Interstitial Lung Disease Proteomic

Grace Malonga, Jingyi Liang, Xu Chen

Abstract

Interstitial lung disease (ILD) is a diverse group of diseases characterized by progressive scarring and inflammation of the lung. This study aimed to analyze the serum expression levels of the corresponding proteins and find protein signatures that differentiate ILD phenotypes. 160 patients who with or without ILD were enrolled to this study. Differential expression analysis identified that 58 proteins differentially expressed between RA ILD vs. RA no ILD, 35 proteins differentially expressed between IPF vs. Healthy group (|logFC|>=0.5, adjusted p<0.01). In addition, age distributions in groups of RA ILD vs. RA no ILD were different. After reducing the influence of age, differential expression analysis identified that 82 proteins differentially expressed in patients under 60 between RA ILD vs RA no ILD (|logFC|>=0.5, adjusted p<0.01). In conclusion, identify protein signatures that differentiate ILD phenotypes may can be used for serum biomarker assay to improve diagnostic accuracy for ILD in remote area.

Introduction

Interstitial lung diseases (ILD), also defined as diffuse parenchymal lung diseases (DPLD) include a heterogeneous group of pulmonary disorders. ILD may be caused by five major factors: (1) Environmental triggers such as prolonged exposure to asbestos, (2) Connective tissue diseases such as rheumatoid arthritis (RA), (3) Sarcoidosis, (4) Idiopathic interstitial pneumonias such as idiopathic pulmonary fibrosis (IPF), (5) and other triggers such as Langerhans cell histiocytosis [1-4]. Ailments generated by these factors contributes widely to morbidity and mortality; these conditions are serious, irreversible and progressive with poor prognosis [5, 6]. Because of the heterogeneity in ILD triggers and conditions, diagnosing this disorder is just as complex since subtype classification needs to be taken into account [7]. Currently, the gold standard for ILD diagnosis and subtype classification moved from open lung biopsy to high resolution CT (HRCT) scan [8, 9, 10]. ILD subtypes are often classified by the display of their phenotypes in terms shapes and sizes that their corresponding imaging scans may take [11, 12]. However, appropriately reading CT scans, identifying ILD and classifying ILD subtypes requires specialist expertise and training. Even among pulmonary specialists, many physicians are ill equipped to recognize and interpret subtle differences on imaging scans between ILD phenotypes, often leading to misdiagnosis and inappropriate treatment.

Development of an accurate serum biomarker assay to differentiate between ILD phenotypes could improve diagnostic accuracy for ILD and lead to better treatments, particularly for populations and in geographic areas where ILD specialists are inaccessible, and in situations where HCRT is limited [13]. The object of this project is to identify protein signatures that differentiate ILD phenotypes; the focus will be on the most clinically relevant comparisons such as: RA ILD vs. RA no ILD, IPF vs. healthy, RA ILD vs. IPF, RA fibrotic vs. RA cellular.

Methodology

This is an analysis of 160 patients with either idiopathic pulmonary fibrosis (IPF), or rheumatoid arthritis ILD (RA ILD), including patients with fibrotic nonspecific interstitial pneumonia (NSIP), cellular NSIP, and chronic organizing pneumonia (COP) patterns. The investigators also enrolled healthy volunteers and subjects with rheumatoid arthritis, but no ILD for comparison. The dataset included disease status, ILD subtype, personal information and lifestyle, lung function and transcription levels of 1321 proteins tested from blood sample of each patient.

All data manipulation and statistical analysis were done in R. Protein signature variables were initially normalized with a log2-transformation; to assess which proteins were differentially expressed between ILD phenotype, student's T-tests were performed using rowttest, a function from genefilter package, to compare ILD phenotypes across all protein biomarkers. However, because several tests are run on a small sample size, this type of analysis is prone to generate false positives, which is also known as 'type I error'; an error occuring when a positive result corresponds to rejecting the null hypothesis.

To further evaluate the expression of protein signatures between groups, a linear model was fitted on the data through the Bioconductor limma package to reduce the false positive rate.

Limma is a R/Bioconductor software package that provides an integrated solution for analysing data similar to gene expression experiments [14], such as microarrays, and protein arrays. First, we did several steps of preprocessing work: (1) Construct a design matrix to explain observed data on different protein biomarkers in terms of different ILD phenotypes. (3) Generate a contrast matrix to evaluate differential expression under two experimental conditions. In this case, the experimental conditions are the different ILD phenotypes: RA ILD vs. RA no ILD, IPF vs. healthy, RA ILD vs. IPF, RA fibrotic vs RA cellular.

In limma, we used the ‘lmFit’ function to fit the linear model and calculate the difference between groups according to the contrast matrix. Based on the fitting result, an empirical Bayes procedure was applied to compute moderate t-statistics, and log-odds of differential expression. The results were summarized with the toptable function in which we performed type I error adjustment by selecting the False Discovery Rate (FDR) method. FDR-controlling procedures are designed to control the expected proportion of "discoveries" (rejected null hypotheses) that are false (incorrect rejections) [15]. For a robust adjustment of false positives, a threshold on LogFoldChange was also set; this method is often used in analyses of gene expression data from microarray and RNA-Seq experiments for measuring change in the expression level of a gene [16]. In this case, absolute log2 fold change (log2FC) cut-offs were set to > 0.5, which equaled to absolute fold change cuts-off > 1.414; also only p-values of <0.01 were considered as significant. These cut-offs force the statistical model to look only at proteins biomarkers that vary wildly amongst other proteins signatures.

After identifying what proteins were differentially expressed between ILD phenotypes, the next step was to perform a multivariate analysis to adjust for significant covariates that can introduce confounding effects in the linear model. To identify covariates to control in the statistical model, differences in demographic, body size, or lung function variables were evaluated between ILD phenotype comparison groups. For normally distributed variables, T-test or Fisher’s Exact test was used; then confirmed with Shapiro-Wilk Normality Test. For non-normal variables, we applied Wilcoxon Rank Sum test. A Chi-squared test was used for categorical variables. For the multivariable model we planned to categorize continuous variables.

Patients were grouped into different subgroups to find out what proteins are differentially expressed between ILD phenotypes. Age was stratified following a new proposed ILD classification in 2013[17], we divided the patients into two age groups, 60 years old and older versus 60 years old and younger.

Results

From T-tests generated with rowttest, we had 309 differentially expressed proteins between RA ILD vs. RA no ILD, 197 proteins between IPF and healthy, 31 proteins between IPF and RA ILD, and 6 proteins were differentially expressed between Fibrotic NSIP and Cellular NSIP (See detailed test results in Figure 2 and Table 1, Appendix). Because of the large amount of false positives this test generated, we used the lmFit function in Limma to adjust for the type I error.

After reducing the false positive rate, 58 proteins differentially expressed between RA ILD vs. RA no ILD, 35 proteins differentially expressed between IPF vs. Healthy, no proteins differentially expressed between RA ILD vs. IPF and RA fibrotic vs RA cellular. (See detailed test results in Figure 3 and Table 2,3, Appendix).

To further assess proteins that differentially expressed between groups, we decided to perform a multivariable analysis for RA ILD vs. RA no ILD, IPF vs. Healthy, RA ILD vs. IPF and RA fibrotic vs. RA cellular groups to control for significant covariates. The RA no ILD and RA ILD groups had significant differences in Age (P-Value<0.001), FVCPP (P-Value=0.0058), and Emphysema (P-Value<0.001). There was no significant difference in Gender, BMI, EverSmoker, PkYr, FEV1, FEV1PP, FVC, RatioPP, TLC, TLCPP, X6MWD. (See detailed distribution in Figure 4, detailed test results in Table 4, Appendix).

Because IPF vs. Healthy and RA ILD vs. IPF did not have any significant covariates, and because no proteins were differentially expressed in RA fibrotic vs RA cellular groups after adjusting type I error, we only focused on RA ILD vs. RA no ILD.

We decided to adjust for age to eliminate the influence of covariates, and found that between RA no ILD and RA ILD, those who were 60 years old and younger had 82 differentially expressed proteins compared to people who are over 60 years old. (See detailed test results in Figure 5 and Table 5, Appendix).

Conclusion

The serum biomarker assay was effective in identifying relevant sets of proteins signatures attributed to the presence of RA ILD, RA no ILD, IPF, and Healthy status. Similar results were noted in a study of Daniel J. Kass et al. in which Multiplex ELISA, a biomarker assay, had successfully identified protein biomarkers related to RA ILD vs. RA no ILD, and IPF [18]. However, no differentially expressed protein biomarkers were identifiable between RA ILD and IPF, as well as between RA fibrotic and RA cellular. The variable Age was not the only significant covariate; however due to the large amount of missing data in demographic and lung functions variables, controlling for all potential confounders might have led to overfitting the statistical model.

Thus, more studies appropriately designed to assess biomarkers performance in identifying ILD diagnosis, predicting ILD prognosis and classifying ILD subtypes are in great need to not only evaluate whether our statistical analysis is reproducible, but also to confirm biomarker assays efficiency in determining ILD diagnosis and prognosis.

References

[1]Collard HR, King TE Jr. Demystifying idiopathic interstitial pneumonia. Arch Intern Med 2003; 163:17 29.

[2]Bradley B, Branley HM, Egan JJ, et al. Interstitial lung disease guideline: the British Thoracic Society in collaboration with the Thoracic Society of Australia and New Zealand and the Irish Thoracic Society. Thorax 2008; 63 (Suppl 5):v1-v58.

[3]Baughman RP, Culver DA, Judson MA. A concise review of pulmonary sarcoidosis. Am J Respir Crit Care Med 2011; 183:573-581.

[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]Rheumatoid arthritis-related interstitial lung disease (RA-ILD): methotrexate and the severity of lung disease are associated to prognosis. Rojas-Serrano J et al. Clin Rheumatol. (2017)

[6]Kim EJ, Elicker BM, Maldonado F, Webb WR, Ryu JH, Van Uden JH, et al. Usual interstitial pneumonia in rheumatoid arthritis-associated interstitial lung disease. Eur Respir J. 2010;35(6):1322–8.

[7]Update in diagnosis and management of interstitial lung disease . Mikolasch TA, Garthwaite HS, Porter JC. Clin Med (Lond). 2017 Apr;17(2):146-153.

[8]Flaherty KR, Mumford JA, Murray S, Kazerooni EA, Gross BH, Colby TV, et al. Prognostic implications of physiologic and radiographic changes in idiopathic interstitial pneumonia. Am J RespirCritCareMed.2003

[9]Hunninghake GW, Zimmerman MB, Schwartz DA, King TE Jr, Lynch J, Hegele R, et al. Utility of a lung biopsy for the diagnosis of idiopathic pulmonary fibrosis. Am J Respir Crit Care Med. 2001;164(2):193–6.

[10]Kim EJ, Collard HR, King TE Jr. Rheumatoid arthritis-associated interstitial lung disease: the relevance of histopathologic and radiographic pattern. Chest. 2009;136(5):1397–405.

[11]High-resolution computed tomography of the lung in patients with rheumatoid arthritis: Prevalence of interstitial lung disease involvement and determinants of abnormalities Fausto Salaffi, Marina Carotti, Marco Di Carlo, Marika Tardella, Andrea Giovagnoni. Medicine (Baltimore) 2019 Sep; 98(38): e17088. Published online 2019 Sep 20.

[12]Interpretation of HRCT Scans in the Diagnosis of IPF: Improving Communication Between Pulmonologists and Radiologists.Chung JH et al. Lung. (2018)

[13]Prognostic implications of physiologic and radiographic changes in idiopathic interstitial pneumonia. Flaherty KR, Mumford JA, Murray S, Kazerooni EA, Gross BH, Colby TV, Travis WD, Flint A, Toews GB, Lynch JP 3rd, Martinez FJ. Am J Respir Crit Care Med. 2003 Sep 1; 168(5):543-8.

[14]limma powers differential expression analyses for RNA-sequencing and microarray studies Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, Gordon K. Smyth Nucleic Acids Research, Volume 43, Issue 7, 20 April 2015, Page e47.

[15]Benjamini, Yoav; Hochberg, Yosef (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing" (PDF). Journal of the Royal Statistical Society, Series B. 57 (1): 289–300. MR 1325392.

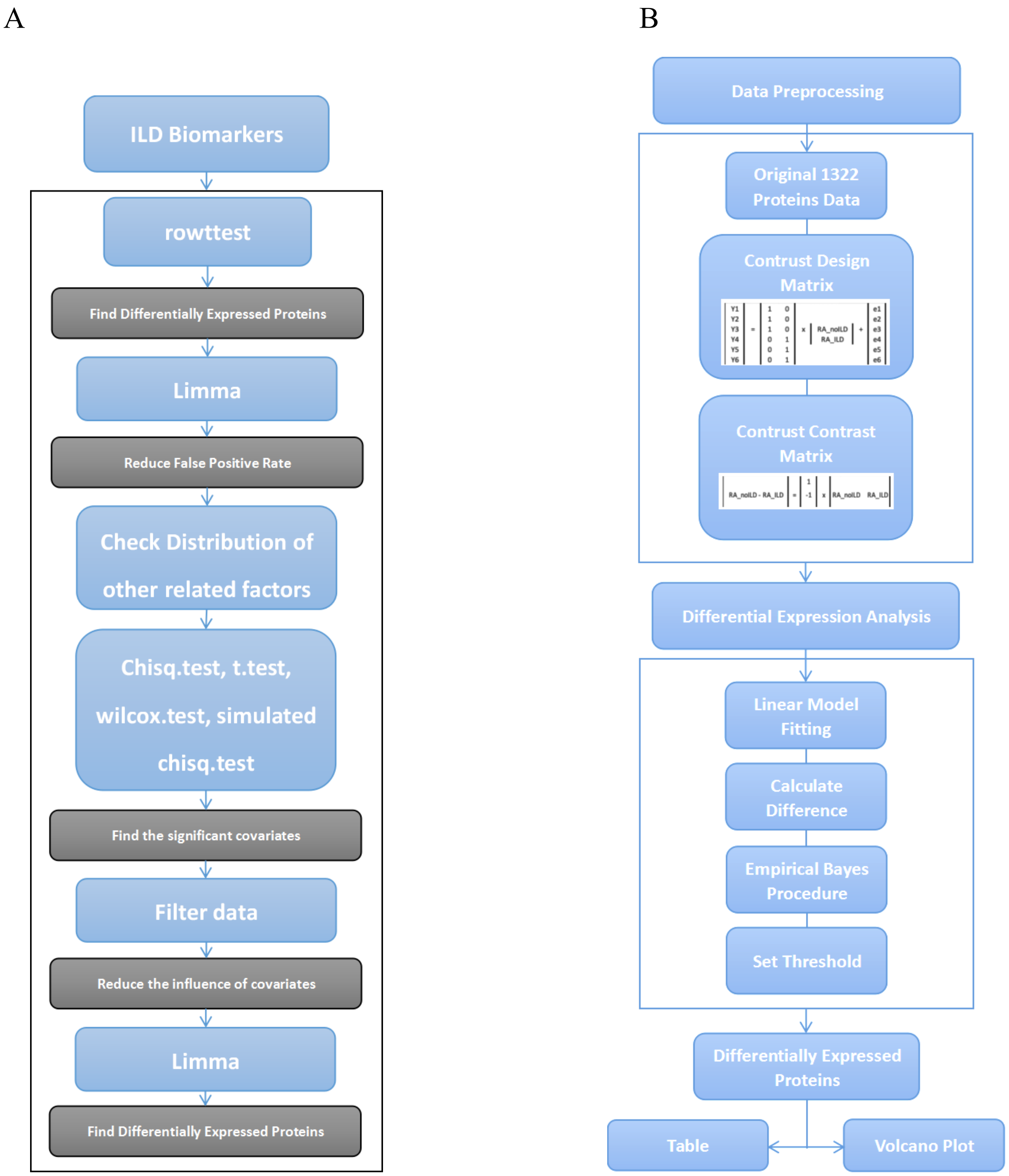
[16]A comparison of fold-change and the t-statistic for microarray data analysis.

[17]Travis WD, Costabel U, Hansell DM, et al. An official American Thoracic Society/European Respiratory Society statement: Update of the international multidisciplinary classification of the idiopathic interstitial pneumonias. Am J Respir Crit Care Med 2013; 188:733-48. 10.1164/rccm.201308-1483ST

[18]Comparative Profiling of Serum Protein Biomarkers in Rheumatoid Arthritis-associated Interstitial Lung Disease and Idiopathic Pulmonary Fibrosis. Kass DJ, Nouraie M, Glassberg MK, Ramreddy N, Fernandez K, Harlow L, Zhang Y, Chen J, Kerr GS, Reimold AM, England BR, Mikuls TR, Gibson KF, Dellaripa PF, Rosas IO, Oddis CV, Ascherman DP. Arthritis Rheumatol. 2019 Sep 18.

Appendix

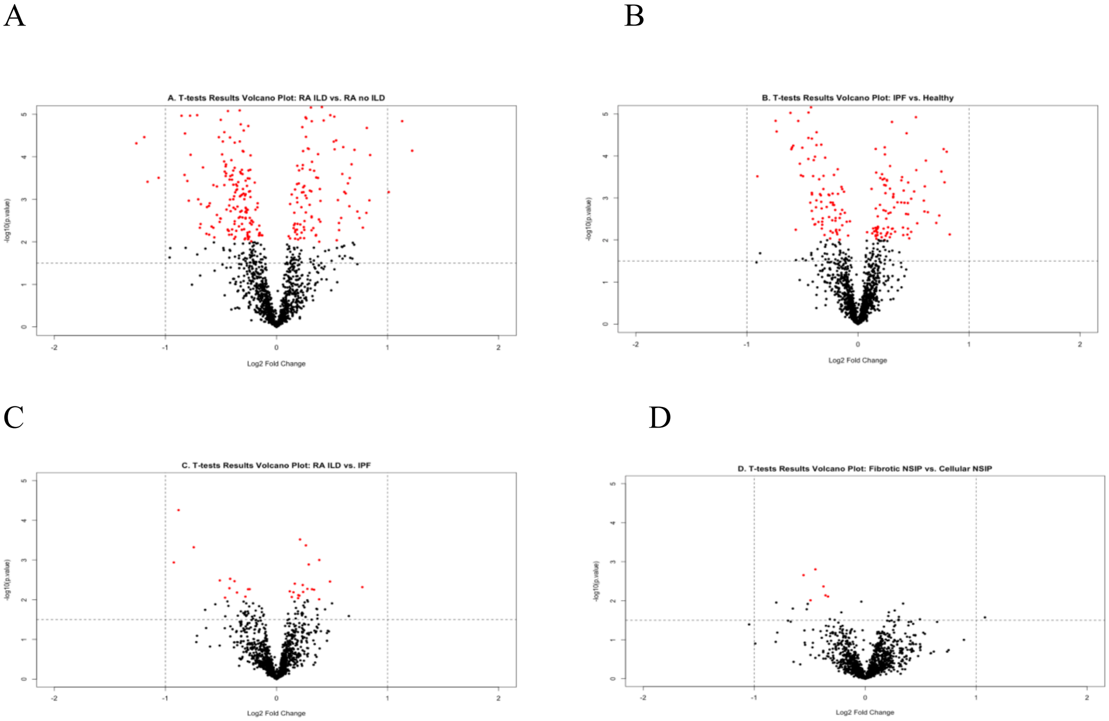
Figure 1


A
ILD Biomarkers
rowttest
Find Differentially Expressed Proteins
Limma
Reduce False Positive Rate
Check Distribution of other related factors
Chisq.test, t.test, wilcox.test, simulated chisq.test
Find the significant covariates
Filter data
Reduce the influence of covariates
Limma
Find Differentially Expressed Proteins
B
Data Preprocessing
Original 1322 Proteins Data
Contrust Design Matrix
Y1 Y2 Y3 Y4 Y5 Y6 = 1 0 1 0 1 0 0 1 0 1 0 1 x RA_noILD RA_ILD + e1 e2 e3 e4 e5 e6
Contrust Contrast Matrix
RA_noILD - RA_ILD = 1 -1 x RA_noILD RA_ILD
Differential Expression Analysis
Linear Model Fitting
Calculate Difference
Empirical Bayes Procedure
Set Threshold
Differentially Expressed Proteins
Table
Volcano Plot

Flow chart of analysis.

(1-A) is the flow chart of whole analysis, (1-B) is the flow chart of differential expression analysis by limma.

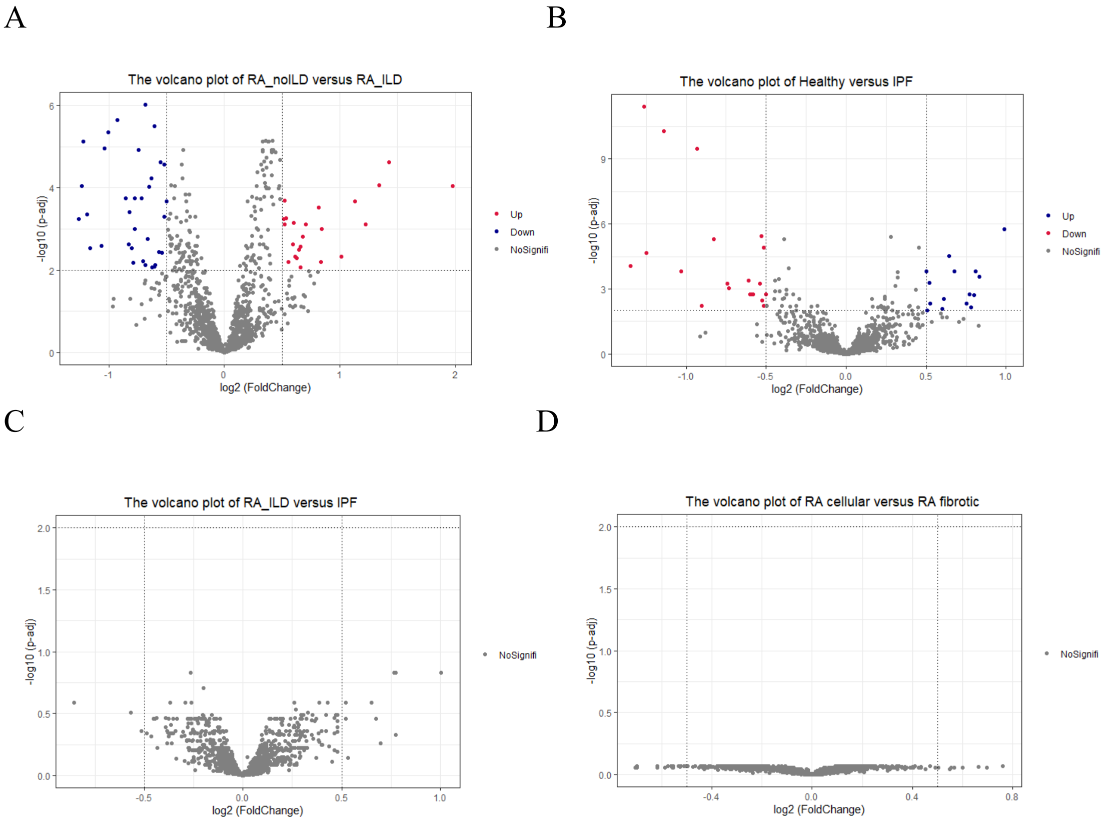
Figure 2



Volcano plots of different expression proteins between groups.

(2-A) is the volcano plot of RA no ILD vs. RA ILD (RA no ILD is the baseline), (2-B) is the volcano plot of Healthy vs. IPF (Healthy is the baseline), (2-C) is the volcano plot of RA ILD vs. IPF (RA ILD is the baseline), (2-D) is the volcano plot of RA cellular vs. RA fibrotic (RA cellular is the baseline). The black dots stand for the proteins which are not significantly different between groups. The red dots stand for the proteins which are significantly different between groups.

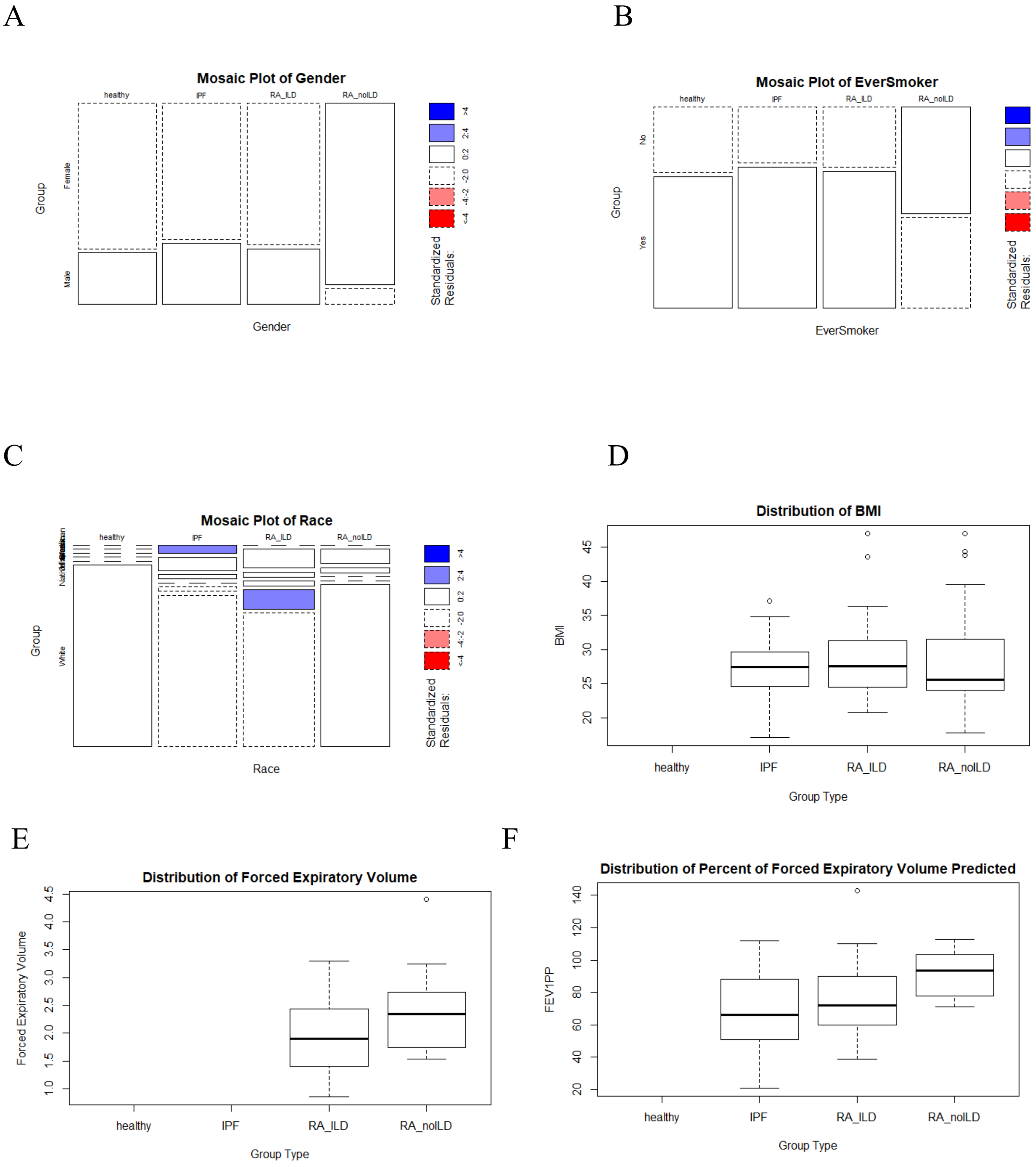
Figure 3



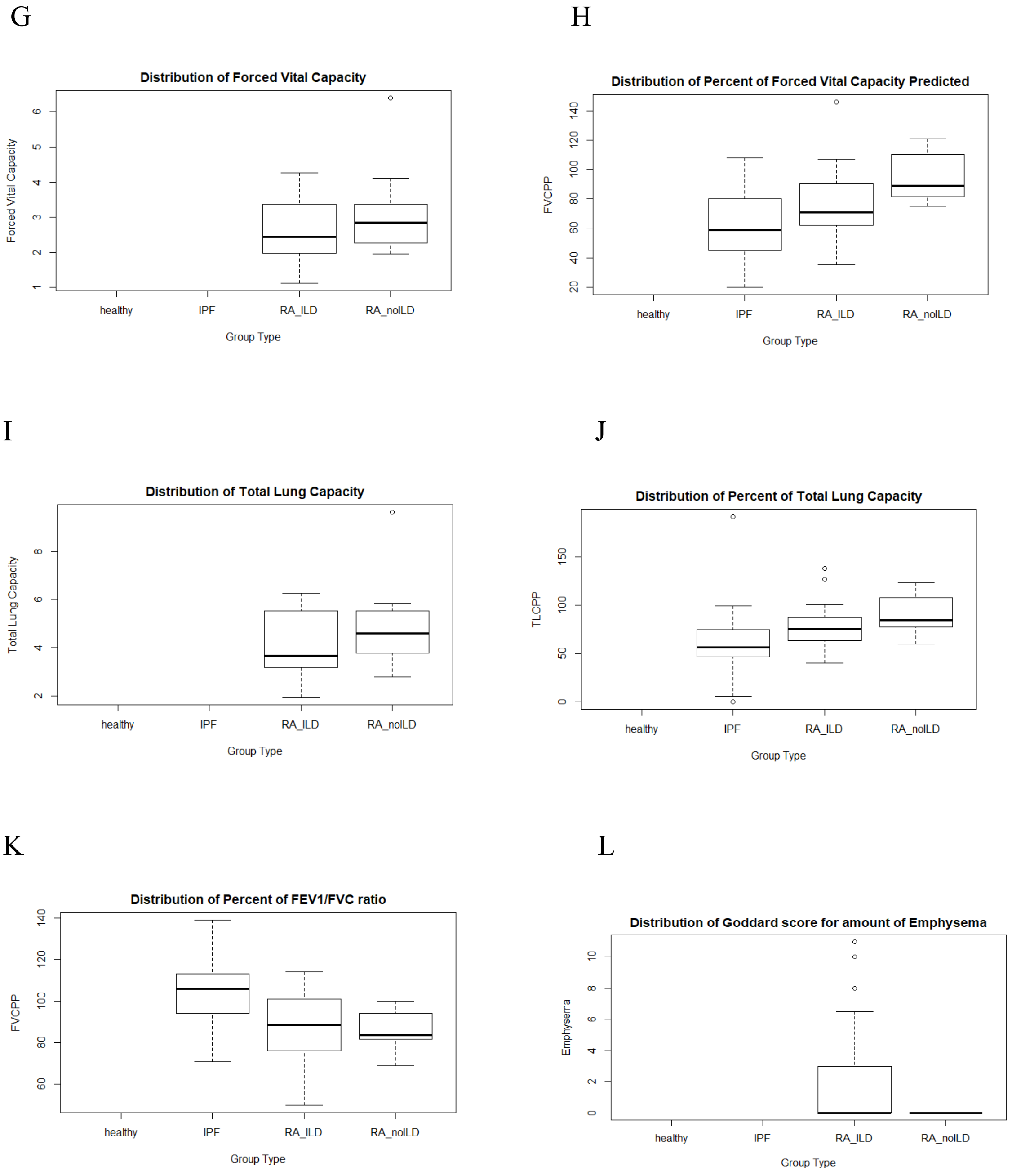
Volcano plots of different expression proteins between groups after reducing false positive rate.

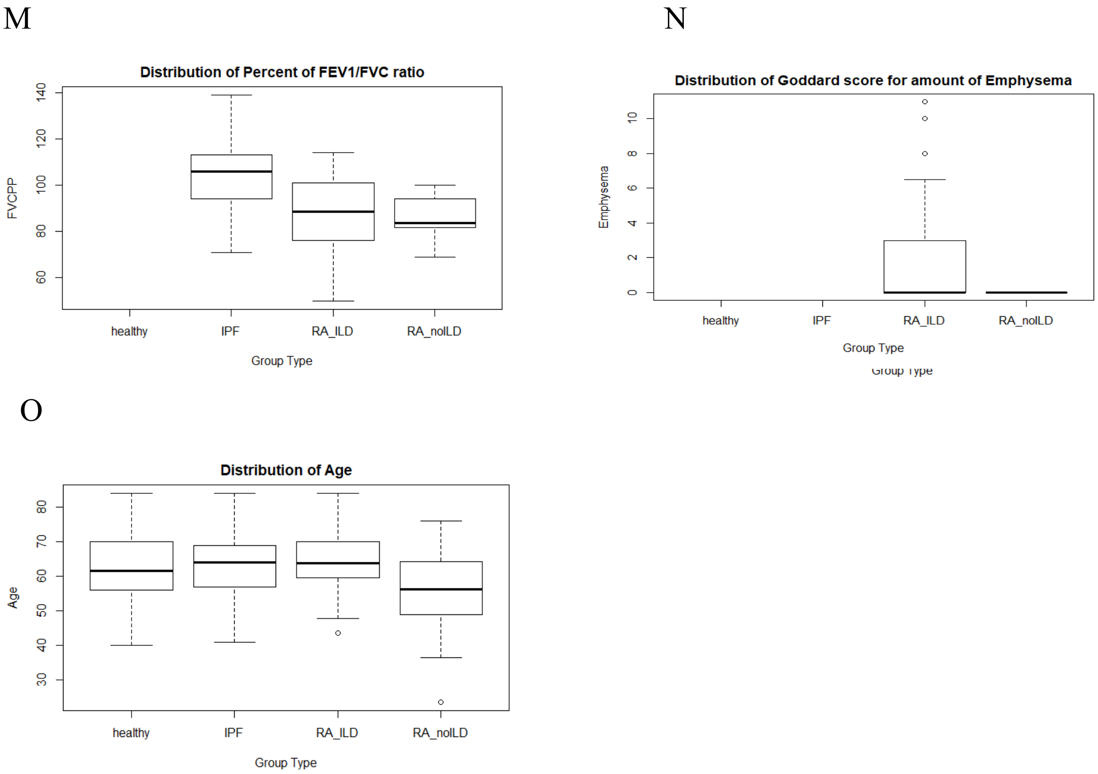
(3-A) is the volcano plot of RA no ILD vs. RA ILD (RA no ILD is the baseline), (3-B) is the volcano plot of Healthy vs. IPF (Healthy is the baseline), (3-C) is the volcano plot of RA ILD vs. IPF (RA ILD is the baseline), (3-D) is the volcano plot of RA cellular vs. RA fibrotic (RA cellular is the baseline). In X-axis, we log2-transformed Fold Change. In Y-axis, we took a negative value of logarithmic adjusted p-value to make the smaller adjusted p-value have a higher y value. The gray dots stand for the proteins which are not differential expressed. The blue dots stand for the proteins whose expression levels of the proteins in the compared group are smaller than those in the basic group. The red dots stand for the proteins whose expression levels of the proteins in the compared group are greater than that in the basic group.

Figure 4



A
Mosaic Plot of Gender
healthy
IPF
RA_ILD
RA_noILD
Group
Female
Male
Gender
>4
2:4
0:2
-2:0
-4:-2
<-4
Standardized Residuals:
B
Mosaic Plot of EverSmoker
healthy
IPF
RA_ILD
RA_noILD
Group
No
Yes
EverSmoker
>4
2:4
0:2
-2:0
-4:-2
<-4
Standardized Residuals:
C
Mosaic Plot of Race
healthy
IPF
RA_ILD
RA_noILD
Group
Native
White
Race
>4
2:4
0:2
-2:0
-4:-2
<-4
Standardized Residuals:
D
Distribution of BMI
BMI
20
25
30
35
40
45
healthy
IPF
RA_ILD
RA_noILD
Group Type
E
Distribution of Forced Expiratory Volume
Forced Expiratory Volume
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
healthy
IPF
RA_ILD
RA_noILD
Group Type
F
Distribution of Percent of Forced Expiratory Volume Predicted
FEV1PP
20
40
60
80
100
120
140
healthy
IPF
RA_ILD
RA_noILD
Group Type

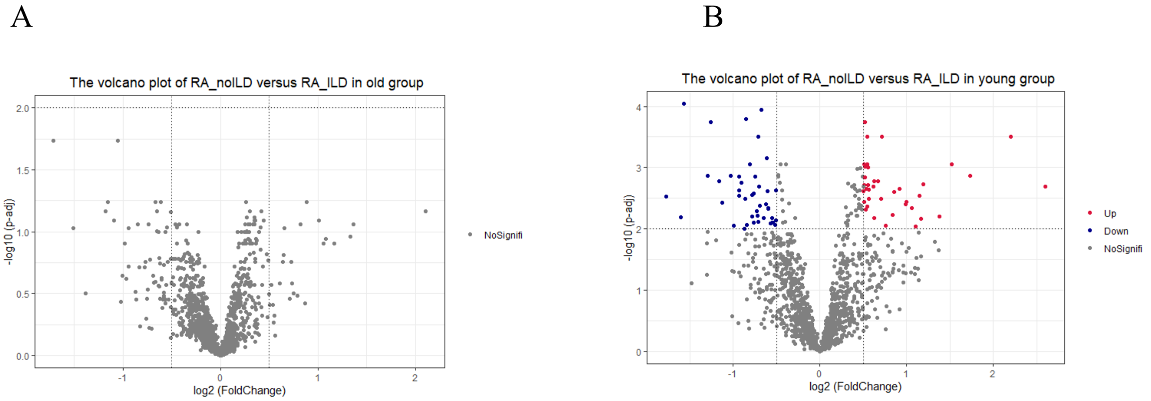

G
Distribution of Forced Vital Capacity
Forced Vital Capacity
1
2
3
4
5
6
healthy
IPF
RA_ILD
RA_noILD
Group Type
H
Distribution of Percent of Forced Vital Capacity Predicted
FVCPP
20
40
60
80
100
120
140
healthy
IPF
RA_ILD
RA_noILD
Group Type
I
Distribution of Total Lung Capacity
Total Lung Capacity
2
4
6
8
healthy
IPF
RA_ILD
RA_noILD
Group Type
J
Distribution of Percent of Total Lung Capacity
TLCPP
0
50
100
150
healthy
IPF
RA_ILD
RA_noILD
Group Type
K
Distribution of Percent of FEV1/FVC ratio
FVCPP
60
80
100
120
140
healthy
IPF
RA_ILD
RA_noILD
Group Type
L
Distribution of Goddard score for amount of Emphysema
Emphysema
0
2
4
6
8
10
healthy
IPF
RA_ILD
RA_noILD
Group Type



Distribution of related variables between groups.

(4-A) is the mosaic plot of Gender, (4-B) is the mosaic plot of EverSmoker, (4-C) is the mosaic plot of Race, (4-D) is the boxplot of BMI, (4-E) is the boxplot of Forced expiratory volume in 1 second, (4-F) is the boxplot of Percent of FEV1 predicted for the subject's gender, age, height, and weight, (4-G) is the boxplot of Forced vital capacity, (4-H) is the boxplot of Percent of FCV predicted for the subject's gender, age, height, and weight, (4-I) is the boxplot of Total lung capacity, (4-J) is the boxplot of Percent of TLC predicted for the subject's gender, age, height, and weight, (4-K) is the boxplot of Percent of FEV1/FVC ratio predicted for the subject's gender, age, height, and weight, (4-L) is the boxplot of Goddard score for amount of emphysema present on high resolution CT, (4-M) is the boxplot of 6-minute walk distance, (4-N) is the boxplot of Pack years of smoking, (4-O) is the boxplot of Age.

Figure 5



Volcano plots of different expression proteins between groups after adjusting covariates.

(5-A) is the volcano plot of RA no ILD versus RA ILD in old group(age>=60), (5-B) is the volcano plot of RA no ILD versus RA ILD in young group(age<60). In X-axis, we log2-transformed Fold Change. In Y-axis, we took a negative value of logarithmic adjusted p-value to make the smaller adjusted p-value have a higher y value. The gray dots stand for the proteins which are not differential expressed. The blue dots stand for the proteins whose expression levels of the proteins in the compared group are smaller than those in the basic group. The red dots stand for the proteins whose expression levels of the proteins in the compared group are greater than that in the basic group.

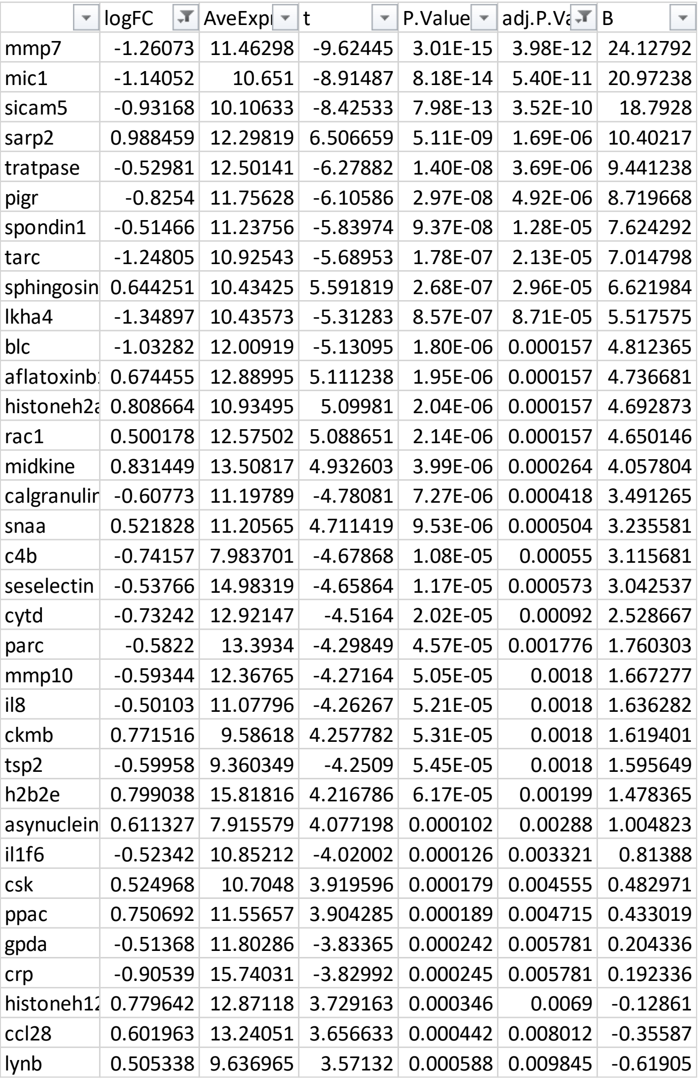
Table 1

The number of differentially expressed proteins between ILD phenotypes.

	RA no ILD &RA ILD (n=76)	Healthy &IPF (n=84)	RA ILD &IPF (n=81)	CellularNSIP &FibroticNSIP (n=34)
Significant	58	35	0	0
Up	23	15	0	0
Down	35	20	0	0
Not significant	1264	1287	1322	1322
Total	1322	1322	1322	1322

Table 2

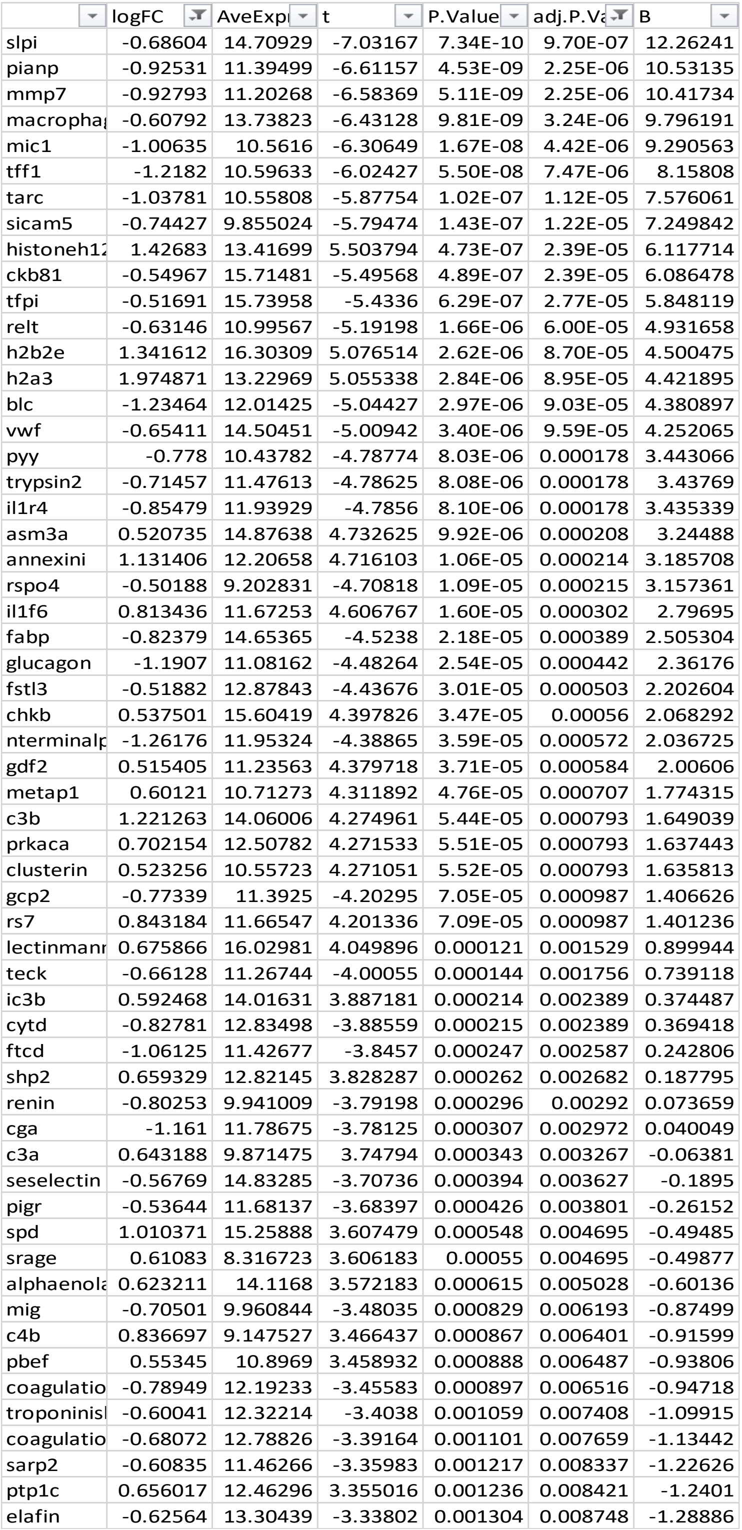
The differentially expressed proteins between IPF vs. Healthy.



	logFC	AveExp	t	P.Value	adj.P.Va	B
mmp7	-1.26073	11.46298	-9.62445	3.01E-15	3.98E-12	24.12792
mic1	-1.14052	10.651	-8.91487	8.18E-14	5.40E-11	20.97238
sicam5	-0.93168	10.10633	-8.42533	7.98E-13	3.52E-10	18.7928
sarp2	0.988459	12.29819	6.506659	5.11E-09	1.69E-06	10.40217
tratpase	-0.52981	12.50141	-6.27882	1.40E-08	3.69E-06	9.441238
pigr	-0.8254	11.75628	-6.10586	2.97E-08	4.92E-06	8.719668
spondin1	-0.51466	11.23756	-5.83974	9.37E-08	1.28E-05	7.624292
tarc	-1.24805	10.92543	-5.68953	1.78E-07	2.13E-05	7.014798
sphingosin	0.644251	10.43425	5.591819	2.68E-07	2.96E-05	6.621984
lkha4	-1.34897	10.43573	-5.31283	8.57E-07	8.71E-05	5.517575
blc	-1.03282	12.00919	-5.13095	1.80E-06	0.000157	4.812365
aflatoxinb1	0.674455	12.88995	5.111238	1.95E-06	0.000157	4.736681
histoneh2a	0.808664	10.93495	5.09981	2.04E-06	0.000157	4.692873
rac1	0.500178	12.57502	5.088651	2.14E-06	0.000157	4.650146
midkine	0.831449	13.50817	4.932603	3.99E-06	0.000264	4.057804
calgranulin	-0.60773	11.19789	-4.78081	7.27E-06	0.000418	3.491265
snaa	0.521828	11.20565	4.711419	9.53E-06	0.000504	3.235581
c4b	-0.74157	7.983701	-4.67868	1.08E-05	0.00055	3.115681
seselectin	-0.53766	14.98319	-4.65864	1.17E-05	0.000573	3.042537
cytd	-0.73242	12.92147	-4.5164	2.02E-05	0.00092	2.528667
parc	-0.5822	13.3934	-4.29849	4.57E-05	0.001776	1.760303
mmp10	-0.59344	12.36765	-4.27164	5.05E-05	0.0018	1.667277
il8	-0.50103	11.07796	-4.26267	5.21E-05	0.0018	1.636282
ckmb	0.771516	9.58618	4.257782	5.31E-05	0.0018	1.619401
tsp2	-0.59958	9.360349	-4.2509	5.45E-05	0.0018	1.595649
h2b2e	0.799038	15.81816	4.216786	6.17E-05	0.00199	1.478365
asynuclein	0.611327	7.915579	4.077198	0.000102	0.00288	1.004823
il1f6	-0.52342	10.85212	-4.02002	0.000126	0.003321	0.81388
csk	0.524968	10.7048	3.919596	0.000179	0.004555	0.482971
ppac	0.750692	11.55657	3.904285	0.000189	0.004715	0.433019
gpda	-0.51368	11.80286	-3.83365	0.000242	0.005781	0.204336
crp	-0.90539	15.74031	-3.82992	0.000245	0.005781	0.192336
histoneh12	0.779642	12.87118	3.729163	0.000346	0.0069	-0.12861
ccl28	0.601963	13.24051	3.656633	0.000442	0.008012	-0.35587
lynb	0.505338	9.636965	3.57132	0.000588	0.009845	-0.61905

Table 3

The differentially expressed proteins between RA ILD vs. RA no ILD.



	logFC	AveExp	t	P.Value	adj.P.Va	B
slpi	-0.68604	14.70929	-7.03167	7.34E-10	9.70E-07	12.26241
pianp	-0.92531	11.39499	-6.61157	4.53E-09	2.25E-06	10.53135
mmp7	-0.92793	11.20268	-6.58369	5.11E-09	2.25E-06	10.41734
macropha	-0.60792	13.73823	-6.43128	9.81E-09	3.24E-06	9.796191
mic1	-1.00635	10.5616	-6.30649	1.67E-08	4.42E-06	9.290563
tff1	-1.2182	10.59633	-6.02427	5.50E-08	7.47E-06	8.15808
tarc	-1.03781	10.55808	-5.87754	1.02E-07	1.12E-05	7.576061
sicam5	-0.74427	9.855024	-5.79474	1.43E-07	1.22E-05	7.249842
histoneh12	1.42683	13.41699	5.503794	4.73E-07	2.39E-05	6.117714
ckb81	-0.54967	15.71481	-5.49568	4.89E-07	2.39E-05	6.086478
tfpi	-0.51691	15.73958	-5.4336	6.29E-07	2.77E-05	5.848119
relt	-0.63146	10.99567	-5.19198	1.66E-06	6.00E-05	4.931658
h2b2e	1.341612	16.30309	5.076514	2.62E-06	8.70E-05	4.500475
h2a3	1.974871	13.22969	5.055338	2.84E-06	8.95E-05	4.421895
blc	-1.23464	12.01425	-5.04427	2.97E-06	9.03E-05	4.380897
vwf	-0.65411	14.50451	-5.00942	3.40E-06	9.59E-05	4.252065
pyy	-0.778	10.43782	-4.78774	8.03E-06	0.000178	3.443066
trypsin2	-0.71457	11.47613	-4.78625	8.08E-06	0.000178	3.43769
il1r4	-0.85479	11.93929	-4.7856	8.10E-06	0.000178	3.435339
asm3a	0.520735	14.87638	4.732625	9.92E-06	0.000208	3.24488
annexini	1.131406	12.20658	4.716103	1.06E-05	0.000214	3.185708
rspo4	-0.50188	9.202831	-4.70818	1.09E-05	0.000215	3.157361
il1f6	0.813436	11.67253	4.606767	1.60E-05	0.000302	2.79695
fabp	-0.82379	14.65365	-4.5238	2.18E-05	0.000389	2.505304
glucagon	-1.1907	11.08162	-4.48264	2.54E-05	0.000442	2.36176
fstl3	-0.51882	12.87843	-4.43676	3.01E-05	0.000503	2.202604
chkb	0.537501	15.60419	4.397826	3.47E-05	0.00056	2.068292
nterminalp	-1.26176	11.95324	-4.38865	3.59E-05	0.000572	2.036725
gdf2	0.515405	11.23563	4.379718	3.71E-05	0.000584	2.00606
metap1	0.60121	10.71273	4.311892	4.76E-05	0.000707	1.774315
c3b	1.221263	14.06006	4.274961	5.44E-05	0.000793	1.649039
prkaca	0.702154	12.50782	4.271533	5.51E-05	0.000793	1.637443
clusterin	0.523256	10.55723	4.271051	5.52E-05	0.000793	1.635813
gcp2	-0.77339	11.3925	-4.20295	7.05E-05	0.000987	1.406626
rs7	0.843184	11.66547	4.201336	7.09E-05	0.000987	1.401236
lectinmanr	0.675866	16.02981	4.049896	0.000121	0.001529	0.899944
teck	-0.66128	11.26744	-4.00055	0.000144	0.001756	0.739118
ic3b	0.592468	14.01631	3.887181	0.000214	0.002389	0.374487
cytd	-0.82781	12.83498	-3.88559	0.000215	0.002389	0.369418
ftcd	-1.06125	11.42677	-3.8457	0.000247	0.002587	0.242806
shp2	0.659329	12.82145	3.828287	0.000262	0.002682	0.187795
renin	-0.80253	9.941009	-3.79198	0.000296	0.00292	0.073659
cga	-1.161	11.78675	-3.78125	0.000307	0.002972	0.040049
c3a	0.643188	9.871475	3.74794	0.000343	0.003267	-0.06381
seselectin	-0.56769	14.83285	-3.70736	0.000394	0.003627	-0.1895
pigr	-0.53644	11.68137	-3.68397	0.000426	0.003801	-0.26152
spd	1.010371	15.25888	3.607479	0.000548	0.004695	-0.49485
srage	0.61083	8.316723	3.606183	0.00055	0.004695	-0.49877
alphaenola	0.623211	14.1168	3.572183	0.000615	0.005028	-0.60136
mig	-0.70501	9.960844	-3.48035	0.000829	0.006193	-0.87499
c4b	0.836697	9.147527	3.466437	0.000867	0.006401	-0.91599
pbef	0.55345	10.8969	3.458932	0.000888	0.006487	-0.93806
coagulatio	-0.78949	12.19233	-3.45583	0.000897	0.006516	-0.94718
troponinis	-0.60041	12.32214	-3.4038	0.001059	0.007408	-1.09915
coagulatio	-0.68072	12.78826	-3.39164	0.001101	0.007659	-1.13442
sarp2	-0.60835	11.46266	-3.35983	0.001217	0.008337	-1.22626
ptp1c	0.656017	12.46296	3.355016	0.001236	0.008421	-1.2401
elafin	-0.62564	13.30439	-3.33802	0.001304	0.008748	-1.28886

Table 4

The difference of covariates between ILD phenotypes.

	RA_ILD	RA_noILD		IPF	healthy		IPF	RA_ILD		CellularNSIP	FibroticNSIP	
	(n=39)	(n=37)	P.Value	(n=42)	(n=42)	P.Value	(n=42)	(n=39)	P.Value	(n=15)	(n=19)	P.Value
Gender												
Female	28 (71.8%)	34 (91.9%)	0.0239	29 (69.0%)	31 (73.8%)	0.629	29 (69.0%)	28 (71.8%)	0.787	13 (86.7%)	12 (63.2%)	0.123
Male	11 (28.2%)	3 (8.1%)		13 (31.0%)	11 (26.2%)		13 (31.0%)	11 (28.2%)		2 (13.3%)	7 (36.8%)	
Age (years)												
Mean (SD)	65.0 (9.16)	56.4 (11.6)	<0.001	63.9 (9.22)	62.7 (9.94)	0.563	63.9 (9.22)	65.0 (9.16)	0.579	65.7 (6.54)	64.6 (11.5)	0.607
Median [Min, Max]	63.8 [43.7, 84.1]	56.3 [23.7, 76.0]		64.0 [41.0, 84.	61.5 [40.0, 84.0]		64.0 [41.0, 84.	63.8 [43.7, 84.1]		65.3 [53.5, 79.1]	62.0 [43.7, 84.1]	
Race												
Asian	0 (0%)	0 (0%)	NA	2 (4.8%)	0 (0%)	NA	2 (4.8%)	0 (0%)	0.399	0 (0%)	0 (0%)	NA
Black	4 (10.3%)	3 (8.1%)		3 (7.1%)	0 (0%)		3 (7.1%)	4 (10.3%)		1 (6.7%)	3 (15.8%)	
Hispanic	1 (2.6%)	1 (2.7%)		1 (2.4%)	0 (0%)		1 (2.4%)	1 (2.6%)		0 (0%)	1 (5.3%)	
Native American	1 (2.6%)	0 (0%)		0 (0%)	0 (0%)		0 (0%)	1 (2.6%)		0 (0%)	0 (0%)	
Other	4 (10.3%)	0 (0%)		1 (2.4%)	0 (0%)		1 (2.4%)	4 (10.3%)		2 (13.3%)	2 (10.5%)	
White	28 (71.8%)	33 (89.2%)		35 (83.3%)	42 (100%)		35 (83.3%)	28 (71.8%)		12 (80.0%)	12 (63.2%)	
Missing	1 (2.6%)	0 (0%)		--	--		--	1 (2.6%)		0 (0%)	1 (5.3%)	
BMI (kg/m^2)												
Mean (SD)	28.4 (5.70)	28.4 (7.11)	0.585	--	--	--	26.9 (4.74)	28.4 (5.70)	0.468	26.9 (4.74)	29.6 (6.65)	0.325
Median [Min, Max]	27.6 [20.7, 47.1]	25.6 [17.8, 47.0]		--	--		27.4 [17.1, 37.	27.6 [20.7, 47.1]		26.1 [20.7, 36.3]	28.6 [20.7, 47.1]	
Missing	2 (5.1%)	0 (0%)		--	--		1 (2.4%)	2 (5.1%)		1 (6.7%)	0 (0%)	
EverSmoker												
No	12 (30.8%)	20 (54.1%)	0.0399	12 (28.6%)	14 (33.3%)	0.637	12 (28.6%)	12 (30.8%)	0.829	5 (33.3%)	7 (36.8%)	0.832
Yes	27 (69.2%)	17 (45.9%)		30 (71.4%)	28 (66.7%)		30 (71.4%)	27 (69.2%)		10 (66.7%)	12 (63.2%)	
Pack years of smoking												
Mean (SD)	24.5 (26.8)	14.9 (19.4)	0.745	--	--	--	12.0 (14.9)	24.5 (26.8)	0.0452	16.3 (15.3)	27.4 (31.3)	0.528
Median [Min, Max]	18.0 [0.00, 100]	5.00 [0.400, 55.5]		--	--		4.00 [0.00, 55.	18.0 [0.00, 100]		16.0 [0.00, 41.0]	15.0 [0.00, 100]	
Missing	7 (17.9%)	29 (78.4%)		--	--		0 (0%)	7 (17.9%)		4 (26.7%)	2 (10.5%)	
Forced expiratory volume												
Mean (SD)	1.96 (0.662)	2.43 (0.824)	0.109	--	--	--	--	--	--	1.59 (0.455)	2.36 (0.626)	0.0043
Median [Min, Max]	1.90 [0.860, 3.30]	2.34 [1.54, 4.41]		--	--		--	--		1.59 [0.860, 2.47]	2.32 [1.38, 3.30]	
Missing	9 (23.1%)	25 (67.6%)		--	--		--	--		3 (20.0%)	5 (26.3%)	
Percent of FEV1												
Mean (SD)	76.2 (22.7)	92.3 (14.6)	0.0119	--	--	--	67.8 (22.2)	76.2 (22.7)	0.149	68.7 (20.3)	85.4 (21.8)	0.0539
Median [Min, Max]	72.0 [39.0, 143]	93.5 [71.0, 113]		--	--		66.0 [21.0, 11	72.0 [39.0, 143]		67.0 [39.0, 107]	87.0 [52.0,143]	
Missing	5 (12.8%)	25 (67.6%)		--	--		1 (2.4%)	5 (12.8%)		2 (13.3%)	2 (10.5%)	
Forced vital capacity												
Mean (SD)	2.58 (0.909)	3.12 (1.22)	0.271	--	--	--	--	--	--	2.08 (0.661)	3.09 (0.803)	0.00305
Median [Min, Max]	2.45 [1.13, 4.27]	2.84 [1.95, 6.40]		--	--		--	--		2.09 [1.25, 3.37]	2.90 [2.14,4.27]	
Missing	9 (23.1%)	25 (67.6%)		--	--		--	--		3 (20.0%)	5 (26.3%)	
Percent of FCV												
Mean (SD)	75.6 (22.1)	94.3 (15.9)	0.00582	--	--	--	63.0 (20.8)	75.6 (22.1)	0.018	68.8 (18.6)	83.2 (22.9)	0.0373
Median [Min, Max]	71.0 [35.0, 146]	89.0 [75.0, 121]		--	--		59.0 [20.0, 10	71.0 [35.0, 146]		63.0 [46.0, 107]	85.5 [36.0, 146]	
Missing	4 (10.3%)	25 (67.6%)		--	--		1 (2.4%)	4 (10.3%)		2 (13.3%)	1 (5.3%)	
Percent of FEV1/FVC ratio												
Mean (SD)	87.7 (17.3)	86.3 (9.27)	0.832	--	--	--	103 (13.8)	87.7 (17.3)	<0.001	90.1 (19.2)	88.4 (16.3)	0.753
Median [Min, Max]	88.5 [50.0, 114]	83.5 [69.0, 100]		--	--		106 [71.0, 139	88.5 [50.0, 114]		92.0 [50.0, 113]	80.0 [61.0,114]	
Missing	5 (12.8%)	25 (67.6%)		--	--		1 (2.4%)	5 (12.8%)		2 (13.3%)	2 (10.5%)	
Total lung capacity												
Mean (SD)	4.14 (1.32)	5.02 (2.10)	0.378	--	--	--	--	--	--	3.47 (1.06)	4.87 (1.13)	0.00483
Median [Min, Max]	3.67 [1.94, 6.25]	4.59 [2.79, 9.64]		--	--		--	--		3.32 [1.94, 5.97]	4.88 [3.34, 6.25]	
Missing	11 (28.2%)	29 (78.4%)		--	--		--	--		4 (26.7%)	6 (31.6%)	
Percent of TLC												
Mean (SD)	75.4 (22.2)	90.4 (21.8)	0.0843	--	--	--	61.4 (31.2)	75.4 (22.2)	0.00792	69.3 (17.0)	83.1 (24.9)	0.227
Median [Min, Max]	75.0 [40.0, 138]	84.5 [60.0, 123]		--	--		56.0 [0.00, 19	75.0 [40.0, 138]		74.0 [41.0, 92.0]	81.0 [40.0, 138]	
Missing	7 (17.9%)	29 (78.4%)		--	--		6 (14.3%)	7 (17.9%)		3 (20.0%)	3 (15.8%)	
Goddard score for amount of emphysema												
Mean (SD)	1.87 (3.17)	0.00 (0.00)	<0.001	--	--	--	--	--	--	2.40 (3.72)	1.74 (3.04)	0.903
Median [Min, Max]	0.00 [0.00, 11.0]	0.00 [0.00, 0.00]		--	--		--	--		0.00 [0.00, 10.0]	0.00 [0.00, 11.00]	
6 minute walk distance												
Mean (SD)	358 (127)	537 (62.7)	0.0424	--	--	--	228 (174)	358 (127)	0.11	249 (181)	407 (101)	0.267
Median [Min, Max]	380 [121, 539]	524 [475, 624]		--	--		255 [0.00, 537	380 [121, 539]		249 [121, 377]	399 [293. 539]	
Missing	32 (82.1%)	33 (89.2%)		--	--		21 (50.0%)	32 (82.1%)		13 (86.7%)	15 (78.9%)	

Table 5

The number of differentially expressed proteins stratified by age.

	RA no ILD &RA ILD		IPF& Health	
Age	>=60	<60	>=60	<60
Significant	0	82	20	0
Up	0	38	5	0
Down	0	44	15	0
Not significant	1322	1250	1302	1322
Total	1322	1322	1322	1322