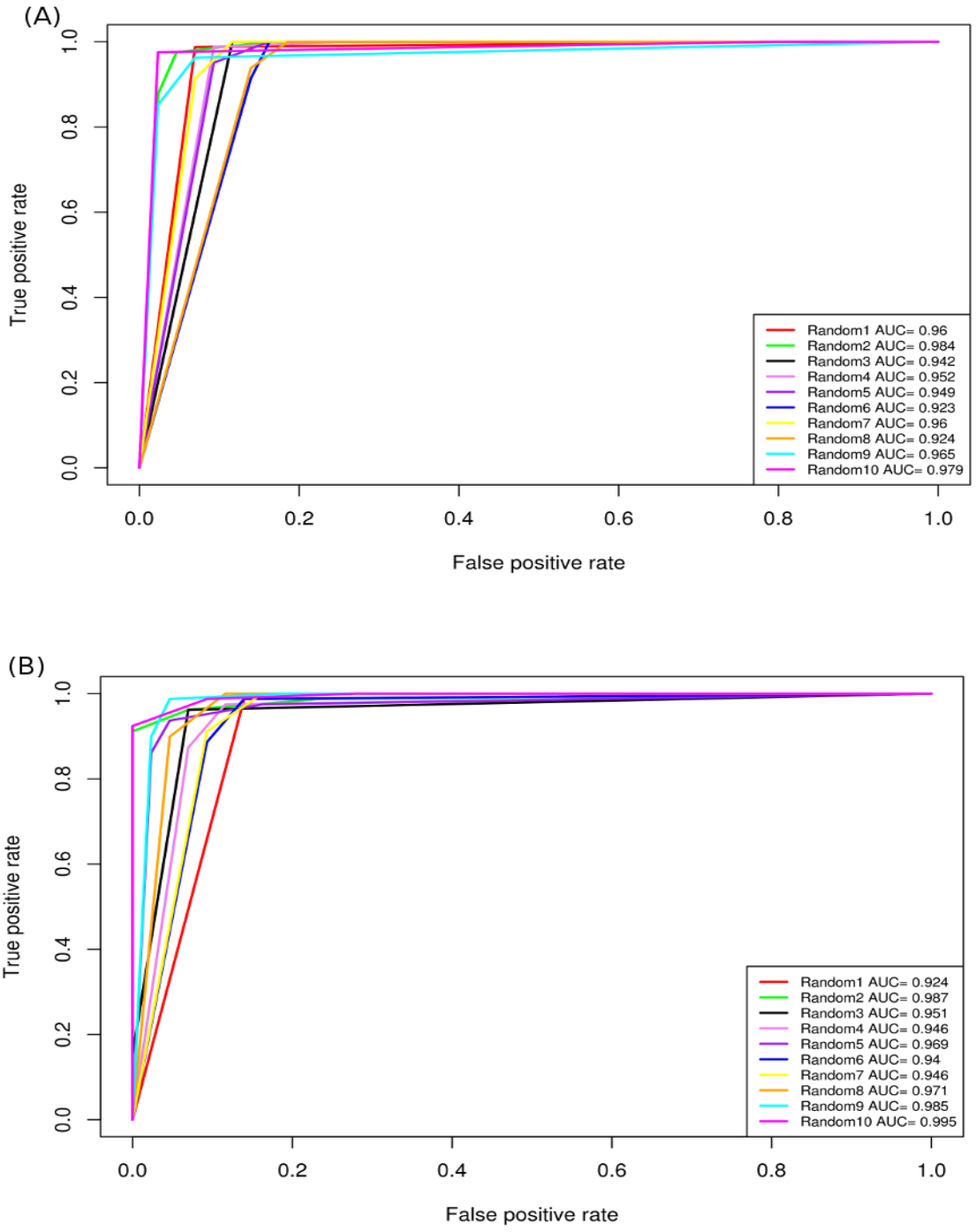

Figure S6: ROC plots of two gene sets against ten random models **(A)** Set A **(B)** Set S1

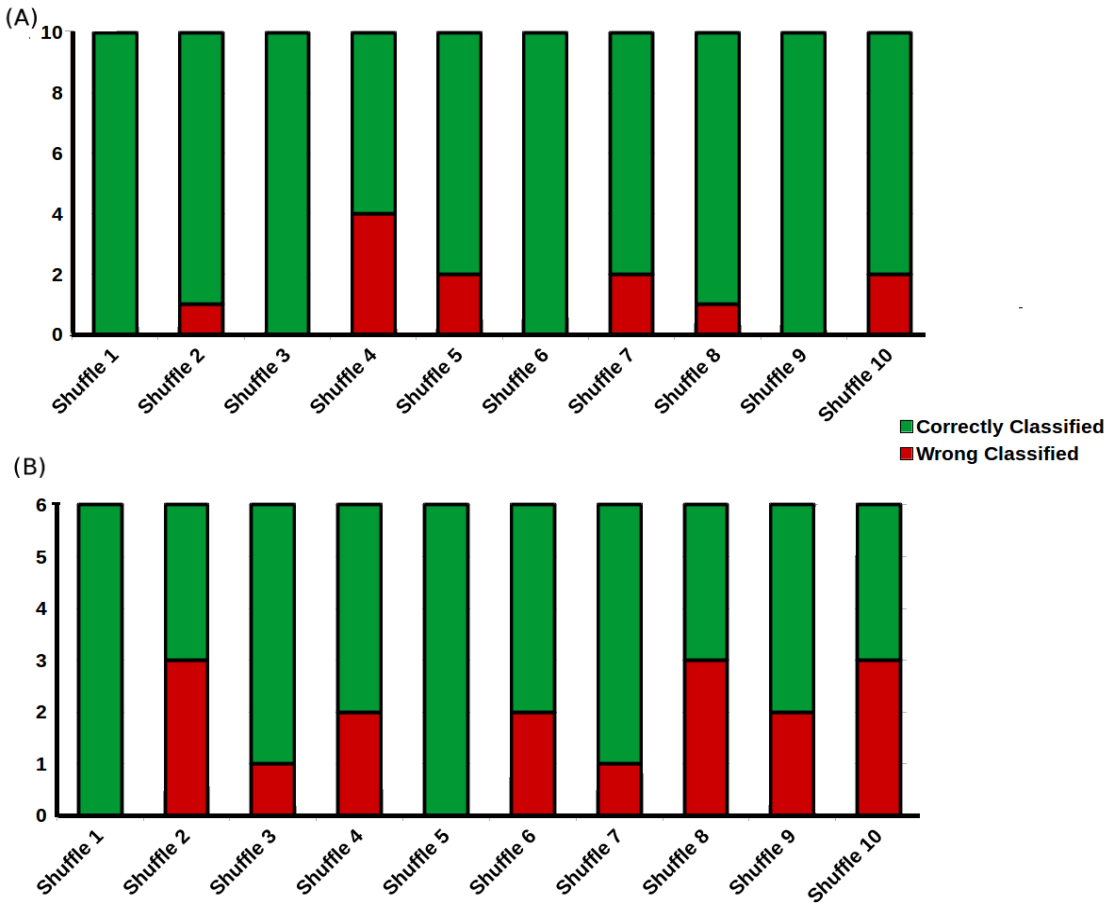

Figure S7: NSIP and Sarcoidosis individauls classification on 10 different models **(A)** Sarcoidosis (10 Individuals) **(B)** NSIP (6 Individuals)

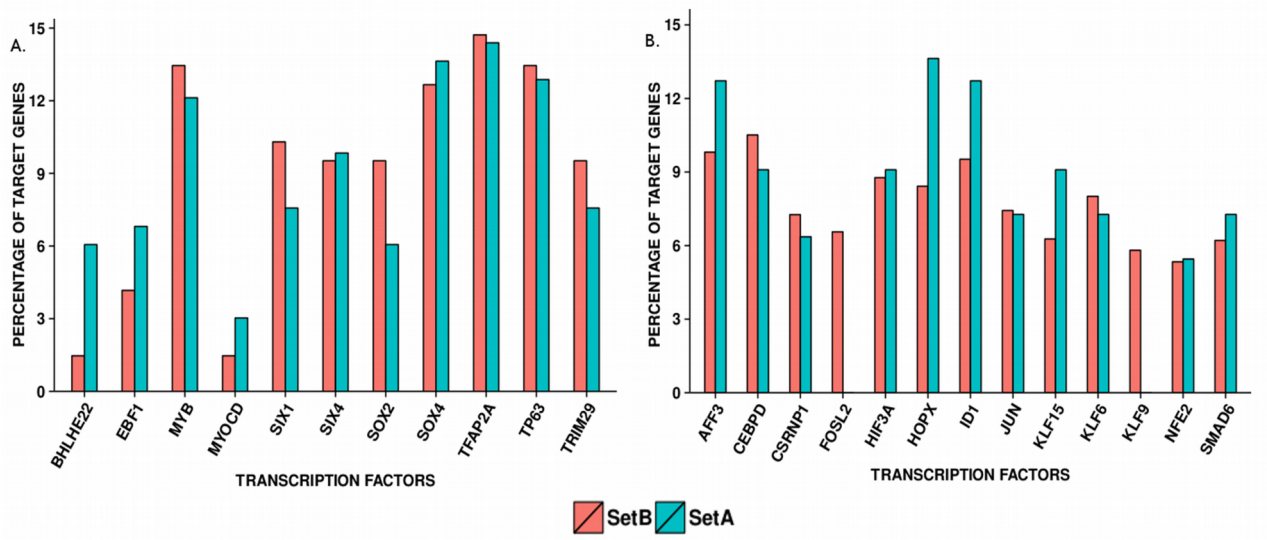

Figure S8: (A) Ranking of differentially expressed upregulated transcription factors regulating DEGs, **(B)** Ranking of differentially expressed downregulated transcription factors regulating DEGs specific to IPF

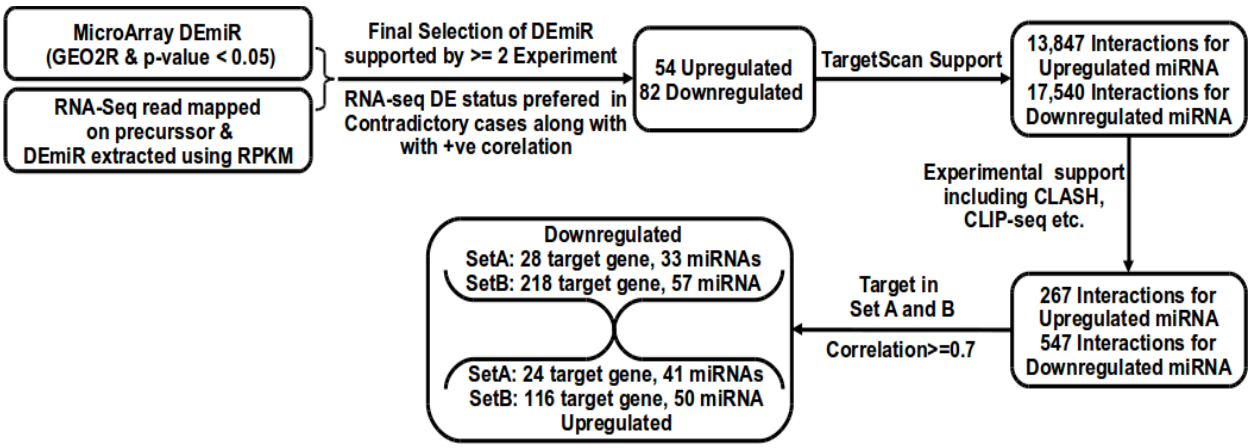

Figure S9: Workflow of miRNA analysis from microarray and RNA-seq study, miRNA target finding, support analysis and involvement of miRNA in different sets of genes

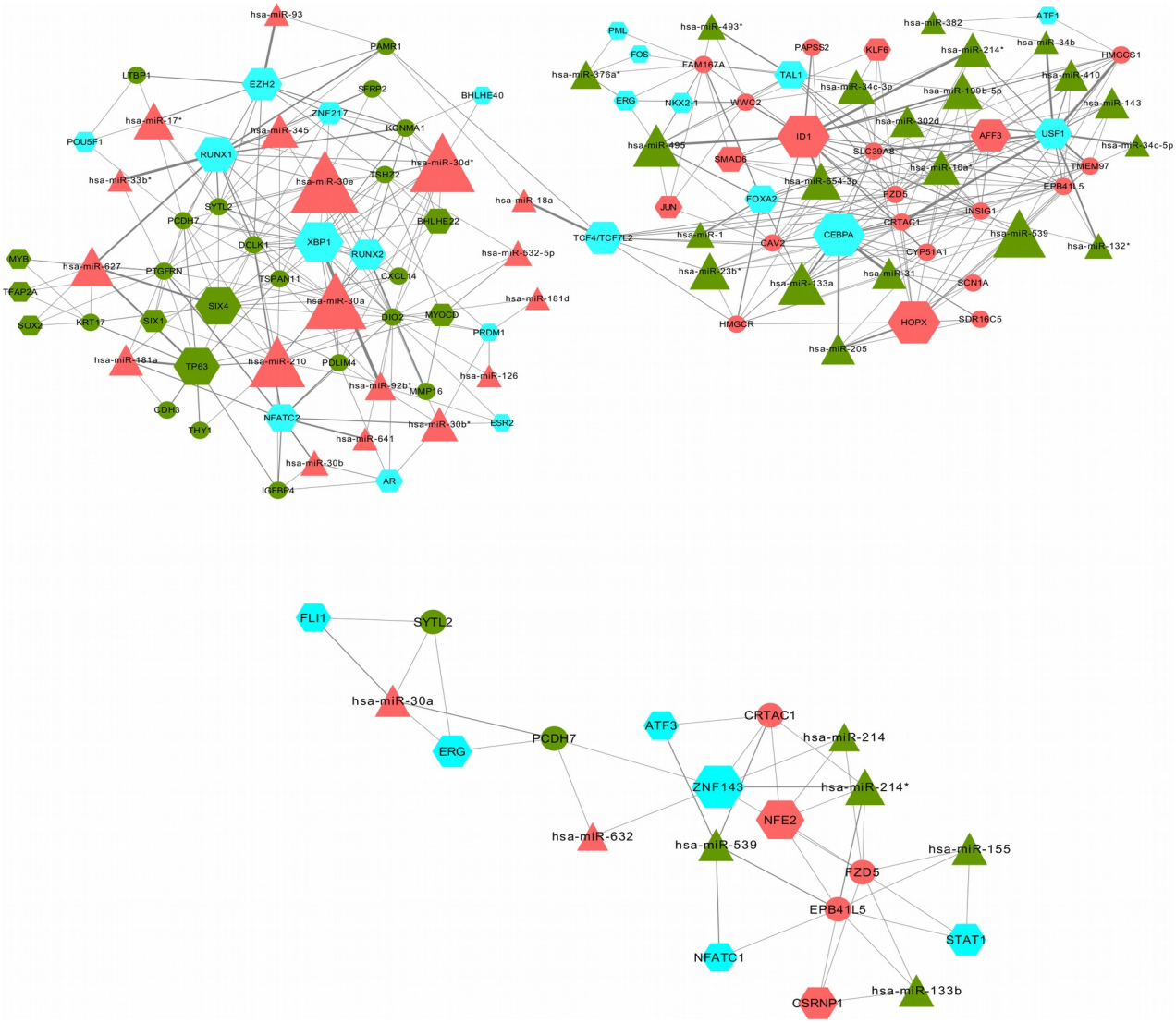

Figure S10: Graphical view of IPF regulatory networks generated for Set A DEGs representing important nodes in the form of IPF genes and miRNAs at each stage (down and up regulated genes and miRNAs are shown in red and green color, respectively). Edges are highlighted based on the edge betweenness and nodes size is adjusted based on outdegree. TFs are shown in sea green color and hexagon shape.

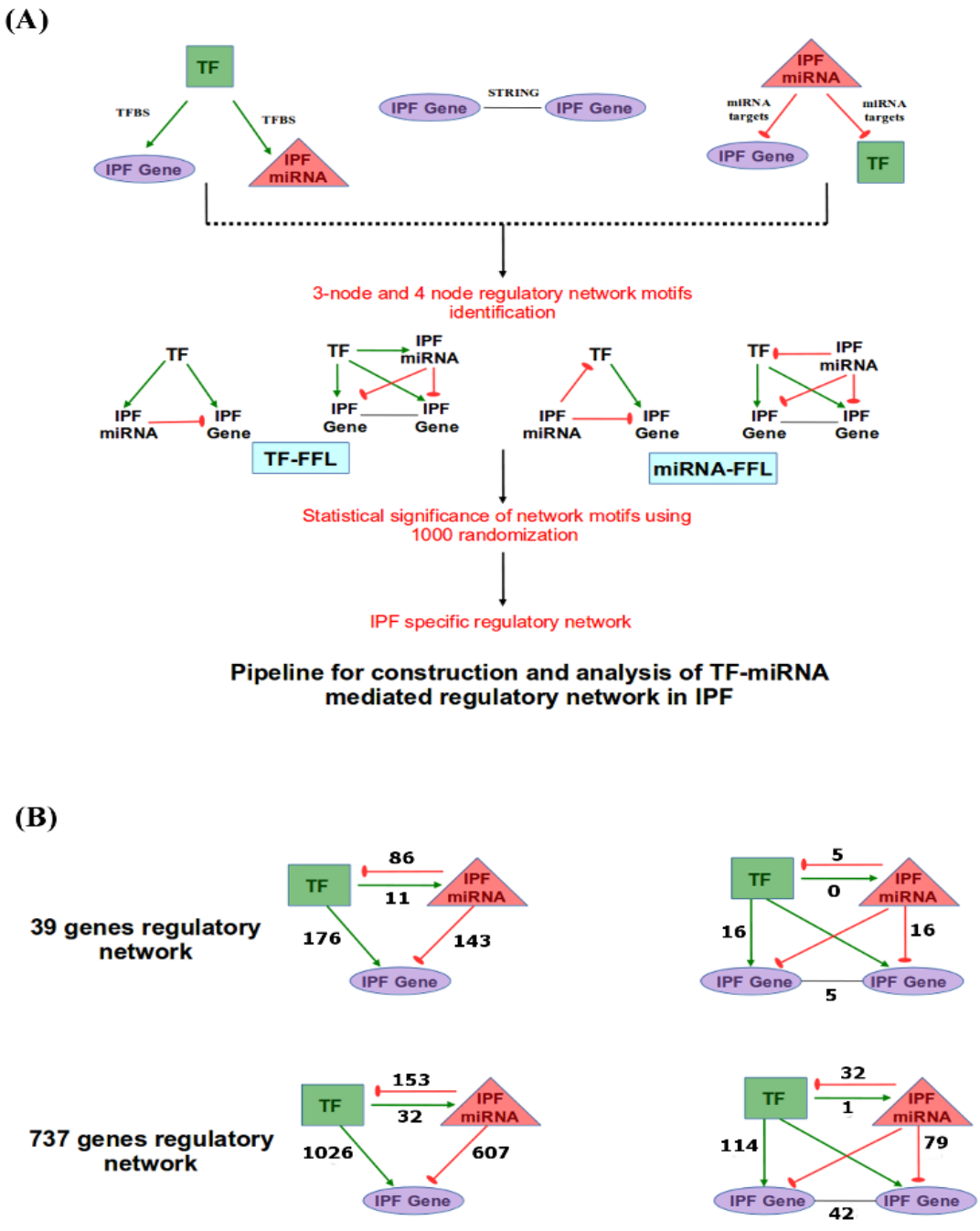

Figure S11: (A) Work flow for generation of various Feed-Forward loops and construction of TF-miRNA mediated regulatory network in IPF, **(B)** Statistics of three and four node FFLs for the three sets of DEGs in IPF.

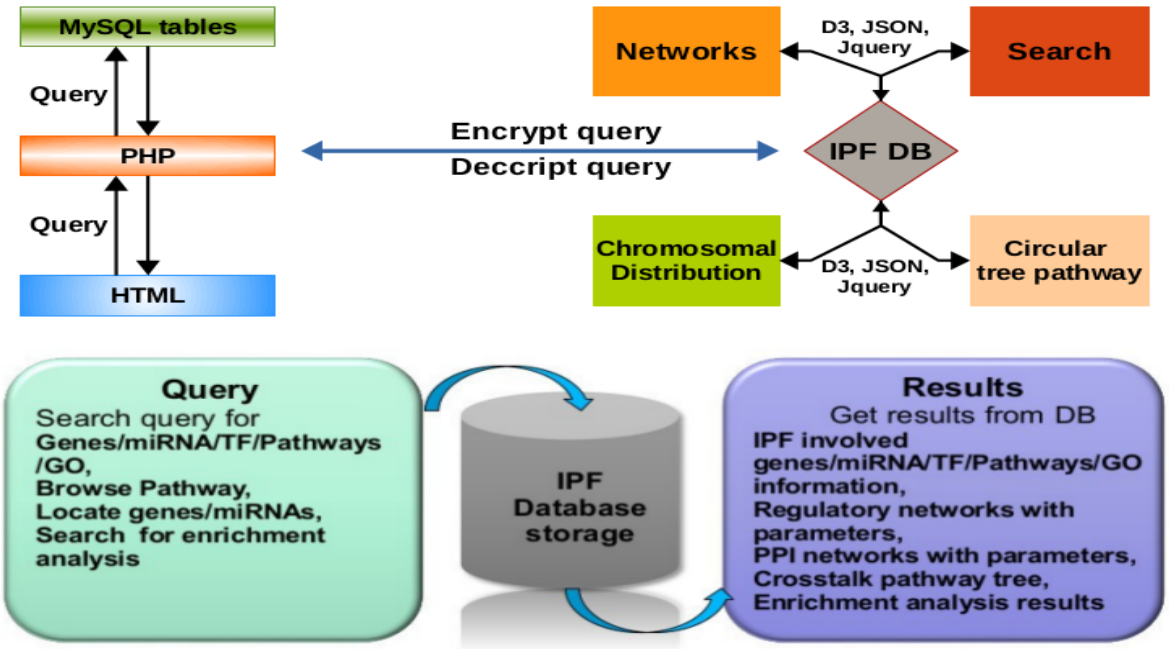

Figure S12: The framework and workflow of IPF portal