HHS Public Access

Author manuscript

Transl Res. Author manuscript; available in PMC 2022 February 01.

Published in final edited form as:

Transl Res. 2021 February ; 228: 1–12. doi:10.1016/j.trsl.2020.07.009.

MicroRNA and Protein-Coding Gene Expression Analysis in Idiopathic Pulmonary Fibrosis Yields Novel Biomarker Signatures Associated to Survival

Nancy G. Casanova[1,†], **Tong Zhou**[2,†], **Manuel Gonzalez-Garay**[1], **Yves A. Lussier**[3], **Nadera Sweiss**[4], **Shwu-Fan Ma**[5], **Imre Noth**[6], **Kenneth S. Knox**[7], **Joe G.N. Garcia**[1,*]

[1]Department of Medicine, University of Arizona Health Sciences, Tucson, AZ 85721, USA

[2]Department of Physiology and Cell Biology, University of Nevada, Reno School of Medicine, Reno, NV, 89557, USA

[3]Center for Biomedical Informatics and Biostatistics, University of Arizona, Tucson, AZ 85721, USA

[4]Section of Rheumatology Department of Medicine, University of Illinois at Chicago, Chicago, IL 60612, USA

[5]Section of Pulmonary/Critical Care, Department of Medicine, University of Chicago, Chicago, IL 60637, USA

[6]Division of Pulmonary and Critical Care Department of Medicine, University of Virginia, Charlottesville, VA 22903, USA

[7]Department of Medicine, College of Medicine-Phoenix, University of Arizona Health Sciences, Phoenix, AZ USA

Abstract

Idiopathic pulmonary fibrosis (IPF) is a chronic, progressive interstitial lung disease of unknown etiology that poses significant challenges in early diagnosis and prediction of progression. Analyses of microRNA and gene expression in IPF have yielded potentially predictive information. However, the relationship between microRNA/gene expression and quantitative phenotypic value in IPF remains controversial, as is the added value of this approach to current molecular signatures in IPF. To identify biomarkers predictive of survival in IPF via a microRNA-driven strategy. We profiled microRNA and protein-coding gene expression in peripheral blood mononuclear cells from 70 IPF subjects in a discovery cohort. We linked the microRNA/gene expression level with the quantitative phenotypic variation in IPF, including diffusing capacity of the lung for carbon monoxide and the forced vital capacity percent predicted. In silico analyses of

[*]Corresponding author: Joe G.N. Garcia, MD, 1230 North Cherry Avenue, suite 441, Tucson, AZ 85721, USA; Tel: 520-626-1197; skipgarcia@email.arizona.edu.

[†]These authors contributed equally to this work.

Disclosures: All authors have disclosed any potential conflict of interest relevant to the subject of this manuscript at the time of submission. Dr. Jo G.N. Garcia is the founder, and CEO of Aqualung therapeutics, Corp. All other authors do not have any financial or personal relationship with organizations that could potentially be perceived as influencing the present research. All authors have read the journal's authorship agreement.

expression profiles and quantitative phenotypic data allowed the generation of two sets of IPF molecular signatures (unique for microRNAs and protein-coding genes) that predict IPF survival. Each signature performed well in a validation cohort comprised of IPF patients aggregated from distinct patient populations recruited from different sites. Resampling test suggests that the protein-coding gene based signature is comparable and potentially superior to published IPF prognostic gene signatures. In conclusion, these results highlight the utility of microRNA-driven peripheral blood molecular signatures as valuable and novel biomarkers associated to individuals at high survival risk and for potentially facilitating individualized therapies in this enigmatic disorder.

Keywords

idiopathic pulmonary fibrosis; gene expression; microRNA expression; molecular signature; microarray

Introduction

Idiopathic pulmonary fibrosis (IPF) is a relentlessly progressive, chronic interstitial lung disease of unknown cause [1,2]. IPF primarily occurs in older adults, and is associated with the histopathologic and/or radiologic pattern of usual interstitial pneumonia (UIP) [3] and is the most common among the idiopathic interstitial pneumonias listed by the American Thoracic Society/European Respiratory Society [4]. The time course for IPF progression is highly variable: ranging from stability for several years to acute exacerbations with rapid deterioration. IPF is a terminal disease with a life expectancy of 3–5 years after an early diagnosis [5]. Current drugs only moderately slow the progression of the disease [6] and no drug reverses fibrosis development. Lung transplantation is the only curative therapy available for the treatment of IPF [7].

One critical barrier to monitor improvement and outcomes in IPF and to the identification of effective therapies has been the paucity of reliable biomarkers for diagnosis, prognosis, and responses to therapy [1,2]. Unfortunately, the search for effective IPF biomarkers is hindered by the inherent heterogeneity of the disease along with the consistent lack of correlation between biochemical markers, pathophysiologic variables and clinical outcomes [5,8,9]. Traditional methods to predict survival in IPF relies on the clinician integrating the clinical, laboratory, radiologic, and/or pathologic data to make a clinical-radiologic-pathologic correlation [5,10]. Current IPF biomarkers include Mucin protein 1 (MUC-1)[11], CC chemokine ligand 18 (CCL18)[12], serum surfactant proteins A (SP-A)[13] and chitinase-like glycoprotein YKL-40[14]. However, most biomarker studies are limited by the scale of candidate biomarkers, cohort size, and lack replication [8]. More recently, there has been a growing interest in the utility of genome-wide expression profiling to generate molecular signatures either from diseased tissues [15-22], whole blood or peripheral blood mononuclear cell (PBMC) RNA [23-25]. The clinical utility of peripheral blood gene expression profiling for the identification of molecular signatures as a surrogate for tissue biopsies has been studied in multiple diseases and disorders, including cancers and neurologic disease[26]. Similarly, in lung diseases, studies in COPD and sarcoidosis, have found overlapping gene expression

signatures between blood and lung tissue [27,28]. More recently, we identified a 52-gene expression signature in peripheral blood from IPF patients prospectively [23,29], which identified two molecular endotypes of IPF patients with significant differences in mortality or transplant-free survival in all cohorts indicating this approach is of significant utility for risk stratification in IPF.

Another approach to gene biomarker identification has been to utilize the power of microRNA (miRNA)-co-expressed genes in complex disease. miRNAs influence the gene expression at the posttranscriptional level; they are involved in multiple physiological processes such as metabolism, homeostasis, apoptosis and inflammatory process. Evidence revealed that miRNAs are involved in a wide range of diseases and hence their potential use as biomarkers [30,31]. We recently reported a 17 unique protein-coding gene signature in sarcoidosis derived from PBMC miRNA associated with Jak-STAT signaling pathway as the most significantly represented pathway[32]. This miRNA-regulated gene signature correlated with severity scores and differentiated sarcoidosis patients from healthy controls in independent validation cohorts. In IPF miRNAs are reported to be involved in lung epithelial repair, epithelial-mesenchymal transition, fibroblast activation, myofibroblast differentiation, macrophage polarization, alveolar epithelial cells senescence and collagen production [33].

In the current study, we sought to validate the utility of a miRNA approach utilizing profiled miRNA and protein-coding gene expression in PBMCs of well phenotyped IPF subjects.

We aim to correlate miRNA/gene expression levels with the quantitative phenotypic variation reported the pulmonary function test, such as the lung diffusing capacity (DLCO) and the forced vital capacity (FVC) percent predicted. DLCO values represent the ability of the lung to transfer inhaled air to the blood stream and act as a surrogate marker of lung damage[34] , similarly, FVC is considered a reliable measure of clinical status in patients with IPF[35], declining of the FVC is associated with an increase in mortality [36]. *In silico* analyses of expression profiles and quantitative phenotypic data allowed the generation of two sets of IPF molecular signatures (unique to miRNAs and to protein-coding genes) that predict IPF survival in both the study and validation cohorts. Further resampling test suggests that our protein-coding gene signature is comparable and potentially superior to published IPF prognostic gene signatures [23,29]. These results highlight the utility of miRNA-driven peripheral blood molecular signatures as valuable and novel biomarkers for survival and may potentially facilitate development of individualized therapies in this enigmatic disorder.

Materials and Methods

Subjects and PBMC samples

This research was carried out according to the principles of the Declaration of Helsinki. The Institutional Review Boards from University of Illinois at Chicago and University of Chicago (IRB #2010-0191 and 14104B respectively) approved the study. Written informed consent was obtained from all participants. The diagnosis of IPF was based on established international criteria [37]. IPF subjects were categorized into two datasets (Datasets A and B). Dataset A contained European descent non-Latino Caucasian Americans with IPF and known DLCO and FVC percent predicted (n = 47). Dataset B consisted of IPF patients of

African descent (n=6), Hispanics (n=4), Asian descent (n=1) and European descent non-Latino Caucasian Americans without known DLCO and FVC percent predicted (n = 6), as well as patients with unknown race information (n = 6). In total 23 IPF subjects were included in Dataset B. *In silico* studies focused on Dataset A. Dataset B was used as a validation cohort. Table 1 shows the characteristics of the IPF patients in the study and validation cohorts. For comparison purpose, we also included 32 European descent control subjects in this study (Dataset C).

Profiling transcriptional expression utilizing exon arrays

Isolation of PBMC from peripheral blood from 70 subjects was performed using the Ficoll-Paque method. Total RNA was isolated from PBMCs using standard molecular biology protocols (n=70) without DNA contamination or RNA degradation. Sample processing (e.g. cDNA generation, fragmentation, end labeling, hybridization to Affymetrix GeneChip Human Exon 1.0 ST arrays) was performed by the University of Chicago Functional Genomics Facility per manufacturer's instructions. Expression arrays were analyzed using the Affymetrix Power Tools v.1.12.0 (http://www.affymetrix.com/). The experimental probe masking workflow provided by the Affymetrix Power Tools was utilized to filter the probe set (exon-level) intensity files by removing probes that contain known SNPs in the dbSNP database [38]. Overall, of the ~1.4 million probe sets on the exon array, ~350,000 probe sets were found to contain at least one probe with a SNP (~600,000 probes) [39]. The resulting probe signal intensities were quartile normalized over all 70 samples. Probeset expression signals were summarized with the robust multi-array average (RMA) algorithm [40] and $\log_2$-transformed with a median polish. Expression signals of the ~22,000 transcript clusters (gene-level) were generated with the core set (i.e., with RefSeq-supported annotations) [41] of exons by taking averages of all annotated probe sets for each transcript cluster. Adjustment for possible batch effect was conducted by COMBAT (http://jlab.byu.edu//ComBat/) [42]. A transcript cluster was considered to be reliably expressed in these samples if the Affymetrix implemented DABG (detection above ground) [43] P-value was less than 0.01 in at least 67% of the samples. We further limited our analysis set to the genes with unique annotations (i.e., transcripts corresponding to unique genes) from the Affymetrix NetAffy website (https://www.affymetrix.com/analysis/netaffx/, accessed on Nov. 12, 2011). We also removed genes on chromosomes X and Y to avoid the potential confounding factor of sex. Totally, 11,245 transcript clusters met these criteria and were further analyzed. The microarray data has been uploaded into NCBI GEO database (GEO accession number: GSE38958)

Profiling miRNA expression using the Exiqon array

The expression levels of miRNAs were profiled using the Exiqon miRCURY™ LNA Array v10.0 (~700 human miRNAs). Briefly, total RNA from PBMCs was extracted and prepared per the manufacturer's protocol. Array hybridization was performed by Exiqon with the quantified signals background corrected using normexp with offset value 10 based on a convolution model [44] and normalized using the global Lowess regression algorithm. In total, 243 miRNAs were expressed in our study cohort (i.e., present in at least 66.7% of total samples). The MirSVR score, a support vector regression tool provided by microrna.org, was used to rank miRNAs target sites.

Statistical analysis

Spearman's rank correlation test was used to detect the relationship between miRNA/gene expression level and IPF physiologic features including DLCO and FVC. We calculated correlation coefficients and associated *P*-values using the R function "cor.test()" with method "spearman". The "survival" library of the R was used to conduct survival analysis. Kaplan-Meier survival curves were used to visualize the difference between patients with high- and low-mortality risk. To determine whether the prognostic power of the gene signature was potentially by chance alone, we performed a resampling test by generating 1,000 random gene signatures [45-47]. Cox proportional hazards regression of survival was conducted for each resampled gene signature. The association between each random gene signature and survival was measured by Cox regression coefficient.

Risk score

Based on the relationship between miRNA/gene expression and DLCO and FVC, we assigned each patient miRNA based and gene-based risk scores to predict survival by:

$$s_{miRNA} = -\sum_{i=1}^{n_{miRNA}} sgn(\rho_i^{miRNA}) \cdot e_i^{miRNA} \quad (1)$$

$$s_{gene} = -\sum_{i=1}^{n_{gene}} sgn(\rho_i^{gene}) \cdot e_i^{gene} \quad (2)$$

Here, s_{miRNA} and s_{gene} are the miRNA based and gene based patient risk scores respectively; n_{miRNA} and n_{gene} are the numbers of miRNAs and genes respectively; e_i^{gene} and e_i^{miRNA} are the expression levels of miRNA i and gene i respectively; ρ_i^{miRNA} and ρ_i^{gene} are the Spearman's rank correlation coefficients computed from the Dataset A (study cohort) for miRNA i and gene i respectively; and "sgn" denotes the sign function. A high score indicates a poor disease severity. The risk score was dichotomized based on the observed median for both the study and validation cohorts.

Results

Patient characteristics

The clinical characteristics of all study patients are displayed in Table 1. Significant differences in age and gender did not exist between the study cohorts (Dataset A and Dataset B) a ($P = 0.89$ by t-test for age and $P = 0.936$ by χ^2 test for gender).

Identifying the miRNAs correlated with lung function in IPF PBMCs

To determine the differentially expressed miRNAs correlated with lung function in IPF PBMCs, we analyzed miRNA expression pattern in Dataset A and linked miRNA expression level with IPF quantitative phenotypic data including DLCO and FVC. Spearman's rank correlation test was conducted both between expression value and DLCO and between expression value and FVC. The *P*-value of the correlation of DLCO and FVC with expression was recorded as P_{DLCO}^{miRNA} and P_{FVC}^{miRNA} respectively. In total, twelve miRNAs

yielded a 12-miRNA signature with both $P_{DLCO}^{miRNA} < 0.05$ and $P_{FVC}^{miRNA} < 0.05$, which was largely mirrored by the results generated by linear regression (Table S1). Eight of the twelve miRNAs showed positive correlation between expression and DLCO and FVC whereas four miRNAs showed negative correlation (Table 2, Figures 1 and S1), i.e. 4 up-regulated and 8 down-regulated miRNAs were identified in patients with relatively more severe IPF (Figure 2A).

Survival by the prioritized miRNAs

A risk score was assigned to each patient based on the expression of the 12-miRNA signature using formula (1). A high-risk score indicated a poor outcome. Patients were categorized as having a high-risk miRNA signature or a low-risk miRNA signature, with the median of the risk score as the threshold value. For our study cohorts, we tested the ability of risk score to correlate patients into prognostic groups based on miRNA expression. Kaplan-Meier survival analysis comparing patient groups demonstrated a significantly reduced survival for high-risk patients ($P = 0.005$) (Figure 2B). This association between risk score and survival was confirmed by Cox proportional hazard analysis of survival. High-risk patients exhibited a 5.43-fold increased risk for death ($P = 0.011$).

Identifying the genes correlated with lung function in IPF PBMCs

To determine protein-coding genes in IPF PBMCs that were correlated with lung function, we analyzed gene expression patterns in Dataset A and linked gene expression levels with DLCO and FVC. Spearman's rank correlation test was conducted between gene expression value and DLCO and FVC. The *P*-value of the correlation of DLCO and FVC with expression was recorded as P_{DLCO}^{gene} and P_{FVC}^{gene} respectively. In total, we found 712 genes with both $P_{DLCO}^{gene} < 0.01$ and $P_{FVC}^{gene} < 0.01$. 505 genes showed positive correlation (down-regulated in patients with relatively more severe IPF) between expression and DLCO and FVC while 207 showed negative pattern (up-regulated in patients with relatively more severe IPF). A search for enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) [48] physiological pathways among the prioritized genes using the NIH/DAVID [49,50] revealed 16 KEGG pathways ($P < 0.05$), including "Ubiquitin-mediated proteolysis", "Leukocyte transendothelial migration", "RIG-I-like receptor signaling pathway", "B cell receptor signaling pathway", "T cell receptor signaling pathway", "Focal adhesion", and "ECM-receptor interaction" (Figure 2C). It is noteworthy that the T cell-associated pathway was also found to be enriched among the published IPF prognostic genes [23,29].

miRNA-mRNA in silico analysis

We next searched for a list of predicted gene targets for the 12-miRNA signature using the *in silico* prediction provided by microrna.org [51], with filtering based on a stringent mirSVR score cutoff of −1.2 (this value represents the top 5% of mirSVR scores) to select the top miRNA-mRNA binding pairs [52]. We intersected these predicted mRNA targets against the microarray-observed 712 differentially expressed protein-coding genes revealing a total of 22 miRNA-mRNA pairs (Table 3). These 22 miRNA-mRNA pairs consisted of 21 unique protein-coding genes yielding a 21-gene signature (Figure 2D). Linear regression further

confirmed the relationship between the expression of the 21-gene signature and DLCO/FVC (Table S2). The direction of correlation of DLCO/FVC with expression of these 21 genes is exactly opposite to the direction of expression of the corresponding silencing miRNAs in the 12-miRNA signature (Figure S2 and S3). When the expression of each miRNA and target gene was directly compared, 17 of 22 miRNA-mRNA pairs demonstrated a significant inverse correlation ($P < 0.05$) in expression (Figure S4), which was largely mirrored by the results generated by linear regression (Table S3). We further checked the miRNA-mRNA binding prediction obtained from the miRDB database [53]. We found that 16 out of the 22 miRNA-mRNA pairs have a miRDB score larger than 50, which suggests the highly consistent outputs between the mirSVR and miRDB prediction algorithms (Table 3). Finally, we compared the expression pattern between the control subjects and IPF patients and eight out 21 genes were found to be differentially expressed ($P < 0.05$) (Figure S5).

Survival by the 21-gene signature

Utilizing a strategy employed for generation of the 12-miRNA signature, we next assigned a risk score to each patient based on the expression of the 21-gene signature using the above-cited formula (2). High-risk patients were defined as having a risk score greater than or equal to the group median score. In our study cohorts, we tested the ability of the score to assign patients into distinct prognostic groups based on gene expression. Kaplan-Meier survival analysis comparing patient groups demonstrated a significantly reduced survival for high-risk patients ($P = 0.002$) (Figure 2B). This association between risk score and survival was also confirmed by Cox proportional hazard analysis of survival; High-risk IPF patients exhibited a 7.79-fold increased risk for death ($P = 0.008$).

Validating the 12-miRNA and 21-gene signature

We used Dataset B as validation cohort. Unique risk scores were assigned to each patient based on the expression of the 12-miRNA and 21-gene signature respectively. High-risk patients were again defined as having a risk score greater than or equal to the group median score. Kaplan-Meier survival analysis demonstrated significantly reduced survival for high-risk patients, regardless of which signature was utilized ($P = 0.043$ for 12-miRNA signature and $P = 0.018$ for 21-gene signature) (Figure 3).

Survival analysis controlling for age

Since age is one of the major factors that affect life expectancy in IPF [5], we compared IPF patients older than 65 y/o to patients less than 65 y/o (65 is rough the median age among our patients). Both the 12-miRNA and 21-gene signatures significantly predicted survival within each older and younger IPF group, although the risk score of the 12-miRNA signature was only marginally different ($P = 0.054$) between high-risk and low-risk patients in the younger IPF group (Figure 4).

Comparison with published gene signatures

Recently, Herazo-Maya *et al.* presented an IPF prognostic signature containing 52 genes (52-gene signature, GSE28221), which was identified from PBMC gene expression profiles [23] and utilized an independent validation cohort comprised of 75 IPF subjects. More

recently, it was demonstrated that adding 52-gene risk profiles to the gender, age, and physiology index significantly improved IPF mortality predictive accuracy [29]. We evaluated the performance of our 21-gene signatures in that validation cohort. Kaplan-Meier survival analysis comparing patient groups indicated that the survival curves are different between patients with high- and low-mortality risk. Survival was significantly reduced for high-risk patients ($P = 0.009$) (Figure 5A). This association between risk score and survival was also confirmed by Cox proportional hazard analysis of survival. High-risk IPF patients exhibited a 2.28-fold increased risk for death ($P = 0.011$).

A published computational study indicated that random gene signatures have a high probability to be associated with survival outcome in breast cancer and published signatures are not significantly more associated with outcome than random predictors [47]. To address this potentially confounding issue, we performed a resampling test to determine whether the prognostic power of the 21-gene signature was potentially by chance alone and generated 1,000 random gene signatures with a size identical to the 21-gene signature. Cox proportional hazards regression of survival was conducted for each resampled gene signature. The association between each random gene signature and survival was measured by Cox regression coefficient. We found that we could reject the null hypothesis that the association between 21-gene signature and survival is by chance (one-tailed $P = 0.011$) (Figure 5B).

To compare the performance between the 21-gene and 52-gene signatures, a different resampling test was conducted. We generated 1,000 random gene signatures by randomly selecting 21 genes from the 52-gene signature. Cox proportional hazards regression of survival was conducted for each resampled gene signature. Among the 1,000 times of randomization, there were only 11 resampled gene signatures with higher Cox regression coefficient compared with that of our 21-gene signature (one-tailed $P = 0.011$) (Figure 5B).

Discussion

IPF is a progressive disease refractory to current therapies with an average survival of 3 to 5 years after diagnosis, although the rapidity to disease progression varies significantly among individuals. Traditionally, clinical, radiologic, and physiologic measurements are used as survival determinants [5]. A major goal of the current study was to identify potential miRNA or gene signatures to serve as novel biomarkers for survival in IPF and specifically to assess whether profiled miRNA and protein-coding gene expression in PBMCs added value to current signature approaches as suggested by our prior miRNA and protein-coding studies in sarcoidosis [32]. In contrast to dividing samples into control and patient groups, we only included IPF patient samples in our study and elected not to perform categorical statistics on patient data. Instead, we used Spearman’s rank correlation test to link the miRNA/gene expression level in PBMCs with the quantitative phenotypic variation in IPF including DLCO percent predicted and FVC percent predicted, reflections of disease severity. This approach reflects our belief that both statistical power and accuracy are advantaged by utilization of continuous statistics over categorical statistics. Largely, the result of artificial categorization of continuous variables may significantly bias results. We speculated that use

of these quantitative methods may provide novel insights into understanding of miRNA/gene expression profiles in IPF and yield a biomarker signature with substantial clinical utility.

Based on quantitative analysis, we initially identified 12 miRNAs (4 miRNAs up-regulated, 8 miRNAs down-regulated) that potentially relate to IPF severity including hsa-miR-150, hsa-miR-31, and hsa-miR-144, now recognized to be related to lung cancer [54,55], thalassemia [56], and sepsis [57]. In a recent study, lung miRNA expression profiles were compared between IPF patients and healthy controls with a greater number of up-regulated miRNAs vs down-regulated miRNAs identified in IPF patients [18]. In our studies, we identified only a single miRNA (hsa-miR-31) within our 12-miRNA signature that overlapped with the reported list of differentially expressed miRNAs between IPF patients and healthy controls (10). In addition, the directionality of hsa-miR-31 expression in our analysis was inconsistent between the two studies. This lack of overlap may reflect tissue-specific expression as our study did not evaluate lung tissue expression but rather analysis of PBMC expression, perhaps consistent with the notion that miRNA expression may be more tissue-specific than mRNA expression [58]. Based on the expression of the 12-miRNA signature, a scoring system was developed reflecting the extent and severity of IPF. We demonstrated that this miRNA-based score was significantly predictive of survival in IPF in both discovery and validation cohorts.

We also explored the utility of a protein-coding gene expression signature in IPF. Using Spearman's rank correlation test, we identified 207 up-regulated and 505 down-regulated genes in patients with increased severity of IPF. In agreement with prior reported studies, these prioritized genes were enriched in pathways such as "Ubiquitin mediated proteolysis" [16], "T cell receptor signaling pathway" [23], "Leukocyte transendothelial migration" [59], "B cell receptor signaling pathway" [60], "Focal adhesion" [60-62], and "ECM-receptor interaction" [61]. The potential role of miRNAs in regulating the dysregulated genes in IPF was examined using *in silico* prediction of miRNA target genes. We identified 22 miRNA-mRNA pairs, which consisted of 21 unique protein-coding genes. Several signature genes, *e.g. ABHD10*, *DECR1*, *DST*, *DUSP10*, *EXOSC9*, and *TTC35*, have been previously reported as dysregulated in IPF lung tissue when compared to normal lung fibroblasts [63]. Using the 21-gene signature, we also assigned each patient a risk score and identified that the gene-based risk score significantly associated to IPF outcome in both study and validation cohorts. Utilizing 1,000 times randomization, we found that 21 genes randomly picked from the pool of the 712 dysregulated genes in IPF yielded predictive power only 10% as effective as our 21-gene signature thereby highlighting the contribution of miRNA expression to capture dysregulated genes (data not shown). Moreover, although patient's age is a major factor affecting survival rate [5], we demonstrated that both 12-miRNA and 21-gene signatures performed well even when age was controlled for as a confounding factor.

In our study cohort (Dataset A), we only included Caucasian American patients because the variation among different populations may amplify the noise in expression data. In contrast, our validation cohort (Dataset B) contained patients not only from Caucasian American but also from other races including African, Hispanic, and Asian Americans. However, the racial complexity of the validation cohort did not affect the performance of our 12-miRNA and 21-

gene signatures in survival prediction. Both signatures successfully separated high-risk from the low-risk patients.

We compared our 21-gene signature with our prior published IPF prognostic 52-gene signature, which was also obtained from PBMC genome-wide profiles [23] and determined that the 21-gene signature effectively predicted clinical outcome in the validation cohort of that study. More recently, the 52-gene expression signature in peripheral blood from IPF patients was prospectively evaluated in a six-cohort study [29], which identified two molecular subphenotypes (low and high risk) of IPF patients with significant differences in mortality or transplant-free survival in all cohorts indicating this approach is of significant utility for risk stratification in IPF. Both our results and the 52-gene signature indicate that T cell-associated pathways are closely related to the prognosis of IPF. Resampling tests and comparison of our miRNA-derived 21-gene signature to our prior 52 gene signature indicates that the prognostic power of the 21-gene signature is not by chance. Furthermore, random selection of genes from the 52-gene signature failed to outperform our 21-gene signature. Whether the 21-gene signature identifies a unique sub-phenotype distinct from those identified by the 52-gene signature remains to be explored.

There are limitations to our study. One related to our analysis, although Spearman correlation coefficient can identify relationship between two continuous variables, it is possible that we missed signals if the relationship had a non-linear relationship that was not monotone. Another limitation to our study is primarily due to the retrospective study design, clinical data to correlate disease progression on these patients was absent, which limits the prediction performance of the proposed signature.

In summary, we derived two sets of unbiased molecular signatures from PBMCs for miRNA and protein-coding gene, respectively. Both signatures can be independently used as potential novel molecular biomarkers associated to IPF outcome. Although miRNA/gene expression variation in PBMCs may not necessarily reflect the dynamics of the pulmonary microenvironment in IPF patients, the potential contribution of these circulating cells in biomarker discovery has been well established. These results correlate pulmonary function tests with epigenomics mechanisms underlaying IPF and adds translational value to the clinical parameters commonly used for monitoring clinical status. The potential for the 21-gene signature to have clinical value in IPF studies will serve to validate the utility of this biomarker in the clinical care of IPF patients for survival prediction and serve as springboard to personalized therapies in this chronic and progressive form of lung disease.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This manuscript was supported in part by the NIH grants: HL126609 and HL091889.

N.C., T.Z and J.G.N.G. conceived of the study. T.Z. and N.C. collected and processed the data. T.Z. performed the biostatistical analysis. J.G.N.G. helped interpret the results. N.C.,TZ., drafted the manuscript. J.G.N.G. helped to

critically revise the manuscript. YA.L., M.L.G-G., N. S., S.M.,I. N., and K.S.K., contributed to study conception and implementation of the study. All authors read and approved the final manuscript.

References

1. Kim DS, Collard HR & King TE Jr. Classification and natural history of the idiopathic interstitial pneumonias. Proc Am Thorac Soc 3, 285–292 (2006). [PubMed: 16738191]
2. Visscher DW & Myers JL Histologic spectrum of idiopathic interstitial pneumonias. Proc Am Thorac Soc 3, 322–329 (2006). [PubMed: 16738196]
3. Raghu G, 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 183, 788–824 (2011). [PubMed: 21471066]
4. Society AT & Society ER American Thoracic Society/European Respiratory Society International Multidisciplinary Consensus Classification of the Idiopathic Interstitial Pneumonias. This joint statement of the American Thoracic Society (ATS), and the European Respiratory Society (ERS) was adopted by the ATS board of directors, June 2001 and by the ERS Executive Committee, June 2001. Am J Respir Crit Care Med 165, 277–304 (2002). [PubMed: 11790668]
5. King TE Jr., Tooze JA, Schwarz MI, Brown KR & Cherniack RM Predicting survival in idiopathic pulmonary fibrosis: scoring system and survival model. Am J Respir Crit Care Med 164, 1171–1181 (2001). [PubMed: 11673205]
6. Canestaro WJ, Forrester SH, Raghu G, Ho L & Devine BE Drug Treatment of Idiopathic Pulmonary Fibrosis: Systematic Review and Network Meta-Analysis. Chest 149, 756–766 (2016). [PubMed: 26836914]
7. Frankel SK & Schwarz MI Update in idiopathic pulmonary fibrosis. Curr Opin Pulm Med 15, 463–469 (2009). [PubMed: 19550329]
8. Zhang Y & Kaminski N Biomarkers in idiopathic pulmonary fibrosis. Curr Opin Pulm Med 18, 441–446 (2012). [PubMed: 22847105]
9. Maher TM PROFILEing idiopathic pulmonary fibrosis: rethinking biomarker discovery. Eur Respir Rev 22, 148–152 (2013). [PubMed: 23728868]
10. Mura M, et al. Predicting survival in newly diagnosed idiopathic pulmonary fibrosis: a 3-year prospective study. Eur Respir J 40, 101–109 (2012). [PubMed: 22241745]
11. Ishikawa N, Hattori N, Yokoyama A & Kohno N Utility of KL-6/MUC1 in the clinical management of interstitial lung diseases. Respir Investig 50, 3–13 (2012).
12. Prasse A, et al. Serum CC-chemokine ligand 18 concentration predicts outcome in idiopathic pulmonary fibrosis. Am J Respir Crit Care Med 179, 717–723 (2009). [PubMed: 19179488]
13. Kinder BW, et al. Serum surfactant protein-A is a strong predictor of early mortality in idiopathic pulmonary fibrosis. Chest 135, 1557–1563 (2009). [PubMed: 19255294]
14. Korthagen NM, et al. Serum and BALF YKL-40 levels are predictors of survival in idiopathic pulmonary fibrosis. Respir Med 105, 106–113 (2011). [PubMed: 20888745]
15. Selman M, et al. Gene expression profiles distinguish idiopathic pulmonary fibrosis from hypersensitivity pneumonitis. Am J Respir Crit Care Med 173, 188–198 (2006). [PubMed: 16166619]
16. Boon K, et al. Molecular phenotypes distinguish patients with relatively stable from progressive idiopathic pulmonary fibrosis (IPF). PLoS One 4, e5134 (2009). [PubMed: 19347046]
17. Konishi K, et al. Gene expression profiles of acute exacerbations of idiopathic pulmonary fibrosis. Am J Respir Crit Care Med 180, 167–175 (2009). [PubMed: 19363140]
18. Pandit KV, et al. Inhibition and role of let-7d in idiopathic pulmonary fibrosis. Am J Respir Crit Care Med 182, 220–229 (2010). [PubMed: 20395557]
19. Pandit KV, Milosevic J & Kaminski N MicroRNAs in idiopathic pulmonary fibrosis. Transl Res 157, 191–199 (2011). [PubMed: 21420029]
20. Navab R, et al. Prognostic gene-expression signature of carcinoma-associated fibroblasts in non-small cell lung cancer. Proc Natl Acad Sci U S A 108, 7160–7165 (2011). [PubMed: 21474781]
21. Shedden K, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med 14, 822–827 (2008). [PubMed: 18641660]

22. Pitroda SP, et al. Tumor endothelial inflammation predicts clinical outcome in diverse human cancers. PLoS One 7, e46104 (2012). [PubMed: 23056240]
23. Herazo-Maya JD, et al. Peripheral blood mononuclear cell gene expression profiles predict poor outcome in idiopathic pulmonary fibrosis. Sci Transl Med 5, 205ra136 (2013).
24. Desai AA, et al. A novel molecular signature for elevated tricuspid regurgitation velocity in sickle cell disease. Am J Respir Crit Care Med 186, 359–368 (2012). [PubMed: 22679008]
25. Kiliszek M, et al. Altered gene expression pattern in peripheral blood mononuclear cells in patients with acute myocardial infarction. PLoS One 7, e50054 (2012). [PubMed: 23185530]
26. Mohr S & Liew CC The peripheral-blood transcriptome: new insights into disease and risk assessment. Trends in molecular medicine 13, 422–432 (2007). [PubMed: 17919976]
27. Bhattacharya S, et al. Peripheral blood gene expression profiles in COPD subjects. Journal of clinical bioinformatics 1, 12 (2011). [PubMed: 21884629]
28. Koth LL, et al. Sarcoidosis blood transcriptome reflects lung inflammation and overlaps with tuberculosis. Am J Respir Crit Care Med 184, 1153–1163 (2011). [PubMed: 21852540]
29. Herazo-Maya JD, et al. Validation of a 52-gene risk profile for outcome prediction in patients with idiopathic pulmonary fibrosis: an international, multicentre, cohort study. Lancet Respir Med 5, 857–868 (2017). [PubMed: 28942086]
30. Anindo MI & Yaqinuddin A Insights into the potential use of microRNAs as biomarker in cancer. International journal of surgery (London, England) 10, 443–449 (2012).
31. Osman A MicroRNAs in health and disease--basic science and clinical applications. Clinical laboratory 58, 393–402 (2012). [PubMed: 22783567]
32. Zhou T, et al. Identification of Jak-STAT signaling involvement in sarcoidosis severity via a novel microRNA-regulated peripheral blood mononuclear cell gene signature. Sci Rep 7, 4237 (2017). [PubMed: 28652588]
33. Miao C, Xiong Y, Zhang G & Chang J MicroRNAs in idiopathic pulmonary fibrosis, new research progress and their pathophysiological implication. Exp Lung Res 44, 178–190 (2018). [PubMed: 29683754]
34. Ozeki N, et al. The diffusing capacity of the lung for carbon monoxide is associated with the histopathological aggressiveness of lung adenocarcinoma. Eur J Cardiothorac Surg 52, 969–974 (2017). [PubMed: 28520857]
35. du Bois RM, et al. Forced vital capacity in patients with idiopathic pulmonary fibrosis: test properties and minimal clinically important difference. Am J Respir Crit Care Med 184, 1382–1389 (2011). [PubMed: 21940789]
36. Lassenius MI, et al. Forced Vital Capacity (FVC) decline, mortality and healthcare resource utilization in idiopathic pulmonary fibrosis. Eur Clin Respir J 7, 1702618 (2020). [PubMed: 32002175]
37. Society AT & Society ER American Thoracic Society. Idiopathic pulmonary fibrosis: diagnosis and treatment. International consensus statement. Am J Respir Crit Care Med 161, 646–664 (2000). [PubMed: 10673212]
38. Sherry ST, et al. dbSNP: the NCBI database of genetic variation. Nucleic acids research 29, 308–311 (2001). [PubMed: 11125122]
39. Duan S, Zhang W, Bleibel WK, Cox NJ & Dolan ME SNPinProbe_1.0: a database for filtering out probes in the Affymetrix GeneChip human exon 1.0 ST array potentially affected by SNPs. Bioinformation 2, 469–470 (2008). [PubMed: 18841244]
40. Irizarry RA, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4, 249–264 (2003). [PubMed: 12925520]
41. Pruitt KD, Tatusova T & Maglott DR NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research 35, D61–65 (2007). [PubMed: 17130148]
42. Johnson WE, Li C & Rabinovic A Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127 (2007). [PubMed: 16632515]
43. Affymetrix. Exon array background correction. Affymetrix Whitepaper http://media.affymetrix.com/support/technical/whitepapers/exon_background_correction_whitepaper.pdf(2005).

44. Ritchie ME, et al. A comparison of background correction methods for two-colour microarrays. Bioinformatics 23, 2700–2707 (2007). [PubMed: 17720982]
45. Ko EA, Lee H, Sanders KM, Koh SD & Zhou T Expression of Alpha-type Platelet-derived Growth Factor Receptor-influenced Genes Predicts Clinical Outcome in Glioma. Transl Oncol 13, 233–240 (2020). [PubMed: 31869747]
46. Ko EA, Sanders KM & Zhou T A transcriptomic insight into the impacts of mast cells in lung, breast, and colon cancers. Oncoimmunology 6, e1360457 (2017). [PubMed: 29147625]
47. Venet D, Dumont JE & Detours V Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput Biol 7, e1002240 (2011). [PubMed: 22028643]
48. Kanehisa M, Goto S, Kawashima S, Okuno Y & Hattori M The KEGG resource for deciphering the genome. Nucleic acids research 32, D277–280 (2004). [PubMed: 14681412]
49. Huang da W, Sherman BT & Lempicki RA Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4, 44–57 (2009). [PubMed: 19131956]
50. Dennis G Jr., et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4, P3 (2003). [PubMed: 12734009]
51. Betel D, Wilson M, Gabow A, Marks DS & Sander C The microRNA.org resource: targets and expression. Nucleic acids research 36, D149–153 (2008). [PubMed: 18158296]
52. Betel D, Koppal A, Agius P, Sander C & Leslie C Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol 11, R90 (2010). [PubMed: 20799968]
53. Chen Y & Wang X miRDB: an online database for prediction of functional microRNA targets. Nucleic acids research 48, D127–D131 (2020). [PubMed: 31504780]
54. Yanaihara N, et al. Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. Cancer Cell 9, 189–198 (2006). [PubMed: 16530703]
55. Melkamu T, Zhang X, Tan J, Zeng Y & Kassie F Alteration of microRNA expression in vinyl carbamate-induced mouse lung tumors and modulation by the chemopreventive agent indole-3-carbinol. Carcinogenesis 31, 252–258 (2010). [PubMed: 19748927]
56. Fu YF, et al. Mir-144 selectively regulates embryonic alpha-hemoglobin synthesis during primitive erythropoiesis. Blood 113, 1340–1349 (2009). [PubMed: 18941117]
57. Vasilescu C, et al. MicroRNA fingerprints identify miR-150 as a plasma prognostic marker in patients with sepsis. PLoS One 4, e7405 (2009). [PubMed: 19823581]
58. Su WL, Kleinhanz RR & Schadt EE Characterizing the role of miRNAs within gene regulatory networks using integrative genomics techniques. Mol Syst Biol 7, 490 (2011). [PubMed: 21613979]
59. Bargagli E, Prasse A, Olivieri C, Muller-Quernheim J & Rottoli P Macrophage-derived biomarkers of idiopathic pulmonary fibrosis. Pulm Med 2011, 717130 (2011). [PubMed: 21637368]
60. Cho JH, et al. Systems biology of interstitial lung diseases: integration of mRNA and microRNA expression changes. BMC Med Genomics 4, 8 (2011). [PubMed: 21241464]
61. Thannickal VJ & Horowitz JC Evolving concepts of apoptosis in idiopathic pulmonary fibrosis. Proc Am Thorac Soc 3, 350–356 (2006). [PubMed: 16738200]
62. Selman M, Pardo A & Kaminski N Idiopathic pulmonary fibrosis: aberrant recapitulation of developmental programs? PLoS Med 5, e62 (2008). [PubMed: 18318599]
63. Emblom-Callahan MC, et al. Genomic phenotype of non-cultured pulmonary fibroblasts in idiopathic pulmonary fibrosis. Genomics 96, 134–145 (2010). [PubMed: 20451601]

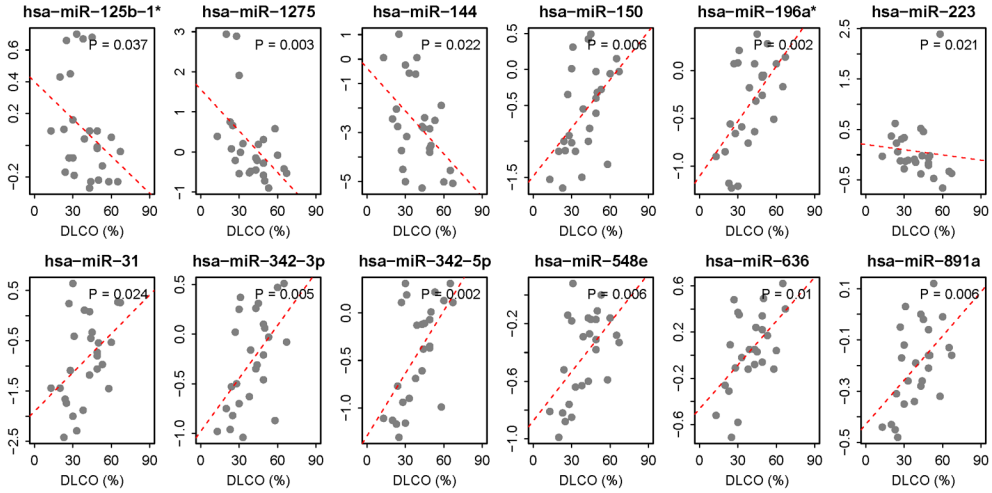

Figure 1.

Expression level as a function of DLCO for the 12-miRNAs signature in IPF. Y-axis denotes the relative miRNA expression level. Each point represents a single IPF subject. The *P*-value for each miRNA is measured by Spearman's rank correlation test.

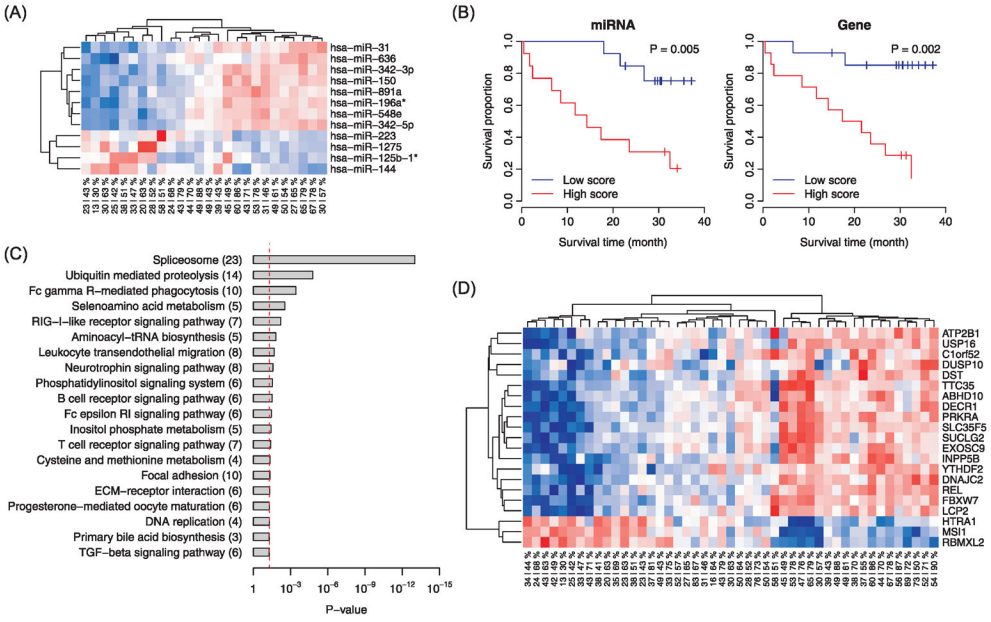

Figure 2.

12-miRNA and 21-gene signatures of IPF. (A) Heatmap of 12-miRNA signature. Red represents increased miRNA expression while blue represents decreased expression. Subjects are labeled by their corresponding DLCO and FVC % predicted (DLCO | FVC). (B) Kaplan-Meier curves for patients from study cohort with high and low risk scores. High scores correspond to the higher predicted risk and low scores correspond to the lower predicted risk. The *P*-values for both the 12-miRNA and 21-gene signatures are measured by log-rank test. (C) Enriched pathways among the prioritized 712 genes in IPF. The top 20 ranked KEGG pathways are listed. The red line indicates the cutoff of significance (0.05). The number of genes in each pathway is shown beside the pathway name. (D) Heatmap of 21-gene signature. Red represents increased gene expression while blue represents gene expression. The patients are labeled by their corresponding DLCO and FVC % predicted (DLCO | FVC).

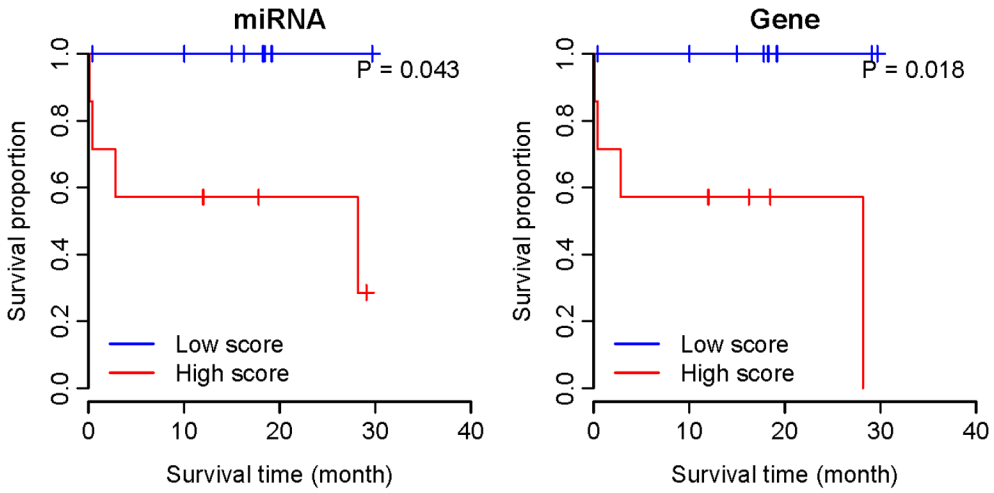

Figure 3.

Kaplan-Meier curves for patients from validation cohort with high- and low-risk scores. High-risk scores correspond to the higher predicted risk and low scores correspond to the lower predicted risk. The P-values for both the 12-miRNA and 21-gene signatures are measured by log-rank test.

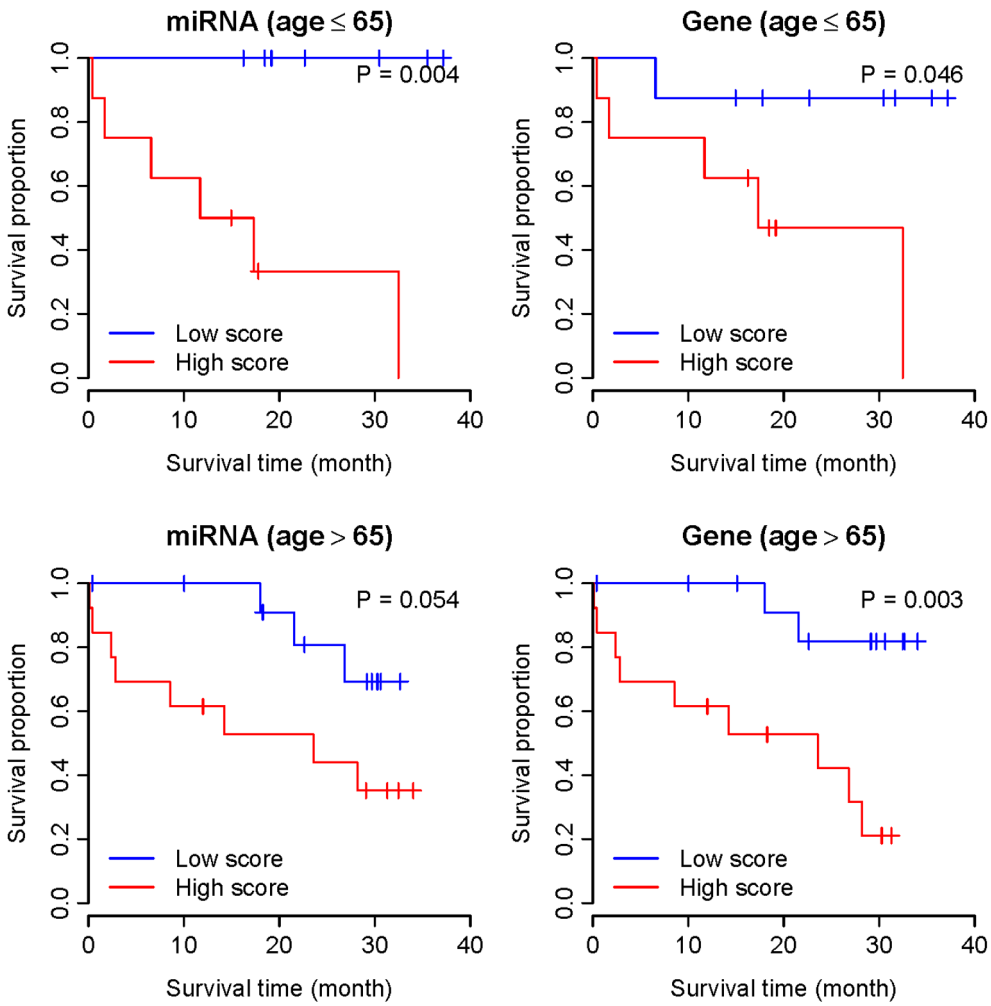

Figure 4.

Kaplan-Meier curves for patients with high and low risk scores stratified by age group. The patients older than 65 and younger or equal to 65 are analyzed separately. High scores correspond to the higher predicted risk and low scores correspond to the lower predicted risk. In total, there were 21 and 26 patients with age ≤ 65 and > 65, respectively. The P-values for both the 12-miRNA and 21-gene signatures are measured by log-rank test.

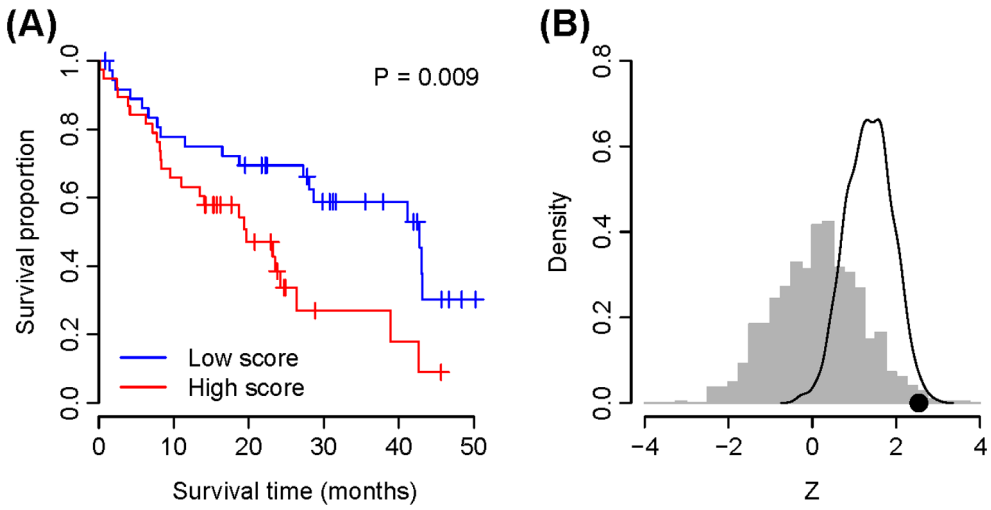

Figure 5.

Comparison between the 21-gene and 52-gene signatures. (A) Kaplan-Meier curves of the 21-gene signature in the validation cohort of Herazo-Maya *et al.*'s study. High scores correspond to the higher predicted risk and low scores correspond to the lower predicted risk. The *P*-values are measured by log-rank test. (B) Non-random predictive power of the 21-gene signature. *Z* denotes the Cox regression coefficient. The black dots stand for the Z values of the 21-gene signature. The grey histograms show the distribution of *Z* values for the 1,000 resampled gene signatures (genes were randomly picked up from human genome) with identical size as 21-gene signature under the null hypothesis of no association between 21-gene signature and survival. The black curve shows the distribution of *Z* values for the 1,000 resampled gene sets obtained from the 52-gene signature (genes were randomly picked up from Herazo-Maya *et al.*'s 52-gene signature).

Table 1.

Patient characteristics

Characteristics	Dataset A (n=47)	Dataset B (n=23)
Age	68.3 ± 6.9	67.9 ± 8.6
Gender (Male/Female)	40/7	19/4
FVC % predicted	62.5 ± 14.6	-
DLCO % predicted	43.5 ± 16.9	-

Data are presented as n, mean ± sd or %. FVC: forced vital capacity, DLCO: diffusing lung capacity for carbon monnoxide

Table 2.

The miRNAs correlated with lung function in IPF PBMCs

	DLCO		FVC	
miRNA	**ρ***	**P_{DLCO_miRNA}**	**ρ***	**P_{FVC_miRNA}**
hsa-miR-125b-1†	−0.41	0.037	−0.49	0.009
hsa-miR-1275	−0.55	0.003	−0.49	0.009
hsa-miR-144	−0.45	0.022	−0.59	0.001
hsa-miR-150	0.52	0.006	0.42	0.028
hsa-miR-196a†	0.58	0.002	0.42	0.030
hsa-miR-223	−0.45	0.021	−0.43	0.024
hsa-miR-31	0.44	0.024	0.41	0.032
hsa-miR-342-3p	0.53	0.005	0.47	0.014
hsa-miR-342-5p	0.59	0.002	0.47	0.014
hsa-miR-548e	0.53	0.006	0.44	0.022
hsa-miR-636	0.50	0.010	0.46	0.016
hsa-miR-891a	0.53	0.006	0.47	0.012

DLCO, lung diffusing capacity; FVC, forced vital capacity; IPF, idiopathic pulmonary fibrosis; PBMC, peripheral blood mononuclear cell.

*Spearman's rank correlation coefficient between miRNA expression level and DLCO/FVC

†The miRNA originates from the opposite arm of the precursor

Table 3.

The 21-gene signature in IPF PBMCs

			DLCO		FVC	
Gene	**miRNA**[a]	**SVR**[b]	ρ^c	P^{gene}_{DLCO}	ρ^c	P^{gene}_{FVC}
ABHD10	hsa-miR-223	−1.2577	0.49	5.30×10^{-4}	0.46	9.89×10^{-4}
ATP2B1	hsa-miR-144	−1.2072	0.51	2.92×10^{-4}	0.41	4.19×10^{-3}
C1orf52	hsa-miR-144	−1.2476	0.46	1.24×10^{-3}	0.49	4.49×10^{-4}
DECR1	hsa-miR-223	−1.2593	0.41	3.78×10^{-3}	0.46	1.06×10^{-3}
DNAJC2	hsa-miR-144	−1.3116	0.53	1.07×10^{-4}	0.48	5.79×10^{-4}
DST	hsa-miR-144	−1.3074	0.47	9.01×10^{-4}	0.57	2.40×10^{-5}
DUSP10	hsa-miR-223	−1.2357	0.39	6.79×10^{-3}	0.42	2.95×10^{-3}
EXOSC9	hsa-miR-144	−1.2346	0.55	6.83×10^{-5}	0.38	7.40×10^{-3}
FBXW7	hsa-miR-223	−1.3105	0.69	7.22×10^{-8}	0.40	4.82×10^{-3}
HTRA1	hsa-miR-31	−1.2449	−0.60	7.89×10^{-6}	−0.42	2.61×10^{-3}
INPP5B	hsa-miR-223	−1.2291	0.40	5.70×10^{-3}	0.56	4.21×10^{-5}
LCP2	hsa-miR-144	−1.2774	0.56	4.12×10^{-5}	0.38	8.30×10^{-3}
MSI1	hsa-miR-548e	−1.3278	−0.56	5.11×10^{-5}	−0.44	2.00×10^{-3}
PRKRA	hsa-miR-144	−1.2368	0.48	7.11×10^{-4}	0.42	3.06×10^{-3}
RBMXL2	hsa-miR-150	−1.2373	−0.56	4.50×10^{-5}	−0.42	2.96×10^{-3}
REL	hsa-miR-144	−1.2180	0.53	1.45×10^{-4}	0.42	2.90×10^{-3}
SLC35F5	hsa-miR-144	−1.2466	0.44	1.90×10^{-3}	0.38	8.02×10^{-3}
SUCLG2	hsa-miR-144	−1.2016	0.53	1.17×10^{-4}	0.37	9.09×10^{-3}
TTC35	hsa-miR-223	−1.3046	0.44	2.16×10^{-3}	0.41	3.99×10^{-3}
	hsa-miR-144	−1.3256				
USP16	hsa-miR-223	−1.2207	0.45	1.32×10^{-3}	0.42	2.96×10^{-3}
YTHDF2	hsa-miR-144	−1.3317	0.51	2.60×10^{-4}	0.46	9.80×10^{-4}

[a] Silencing miRNAs for the protein-coding genes in 21-gene signature

[b] Abbreviation of mirSVR score, ranking miRNA-mRNA binding

[c] Spearman's rank correlation coefficient between gene expression level and DLCO/FVC