Respiratory Research

RESEARCH

Open Access

Blood gene expression risk profiles and interstitial lung abnormalities: COPDGene and ECLIPSE cohort studies

Matthew Moll[1,2,3†], Brian D. Hobbs[1,2,3†], Aravind Menon[2,3], Auyon J. Ghosh[1,2,3], Rachel K. Putman[2,3], Takuya Hino[3,4], Akinori Hata[3,4], Edwin K. Silverman[1,2,3], John Quackenbush[1,5], Peter J. Castaldi[1,3,6], Craig P. Hersh[1,2,3], Michael J. McGeachie[1], Don D. Sin[7], Ruth Tal-Singer[8], Mizuki Nishino[3,4], Hiroto Hatabu[3,4], Gary M. Hunninghake[2,3†] and Michael H. Cho[1,2,3*†]

Abstract

Background: Interstitial lung abnormalities (ILA) are radiologic findings that may progress to idiopathic pulmonary fibrosis (IPF). Blood gene expression profiles can predict IPF mortality, but whether these same genes associate with ILA and ILA outcomes is unknown. This study evaluated if a previously described blood gene expression profile associated with IPF mortality is associated with ILA and all-cause mortality.

Methods: In COPDGene and ECLIPSE study participants with visual scoring of ILA and gene expression data, we evaluated the association of a previously described IPF mortality score with ILA and mortality. We also trained a new ILA score, derived using genes from the IPF score, in a subset of COPDGene. We tested the association with ILA and mortality on the remainder of COPDGene and ECLIPSE.

Results: In 1469 COPDGene (training n = 734; testing n = 735) and 571 ECLIPSE participants, the IPF score was not associated with ILA or mortality. However, an ILA score derived from IPF score genes was associated with ILA (meta-analysis of test datasets OR 1.4 [95% CI: 1.2–1.6]) and mortality (HR 1.25 [95% CI: 1.12–1.41]). Six of the 11 genes in the ILA score had discordant directions of effects compared to the IPF score. The ILA score partially mediated the effects of age on mortality (11.8% proportion mediated).

Conclusions: An ILA gene expression score, derived from IPF mortality-associated genes, identified genes with concordant and discordant effects on IPF mortality and ILA. These results suggest shared, and unique biologic processes, amongst those with ILA, IPF, aging, and death.

Introduction

Interstitial lung abnormalities (ILA) are specific radiologic findings detected on computed tomography (CT) scans [1–3] and, in some instances, may represent, or progress to, pulmonary fibrosis [4, 5]. Both ILA and idiopathic pulmonary fibrosis (IPF) are associated with pulmonary symptoms [2, 6], diminished lung function [1, 2, 4, 5, 7, 8], and mortality [9]. Despite these clinical similarities, genetic analyses reveal that these entities have both overlapping and distinct genetic risk alleles [10, 11].

†Matthew Moll and Brian D. Hobbs are equal contributions and co-first authors.

†Gary M. Hunninghake and Michael H. Cho jointly supervised this work and are co-senior authors.

*Correspondence: remhc@channing.harvard.edu

[1] Channing Division for Network Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA
Full list of author information is available at the end of the article

Taken together, the clinical and genetic evidence suggests that ILA and IPF possess both shared and unique pathobiology.

The risk of IPF and IPF mortality are not due to genetic variants alone [12]. Gene expression data can reflect the combination of both genetic variation and environmental factors that contribute to IPF pathogenesis (e.g. cigarette smoking) [13]. Herazo-Maya et al. [14] used blood microarray data to develop a 52-gene IPF risk score that predicted mortality in multiple IPF cohorts [15]. Whether this IPF mortality risk score, or the specific genes in this risk score, are associated with ILA or ILA mortality in current and former smokers is not known. A peripheral blood signature that predicted ILA and mortality in ILA could be important for identifying early disease and those at risk for worse outcomes. In addition, shared gene expression features of IPF progression and ILA risk would highlight important biologic processes associated with the spectrum of interstitial lung disease from precursor lesions to irreversible fibrosis to death.

Therefore, we hypothesized that the IPF risk score would be associated with ILA and ILA-associated mortality, and that this risk would be driven by a subset of the genes in the IPF mortality risk score; this subset of genes may also lend insight into the biologic mechanisms relating ILA to fibrosis and death. Briefly, the original IPF risk score applied the scoring algorithm of molecular subphenotypes (SAMS) to 52 genes expressed in peripheral blood mononuclear cells to predict transplant-free survival [14]. The SAMS method classifies participants into high- or low-risk groups based on the proportion of genes expected to have increased or decreased expression levels. This score suggested several immune alterations that may be contributing to poor IPF outcomes, including CD4 + T cells with CD28 downregulation and T cell exhaustion, activation of mast cells and fibroblasts. This score was able to improve survival prediction when added to clinical factors in six independent cohorts [14, 15].

Methods

Study populations

All study participants provided written informed consent. Each study center obtained institutional review board approval.

COPDGene

We included participants from the Genetic Epidemiology of COPD (COPDGene) study [16] who had a 5-year follow up visit with blood RNA-sequencing (RNA-seq) data and computed tomography (CT) scans that were assessed for the presence ILA [9]. Details of the COPDGene study have been previously described [16]. Briefly, COPDGene is a prospective cohort study of non-Hispanic white (NHW) and African American (AA) smokers (≥ 10 pack-years of smoking), aged 45–80 years at study initiation, with and without COPD. The study was originally conceived of as a case–control study and has been extended into a longitudinal study with 5- and 10-year follow up visits. Whole blood samples, as well as anthropometric, spirometry, and CT imaging data were collected at each visit.

ECLIPSE

We included Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE) study [17] participants with microarray gene expression data and chest CT scans that were assessed for the presence of ILA [9]. ECLIPSE participants were smokers (≥ 10 pack-years of smoking) aged 45–75 years at study enrollment. Baseline questionnaire, spirometry, CT imaging and blood samples were collected. In ECLIPSE, COPD participants, but not controls, were followed longitudinally for 3 years.

ILA phenotyping

In both COPDGene and ECLIPSE, thoracic CT scans were assessed for ILA using a sequential method by three readers as previously described [10, 11]. The definition of ILA included in this manuscript conforms to the updated definition utilized by the Fleischner society [3, 18–20].

Preparation of gene expression data

RNA sequencing

COPDGene whole blood RNA-seq data was available at the 5-year follow up visit, and data generation was previously described [13]. Briefly, PAXgene Blood RNA tubes were used to collect whole blood samples, and the Qiagen PreAnalytiX PAXgene Blood miRNA Kit (Qiagen, Valencia, CA) was used to extract total RNA. Samples with concentrations > 25 ug/uL and RNA integrity number (RIN) > 6 were eligible for sequencing. TruSeq Stranded Total RNA with Ribo-Zero Globin kit (Illumina, Inc., San Diego, CA) were used for globin-reduction and cDNA library preparation. The Illumina HiSeq 2500 sequencer was used to generate 75 bp reads with a mean of 20 million reads per sample. Skewer [21] was used to trim TruSeq adapter sequences.

Additional quality control was performed using FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and RNA-SeQC [22]. Reads were aligned to the human GRCh38 reference genome using STAR 2.5 [23]. Count data were adjusted for library depth and batch effects were removed using the limma removeBatchEffects function [24].

Microarray gene expression data

ECLIPSE whole blood microarray data were available at the initial enrollment visit. Details regarding microarray data collection and processing were previously published [25]. Total RNA was extracted using PAXgene Blood miRNA kits and hybridized to the Affymetrix Human Gene 1.1 ST array. Quality control metrics were implemented using the Bioconductor oligo [26] and RMA Express [27] packages. The Factor Analysis for Robust Microarray Summarization [28] package was utilized for background correction and normalization. Batch effects were removed using the limma removeBatchEffects function [24]. As some gene transcripts were represented by multiple probes, we chose the probe with the greatest interquartile range, as previously described [14, 15].

We then took additional steps to facilitate comparability across RNA-seq and microarray data technologies: (1) we limited transcripts to those present in both data sets based on HGNC symbols; (2) we scaled and centered all gene expression data to have a mean of zero and a standard deviation of 1.

Statistical analyses

Overview of study design

COPDGene was used as a discovery and testing cohort, and ECLIPSE was used for independent replication (Fig. 1). In COPDGene, risk score training was performed in half the participants with the other half of the participants used for risk score testing. We assessed the association of three transcriptome-based risk scores (see Predictors) with outcomes (see Outcomes) in the COPDGene testing sample. We attempted to replicate the risk score associations in ECLIPSE.

Predictors

We calculated the Herazo-Maya et al. [14, 15] IPF prognosis signature using transcriptomic data from COPDGene and ECLIPSE and applied this risk signature to the prediction of ILA and mortality in these cohorts. Two transcripts were not available in our cohorts (*C2ORF27A* and *SNHG1*), resulting in 50 genes available for testing. Briefly, calculating this risk score involves calculating the proportion of up- and down-regulated genes, summing the normalized expression values, calculating the product between the summed normalized expression values and proportion of decreased and increased genes, and comparing the up- and down- scores to the median for the population. If both the up- and down- scores are higher than the median of the population in which the score is being calculated, then the person is considered to be at high IPF mortality risk (1 = high IPF mortality risk, 0 = low IPF mortality risk).

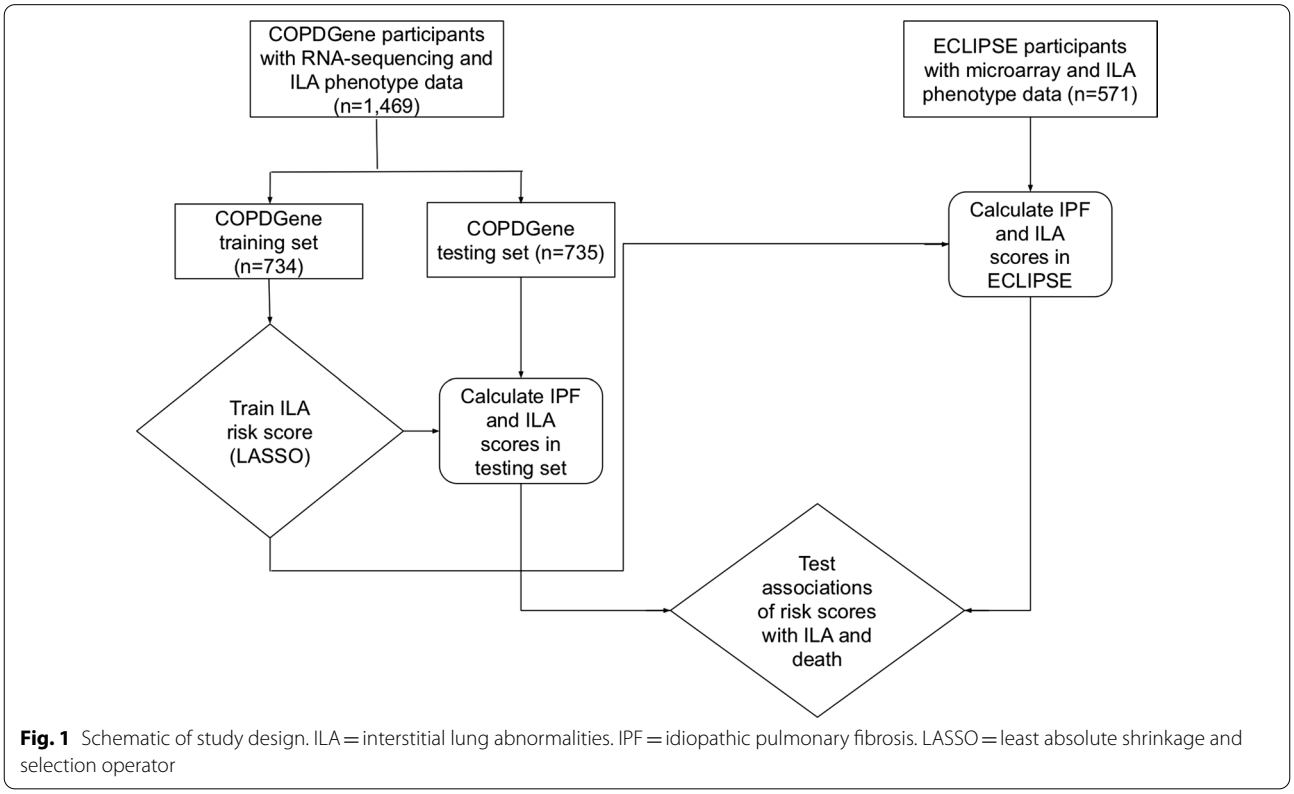

Fig. 1 Schematic of study design. ILA = interstitial lung abnormalities. IPF = idiopathic pulmonary fibrosis. LASSO = least absolute shrinkage and selection operator

To test the performance of these same genes using an alternative method, we developed a risk score optimized to the outcome of ILA by taking the following steps:

(1) We used the 50 available transcripts as inputs to construct a penalized regression (Least Absolute Shrinkage and Selection Operator (LASSO)) model in the COPDGene training set optimized to the outcome of ILA. LASSO regression shrinks coefficients toward zero to provide feature selection and minimizes collinearity amongst predictors. We tuned models within the COPDGene training set using fivefold cross-validation, optimizing the area-under-the-receiver-operating-curve (AUC) on the left-out fold.
(2) We used this penalized regression model (weights of gene transcripts) to calculate the log odds for ILA for each individual in the COPDGene testing sample and ECLIPSE.
(3) We scaled and centered the risk scores within the COPDGene testing and ECLIPSE datasets separately.

Thus, we evaluated two main predictors in this study: (1) the original IPF mortality risk score, and (2) a re-weighted risk score optimized to ILA as an outcome (hereafter, "ILA score [IPF transcripts]"). Results were presented as 1 standard deviation increases in risk scores. Each risk score was tested in the COPDGene testing set and the ECLIPSE replication cohort. We additionally compared the direction of effects of genes in these scores. We range-standardized both risk scores and plotted histograms of predicted probabilities for each respective outcome in the COPDGene testing set.

As an additional analysis, we created an ILA-optimized risk score using genome-wide transcripts (hereafter, "ILA score [all transcripts]"). Prior to training the ILA score [all transcripts], we limited our analyses to highly expressed transcripts (>1 count per million in 99% of samples), resulting in 9100 transcripts. These 9100 transcripts were used as inputs into LASSO regression. The model was tuned within the COPDGene training set using tenfold cross-validation, minimizing misclassification error on the left-out fold. We then calculated and standardized the risk scores above as for the ILA score [IPF transcripts]. We examined the correlation between the ILA score [all transcripts] and IPF score genes using Pearson correlation coefficients and considered those with Bonferroni-adjusted p-values to be significant. We tested the association of each risk score with outcomes in the COPDGene testing set and ECLIPSE.

We also performed differential gene expression analyses for ILA in the COPDGene training set using limma [24]; for this analysis, RNA-seq data were processed and normalized as described above except that we included transcripts with more than 1 count per million in half the samples. We considered a false discovery rate (FDR)-adjusted p-value level of 0.05 to be significant.

Outcomes

We examined two primary outcomes available in both cohorts: (1) prevalent ILA (at the time of blood sample collection), and (2) time-to-death. With respect to time-to-death, COPDGene participants were followed for up to 5 years after collection of RNA-seq data, and ECLIPSE participants for up to 8 years.

Models and model specifications

We used logistic regression to test the association of each of the 50 transcripts with ILA and compared the direction of effect for ILA to the transcript direction (up or down) in the IPF mortality score. We also compared the association of the 50 individual transcripts with all-cause mortality and compared the effect direction to both the effect for ILA association and direction in the original IPF mortality score. We assessed the association of each risk score (IPF score, ILA score [IPF transcripts], ILA score [all transcripts]) with ILA and time-to-death. Logistic regression was used to assess associations with ILA. We used Cox regression [29] to test associations with time-to-death (survival R package [30]). Models were adjusted for age, sex, race, body-mass index, pack-years of cigarette smoking, and current smoking status (at the time of blood sample collection). For time-to-death analyses, we also performed stratified analyses in ILA and non-ILA participants. For ILA, the Bonferroni-adjusted threshold was 0.05/3 risk scores/2 cohorts = 0.0083. For time-to-death analyses, the Bonferroni-adjusted threshold was 0.05/3 risk scores/2 cohorts/3 strata = 0.0028. Proportional hazards assumptions were evaluated with Schoenfeld residual plots and tests. To evaluate predictive performance, we performed AUC analyses using the pROC R package and c-indices [31] for survival models using the rms R package. As a sensitivity analysis, we further adjusted models for white blood cell differential counts to assess whether white cell counts attenuated the observed signals. To combine the signals between the two testing cohorts (COPDGene testing, and ECLIPSE), we performed fixed-effects inverse variance-weighted meta-analyses using the meta R package [32].

Mediation analyses

We noted that the associations of the ILA score [IPF transcripts] with time-to-death were substantially attenuated when age was added to models. Therefore, we performed causal mediation analyses to determine

whether the effects of age on mortality were mediated through this score. We used the medflex R package [33] to perform natural effects causal mediation analyses [34, 35] in the COPDGene testing set. We considered age as the exposure, and death (binary) as the outcome in logistic regression analyses. We considered the ILA score [IPF transcripts] as the mediator, and a p-value for the natural indirect effects less than 0.05 was considered significant. We also tested the association of the ILA score [IPF transcripts] with age using a Pearson correlation coefficient.

Characterization of IPF and ILA score genes

To gain biologic insight into the relationship between genes that comprised the risk scores, we examined Pearson correlation coefficients between each of the 50 transcripts and constructed a heatmap of correlation coefficients. The observed correlation structure allowed us to use the sigora R package [36] to perform pathway enrichment analyses for the 50 gene transcripts. We also evaluated how changing the number of genes in the ILA score [IPF transcripts] affected predictive performance (Additional file 1: Methods).

All analyses were performed in R version 4.0.3 (www.r-project.org). Normality for continuous variables was assessed by visual inspection of histograms. Results were reported as mean ± standard deviation or median [interquartile range], as appropriate. Continuous variables were compared with Student t-tests or Wilcoxon tests, and categorical variables were compared with analysis of variance (ANOVA) or Kruskal–Wallis tests, as appropriate. A p-value less than 0.05 was considered nominally significant, and p-values below Bonferroni-adjusted thresholds were considered significant.

Results

Characteristics of study participants

Figure 1 is a schematic of the study design. We included 1,469 COPDGene participants from the 5-year follow up visit with RNA-seq and visual scoring of ILA phenotype data, and 571 ECLIPSE participants with microarray and ILA phenotype data. Table 1 shows demographic characteristics and outcomes in the COPDGene training set (n = 734), COPDGene testing set (n = 735), and ECLIPSE (n = 571). Characteristics were similar across COPDGene training and testing sets. Compared to COPDGene, ECLIPSE participants were more likely to be male, were all European ancestry, had more pack-years of smoking, were less likely to be current smokers, were less likely to have ILA, and had a higher proportion of deaths.

Development of risk scores

Fifty out of the 52 genes from Herazo-Maya et al. [14, 15] were available in both cohorts (*C2ORF27A* and *SNHG1* were missing). First, using these 50 genes, we calculated the IPF score in COPDGene and ECLIPSE participants. We noted that when testing the associations of individual transcripts with ILA or death, 23 genes had a discordant direction of effect for ILA and 3 genes (*HLA-DP1*, *HLA-DP2*, *LPAR6*) had a discordant direction of effect for all-cause mortality compared to the directions of effects reported for the IPF score (Additional file 1: Table S1). Second, we used the 50 available genes to construct a LASSO penalized regression model in the COPDGene training set, which was optimized to the outcome of ILA (ILA score [IPF transcripts]). Additional file 1: Table S2 shows the ILA score [IPF transcripts] gene transcripts, beta coefficients and comparisons of directions of effects to the IPF risk score. Only three transcripts (*LBH*, *GBP4*, *BTN3A1*) were significantly associated with ILA, all of which were included in the ILA score [IPF transcripts].

Table 1 Characteristics of study participants

Characteristics	COPDGene training sample	COPDGene testing sample	ECLIPSE	p
n	734	735	571	
Age in years (mean (SD))	65.08 (8.53)	64.91 (8.60)	64.04 (6.10)	0.054
Sex (No. (%) female)	341 (46.5)	371 (50.5)	195 (34.2)	< 0.001
Race (No. (%) African American)	178 (24.3)	171 (23.3)	0 (0.0)	< 0.001
Pack-years of smoking (mean (SD))	43.71 (23.86)	42.38 (21.85)	47.27 (26.45)	0.002
Current smoking status (No. (%))	253 (34.5)	250 (34.1)	129 (22.6)	< 0.001
Outcomes				
ILA (No. (%)) *	122 (16.6)	126 (17.1)	37 (6.5)	< 0.001
Dead (No. (%))	84 (11.4)	73 (9.9)	107 (30.7)	< 0.001
Days followed (median [IQR])	1881.00 [1603.75, 2104.00]	1916.50 [1631.50, 2140.75]	2848.00 [1425.00, 2926.00]	< 0.001

ILA = interstitial lung abnormalities. *ILA prevalence in the overall population was 5.7% (586/10,364) in COPDGene and 1.3% (37/2,746) in ECLIPSE

The ILA score [IPF transcripts] included 11 transcripts (*BTN3A1, CPED1, CXCR6, GBP4, GPR174, IL7R, LBH, LPAR6, LRRC39, NAP1L2, PLBD1*), six of which had a discordant direction of effect with the IPF score (*BTN3A1, CPED1, GBP4, GPR174, LPAR6, NAP1L2*). The distribution of the ILA score [IPF transcripts] in both cohorts is shown in Additional file 1: Figure S1. We then plotted range-standardized predicted probabilities in the COPDGene testing set and observed overlapping but distinct distributions between the two risk scores (Additional file 1: Figure S2). We additionally used genome-wide transcripts to train a risk score to ILA (ILA score [all transcripts]); this score included 25 transcripts (Additional file 1: Table S3), none of which overlapped with the 50 genes in the IPF score. To examine correlation structure between the ILA score [all transcripts] and IPF score genes, we calculated correlation coefficients for 1200 unique combinations of risk score genes and observed that 547 transcripts were significantly correlated (Additional file 1: Figure S3). We then tested these risk scores (IPF score, ILA score [IPF transcripts], ILA score [all transcripts]) for association with ILA and time-to-death in the COPDGene testing set and ECLIPSE.

Association of risk scores with ILA

The IPF score was not associated with ILA in the COPDGene testing set or ECLIPSE (Table 2). The univariable associations of the ILA score [IPF transcripts] with ILA in both cohorts is shown in Fig. 2A. In multivariable models (Table 2), the ILA score [IPF transcripts] was associated with ILA in the COPDGene testing set and ECLIPSE (meta-analysis OR 1.4 [95% CI: 1.2–1.6], p = 6.8e-5). The multivariable model including clinical factors and the IPF score had AUCs of 0.64 in both the COPDGene testing set and ECLIPSE. The multivariable model including clinical factors and the ILA score [IPF transcripts] had AUCs of 0.65 in the COPDGene testing set and 0.68 in ECLIPSE, respectively. As a sensitivity analysis, we further adjusted models for white blood cell

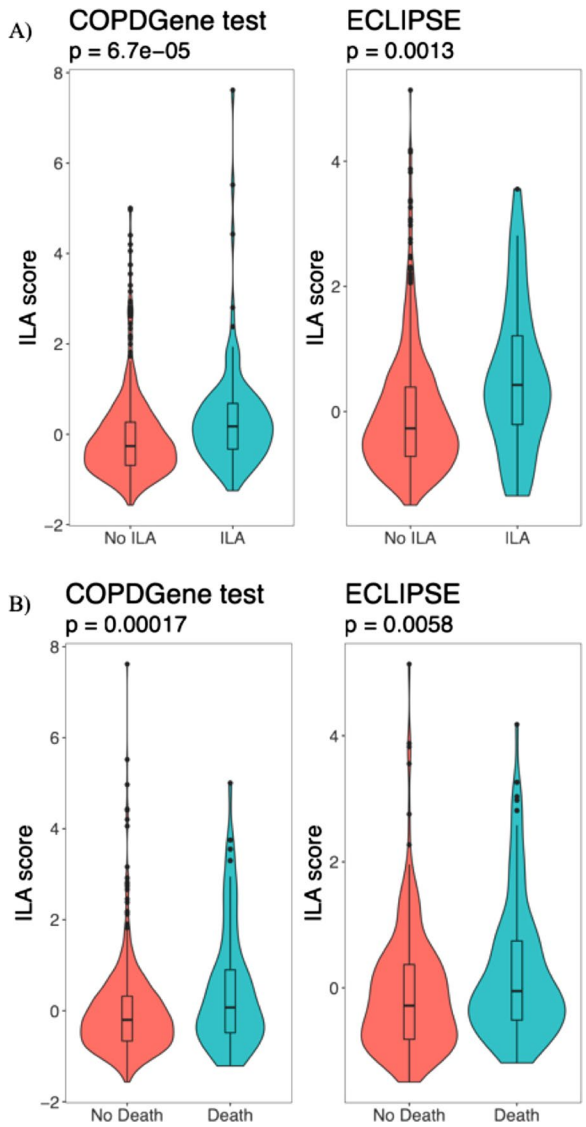

Fig. 2 Violin and box plots showing the unadjusted association of the ILA score [IPF transcripts] with ILA (A) and death (B) in the COPDGene test set (n = 735) and ECLIPSE (n = 571)

Table 2 Odds ratios of the IPF and ILA [IPF transcripts] risk scores with ILA in the COPDGene test set (n = 735) and ECLIPSE (n = 571)

Score	COPDGene				ECLIPSE				Combined (Meta-analyses)	
	Unadjusted		Adjusted		Unadjusted		Adjusted		Adjusted	
	OR (95% CI)	*p*	*OR (95% CI)*	*p*	*OR (95% CI)*	*p*	*OR (95% CI)*	*p*	*OR (95% CI)*	*p*
IPF score	1.8 (1.1–2.8)	0.018	1.7 (1.1–2.8)	0.023	0.64 (0.22–1.9)	0.41	0.63 (0.22–1.8)	0.4	1.5 (0.95—2.3)	0.083
ILA score [IPF transcripts]	1.4 (1.2–1.7)	6.70E-05	1.3 (1.1–1.6)	0.0019	1.5 (1.2–2)	0.001	1.4 (1.1–1.9)	0.011	1.4 (1.2—1.6)	6.80E-05

Logistic regression models were adjusted for age, sex, race, body-mass index, pack-years of smoking, and current smoking status. Inverse variance fixed-effects meta-analyses of the adjusted estimates were performed

ILA interstitial lung abnormalities, *IPF* idiopathic pulmonary fibrosis

counts, and observed similar results (Additional file 1: Table S4).

Association of risk scores with time-to-death

The IPF risk score was not associated with mortality in the COPDGene testing set or ECLIPSE (Table 3). The univariable associations of the ILA score [IPF transcripts] with death in both cohorts is shown in Fig. 2B. In multivariable analyses (Table 3), the ILA score [IPF transcripts] is associated with time-to-death in all participants, but not the ILA participants. Multivariable models with clinical risk factors and the IPF or ILA [IPF transcripts] scores performed similarly in terms of c-indices (Table 3). Meta-analyses demonstrate a significant association of the ILA score [IPF transcripts] with time-to-death (Fig. 3) in all (meta-analysis HR 1.25 [95% CI: 1.12–1.41], p = 1.25e-4) and non-ILA (meta-analysis HR 1.33 [95% CI: 1.17–1.52], p = 2.17e-5) participants. As a sensitivity analysis, we further adjusted models for white blood cell counts and observed similar results (Additional file 1: Table S5). In causal mediation analyses, we observed that the ILA score [IPF transcripts] mediated the effects of age on mortality (natural indirect effect p = 0.003; proportion mediated = 11.8% [95% CI: 4.04%-22.6%]). We also observed that the ILA score [IPF transcripts] was associated with age in both the COPDGene training and testing sets (Additional file 1: Figure S4).

We examined the ILA score [all transcripts] for association with ILA and all-cause mortality. We found that this new score was associated with ILA and all-cause mortality in the COPDGene testing set, but only with all-cause mortality in ECLIPSE (Additional file 1: Table S6).

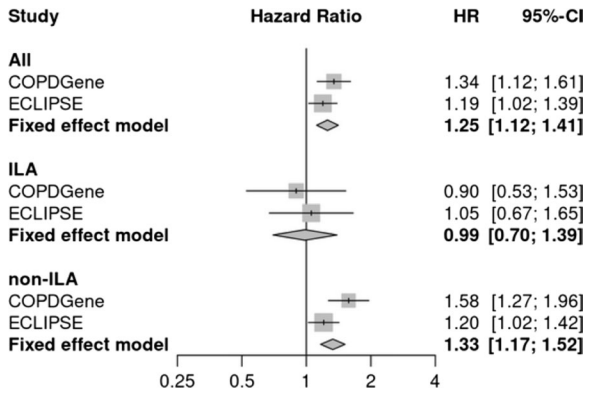

Fig. 3 Forest plots based on time-to-death analyses of the ILA score [IPF transcripts] in the COPDGene test set (n = 735) and ECLIPSE (n = 571). Cox proportional hazards models were fit, adjusting for age, sex, race, body-mass index, pack-years of smoking, and current smoking. Hazard ratios are shown as per standard deviation increase in the ILA score. ILA = interstitial lung abnormalities

In differential gene expression analyses in the COPDGene training set, no transcripts were associated with ILA at an FDR of 0.05.

Characterization of the gene signature

Having demonstrated that the genes in the IPF mortality score can be re-weighted to create an ILA score, and that this ILA score [IPF transcripts] consistently associates with ILA and time-to-death in two cohorts, we sought to understand how biological processes annotated to the IPF mortality genes relate to our outcomes. We observed that these 50 genes are highly correlated with each other

Table 3 Association of the IPF and ILA [IPF transcripts] scores with time-to-death in the COPDGene test set (n = 735) and ECLIPSE (n = 571)

Score	Cohort	Stratum	Unadjusted		Adjusted		
			HR (95% CI)	p	HR (95% CI)	p	c-index
IPF score	COPDGene test set	All	1.3 (0.71–2.3)	0.41	1.3 (0.7–2.3)	0.44	0.63
		ILA	1.3 (0.47–3.8)	0.6	1.2 (0.44–3.6)	0.68	0.66
		non-ILA	1.2 (0.57–2.4)	0.67	1.2 (0.58–2.4)	0.63	0.61
	ECLIPSE	All	1.3 (0.77–2.1)	0.36	1.2 (0.75–2)	0.4	0.56
		ILA	2.2 (0.49–10)	0.3	2.6 (0.54–13)	0.23	0.54
		non-ILA	1.2 (0.72–2.1)	0.47	1.2 (0.7–2)	0.54	0.56
ILA score [IPF transcripts]	COPDGene test set	All	1.4 (1.2–1.7)	3.60E-05	1.3 (1.1–1.6)	0.0012	0.64
		ILA	1 (0.64–1.6)	0.99	0.9 (0.53–1.5)	0.69	0.66
		non-ILA	1.6 (1.3–2)	2.50E-06	1.6 (1.3–2)	4.00E-05	0.62
	ECLIPSE	All	1.3 (1.1–1.4)	0.003	1.2 (1–1.4)	0.023	0.6
		ILA	1.2 (0.78–1.7)	0.47	1.1 (0.67–1.7)	0.82	0.52
		non-ILA	1.3 (1.1–1.5)	0.006	1.2 (1–1.4)	0.027	0.59

Multivariable Cox regression models were adjusted for age, sex, race, body-mass index, pack-years of smoking, and current smoking status. ILA = interstitial lung abnormalities. C-indices for multivariable models including clinical variables and corresponding risk score are shown in the right-hand column

and may represent coordinated biological processes. A heatmap of Pearson correlation coefficients is shown in Additional file 1: Figure S5. In pathway enrichment analyses, the 50 IPF score genes were associated with three pathways: Butyrophilin family interactions, P2Y receptors, and immunoregulatory interactions between lymphoid and non-lymphoid cells (Additional file 1: Table S7). The effects of varying levels of genes in the ILA score [IPF transcripts] on predictive performance for ILA are shown in Additional file 1: Results; Additional file 1: Table S8; Additional file 1: Figure S6).

Discussion

While there has been extensive research into gene expression changes associated with IPF [37–40], no studies, to our knowledge, have examined gene expression profiles in ILA. ILA may progress to pulmonary fibrosis in certain instances [4, 5], and IPF and ILA have demonstrated overlapping yet distinct genetic underpinnings [10]. In this study of over 2000 individuals with blood gene expression and ILA phenotype data, we examined the association of a blood IPF gene expression mortality score with ILA and all-cause mortality. We found that a previously described IPF score was not associated with ILA or mortality. By applying penalized regression to the IPF gene transcripts, we developed an 11-transcript ILA score. Six of these genes had discordant directions of effects compared to the IPF score, which may implicate important mechanisms regulating whether an individual develops ILA versus progressing to pulmonary fibrosis. While there may be some shared pathogenic mechanisms between ILA and IPF, the amount of discordance we observed further suggests that some of those with ILA (among populations of smokers) are likely distinct from IPF. Two ILA scores, derived from IPF score or genome-wide transcripts, were associated with all-cause mortality in both cohorts, suggesting that the transcripts relevant to ILA risk may represent general risk factors for mortality. We identified a peripheral blood signature of ILA, demonstrated overlapping and distinct gene transcripts between ILA and IPF, and lend insight into how gene expression profiles and biological pathways associated with IPF prognosis relate to ILA and all-cause mortality.

Our analyses may lend insight into biological processes relevant to ILA. We developed our risk score using LASSO, a method that reduces collinearity amongst features and optimizes prediction accuracy, but which does not necessarily choose the most biologically relevant features; thus, biological interpretation must be performed with caution. Despite this caveat, many of the genes identified as being predictive of ILA have been implicated in IPF pathogenesis, and the correlation structure amongst IPF risk score genes allowed us to perform pathway enrichment analyses and link the predictive transcripts to specific biological processes.

Five genes demonstrated concordant directions of effects between the ILA [IPF transcripts] and IPF scores *(CXCR6, IL7R, LBH, LRRC39, PLBD1)*, suggesting these genes may represent important biologic processes in promoting the progression of pulmonary fibrosis. *IL7R* had the largest absolute effect size in the ILA score [IPF transcripts] and was inversely associated with ILA. Our results are consistent with in vitro and in vivo molecular evidence demonstrating that IL-7 inhibits fibroblast TGF-ß production in pulmonary fibrosis [41]. *CXCR6* is a chemokine receptor that attracts T cells to the lungs [42] and can promote the epithelial-mesenchymal transition in cell lines [43], an event that may be important to the progression of pulmonary fibrosis. *LBH* is a transcription factor involved in Wnt signaling and is highly expressed in sub-populations of matrix fibroblasts from mouse lung [44]. Thus, there is already evidence that *IL7R*, *LBH,* and *CXCR6* are important for understanding processes that promote pulmonary fibrosis, and *LRRC39* and *PLBD1* are additional targets for future studies.

The transcripts with discordant directions of effects compared to the IPF score (*BTN3A1, GBP4, CPED1, GPR174, LPAR6, NAP1L2*) support the notion that ILA may represent a collection of disorders, some of which are distinct from IPF. This observation is consistent with genetic analyses of IPF and ILA. A genome-wide association study (GWAS) of ILA identified four genome-wide significant variants (in *FCF1P3, IPO11, HTR1E, MUC5B*); while the *MUC5B* rs35705950 promoter polymorphism is well known in IPF, the other three loci were not associated with IPF. Similarly, out of 12 previously reported IPF GWAS variants, only four other variants were significantly associated with ILA (*DPP9, DSP, FAM13A, IVD*), though most of the others were consistent in effect direction. Whether the genetic variants uniquely associated with ILA indirectly regulate the discordant genes identified in the current study is unclear and requires further investigation.

The transcripts with discordant directions of effects between the ILA [IPF transcripts] and IPF scores allude to divergent biologic processes that could play a role in determining whether a person develops ILA or progresses to irreversible fibrosis. *BTN3A1* is part of the Butyrophilin immunoglobulin superfamily. Butyrophilins may play a role in facilitating interactions between adaptive and innate immune cells [45]. *GBP4* is a guanylate binding protein that facilitates second messenger signaling for interferons, and has been observed to increase in response to cytokine stimulation in IPF lungs [46]. Genetic variation in *GPR174* is associated with susceptibility to autoimmune disease,

and *GPR174*-deficient mice were resistant to lipopolysaccharide-induced cytokine storm [47]. *NAP1L2* promotes histone acetylation during neuronal differentiation [48] and other members of the same protein family have demonstrated increased expression in lung fibroblasts from IPF patients [49]. To determine whether peripheral blood cells might be explaining a substantial portion of our associations, we adjusted models for measured white cell counts and observed similar results. These results suggest that peripheral blood cell counts do not explain our observed associations. Taken together, these data suggest that immune responses to environmental/infectious insults, genetic variation, and epigenetic modifications may be important to understanding the distinct pathogenic mechanisms present in ILA that may lead to IPF.

We observed that the ILA score [IPF transcripts] was associated with all-cause mortality, but that this association was driven by non-ILA participants, albeit with small samples sizes for ILA. In single gene association analyses for the 50 IPF score genes, nearly half of all transcripts had discordant effect directions for ILA association compared to associations with all-cause mortality and in comparison to gene weights in the IPF score. Conversely, the 50 IPF gene effect directions in association with all-cause mortality (in all study participants) were highly concordant with the directions of weights in the IPF score with only three genes having discordant effect directions. Further, an ILA score trained using genome-wide transcripts in the COPDGene training data was not associated with ILA in both the COPDGene testing set and ECLIPSE but was associated with all-cause mortality in both cohorts. These data suggest that transcripts associated with ILA risk may also be risk factors for all-cause mortality. When training a gene expression model to ILA, we identified a gene set associated with all-cause mortality; however, an ILA score derived from the limited set of IPF score genes is associated with ILA and mortality. These observations suggest that the IPF score represents a mixture of transcripts relevant to all-cause mortality and the ILA-IPF axis.

LPAR6 was the only non-HLA transcript with concordant directions of effects when tested for association with ILA and all-cause mortality and had discordant directions of effects compared to the IPF score. *LPAR6* is a lysophosphatidic acid (LPA) receptor. LPA has been shown to signal through its G-protein-coupled receptors to induce pro-inflammatory signals from stressed epithelial cells and activate TGF signaling [50]. Thus, LPAR6 could represent a shared mechanism between ILA and all-cause mortality or it could be on the causal pathway between ILA and death.

As aging is a major driver of mortality, we sought to determine whether the effects of the ILA score [IPF transcripts] on mortality represent an aging effect. In causal mediation analyses, we found that about 12% of the effect of age on mortality was mediated through the ILA score [IPF transcripts]. Thus, the 11 genes in the ILA score [IPF transcripts] might be important for overall mortality and partially represent aging effects on mortality. We must interpret our mediation results with caution as the ILA score [IPF transcripts] was associated with age, which makes it unclear whether the observed gene expression changes lead to aging or vice versa; this issue alludes to an assumption of causal mediation analyses that there are no intertwined causal pathways (i.e. "identifiability condition"). Determining which gene expression changes are caused by versus causative of aging requires further investigation. Further, the majority of aging effects are not captured by the ILA score [IPF transcripts], which highlights the need for further research into the shared and divergent biological processes related to ILA, aging, and mortality.

Strengths of this study include replication in two well-characterized cohorts of smokers, cross-technology replication (both RNA-seq and microarray), and the application of causal inference analyses to examine the relationship between IPF and ILA transcriptomic risk with aging and risk of death. Despite the statistical significance of the odds ratio of 1.4 and AUC of 0.65–0.68, the clinical impact is likely low. Even large odds ratios can have poor discriminative performance. Quantifying the clinical impact of risk scores in disease screening, diagnosis, and prevention is a complex issue [51, 52]. Understanding the implications for using our score as a clinical prediction tool was outside the scope of this study. We were not able to assess the association of the ILA risk scores with the development or progression of ILA due to data availability. Our primary analysis was of genes identified in a study of IPF, which allowed an assessment of the contrasts between IPF mortality and ILA risk in a well-validated set of genes. Notably, attempting to use a larger set of genes in the blood transcriptome, including other transcripts that may be more predictive of ILA risk, was not superior. The reasons that the ILA score [all transcripts] performed worse than the ILA score [IPF transcripts] are likely similar to the reasons we did not observe genes differentially expressed with ILA; the reason for the lack of differential gene expression for ILA is multifactorial and is attributable to a combination of limited statistical power, phenotypic heterogeneity of ILA in smokers, and poor reflection of ILA disease processes in peripheral blood gene expression. While our study is consistent with a limited signal for ILA in peripheral blood gene expression, we acknowledge that a wider

range of machine learning and prediction methodologies could improve predictive performance of the risk score. The correlation between the ILA score [all transcripts] and IPF score genes suggests an apparent lack of specificity of the prediction genes and reinforces the need to approach biological interpretations with caution. We were not able to assess the impact of technical factors (e.g. globin-versus poly-A-reduced RNA-seq, harmonizing gene expression platforms) on risk score construction and performance; improved risk scores and biological insights might be attainable by standardizing data collection and processing methods across cohorts. Other approaches using blood gene expression and additional -Omics data (such as methylation, proteomic, microbiome) may allow us to delve deeper into the mechanisms underlying our observed associations, as well as longitudinal studies of ILA and IPF.

In conclusion, a peripheral blood gene signature associated with IPF mortality was not associated with ILA or mortality in two well-characterized cohorts of smokers. An ILA gene expression score, derived from the genes in the IPF score, was reproducibly associated with ILA and all-cause mortality in current and former smokers. Separately, an ILA score derived from genome-wide transcripts was associated with all-cause mortality, but not ILA in validation studies. Approximately half of the genes in the ILA score [IPF transcripts] were of opposite direction in the IPF score, and genes associated with ILA may also be risk factors for mortality and partially represent aging effects on mortality. Genes identified in this study may be important candidates to further examine in the pathogenesis and progression of ILA to IPF and in mortality.

COPDGene.® Investigators – Core Units

Administrative Center: James D. Crapo, MD (PI); Edwin K. Silverman, MD, PhD (PI); Barry J. Make, MD; Elizabeth A. Regan, MD, PhD.

Genetic Analysis Center: Terri Beaty, PhD; Ferdouse Begum, PhD; Peter J. Castaldi, MD, MSc; Michael Cho, MD; Dawn L. DeMeo, MD, MPH; Adel R. Boueiz, MD; Marilyn G. Foreman, MD, MS; Eitan Halper-Stromberg; Lystra P. Hayden, MD, MMSc; Craig P. Hersh, MD, MPH; Jacqueline Hetmanski, MS, MPH; Brian D. Hobbs, MD; John E. Hokanson, MPH, PhD; Nan Laird, PhD; Christoph Lange, PhD; Sharon M. Lutz, PhD; Merry-Lynn McDonald, PhD; Margaret M. Parker, PhD; Dmitry Prokopenko, Ph.D; Dandi Qiao, PhD; Elizabeth A. Regan, MD, PhD; Phuwanat Sakornsakolpat, MD; Edwin K. Silverman, MD, PhD; Emily S. Wan, MD; Sungho Won, PhD.

Imaging Center: Juan Pablo Centeno; Jean-Paul Charbonnier, PhD; Harvey O. Coxson, PhD; Craig J. Galban, PhD; MeiLan K. Han, MD, MS; Eric A. Hoffman, Stephen Humphries, PhD; Francine L. Jacobson, MD, MPH; Philip F. Judy, PhD; Ella A. Kazerooni, MD; Alex Kluiber; David A. Lynch, MB; Pietro Nardelli, PhD; John D. Newell, Jr., MD; Aleena Notary; Andrea Oh, MD; Elizabeth A. Regan, MD, PhD; James C. Ross, PhD; Raul San Jose Estepar, PhD; Joyce Schroeder, MD; Jered Sieren; Berend C. Stoel, PhD; Juerg Tschirren, PhD; Edwin Van Beek, MD, PhD; Bram van Ginneken, PhD; Eva van Rikxoort, PhD; Gonzalo Vegas Sanchez-Ferrero, PhD; Lucas Veitel; George R. Washko, MD; Carla G. Wilson, MS;

PFT QA Center, Salt Lake City, UT: Robert Jensen, PhD.

Data Coordinating Center and Biostatistics, National Jewish Health, Denver, CO: Douglas Everett, PhD; Jim Crooks, PhD; Katherine Pratte, PhD; Matt Strand, PhD; Carla G. Wilson, MS.

Epidemiology Core, University of Colorado Anschutz Medical Campus, Aurora, CO: John E. Hokanson, MPH, PhD; Gregory Kinney, MPH, PhD; Sharon M. Lutz, PhD; Kendra A. Young, PhD.

Mortality Adjudication Core: Surya P. Bhatt, MD; Jessica Bon, MD; Alejandro A. Diaz, MD, MPH; MeiLan K. Han, MD, MS; Barry Make, MD; Susan Murray, ScD; Elizabeth Regan, MD; Xavier Soler, MD; Carla G. Wilson, MS.

Biomarker Core: Russell P. Bowler, MD, PhD; Katerina Kechris, PhD; Farnoush Banaei-Kashani, Ph.D **COPDGene® Investigators – Clinical Centers.**

Ann Arbor VA: Jeffrey L. Curtis, MD; Perry G. Pernicano, MD.

Baylor College of Medicine, Houston, TX: Nicola Hanania, MD, MS; Mustafa Atik, MD; Aladin Boriek, PhD; Kalpatha Guntupalli, MD; Elizabeth Guy, MD; Amit Parulekar, MD;

Brigham and Women's Hospital, Boston, MA: Dawn L. DeMeo, MD, MPH; Alejandro A. Diaz, MD, MPH; Lystra P. Hayden, MD; Brian D. Hobbs, MD; Craig Hersh, MD, MPH; Francine L. Jacobson, MD, MPH; George Washko, MD.

Columbia University, New York, NY: R. Graham Barr, MD, DrPH; John Austin, MD; Belinda D'Souza, MD; Byron Thomashow, MD.

Duke University Medical Center, Durham, NC: Neil MacIntyre, Jr., MD; H. Page McAdams, MD; Lacey Washington, MD.

Grady Memorial Hospital, Atlanta, GA: Eric Flenaugh, MD; Silanth Terpenning, MD

HealthPartners Research Institute, Minneapolis, MN: Charlene McEvoy, MD, MPH; Joseph Tashjian, MD.

Johns Hopkins University, Baltimore, MD: Robert Wise, MD; Robert Brown, MD; Nadia N. Hansel, MD,

MPH; Karen Horton, MD; Allison Lambert, MD, MHS; Nirupama Putcha, MD, MHS.

Lundquist Institute for Biomedical Innovationat Harbor UCLA Medical Center, Torrance, CA: Richard Casaburi, PhD, MD; Alessandra Adami, PhD; Matthew Budoff, MD; Hans Fischer, MD; Janos Porszasz, MD, PhD; Harry Rossiter, PhD; William Stringer, MD.

Michael E. DeBakey VAMC, Houston, TX: Amir Sharafkhaneh, MD, PhD; Charlie Lan, DO.

Minneapolis VA: Christine Wendt, MD; Brian Bell, MD; Ken M. Kunisaki, MD, MS.

National Jewish Health, Denver, CO: Russell Bowler, MD, PhD; David A. Lynch, MB.

Reliant Medical Group, Worcester, MA: Richard Rosiello, MD; David Pace, MD.

Temple University, Philadelphia, PA: Gerard Criner, MD; David Ciccolella, MD; Francis Cordova, MD; Chandra Dass, MD; Gilbert D'Alonzo, DO; Parag Desai, MD; Michael Jacobs, PharmD; Steven Kelsen, MD, PhD; Victor Kim, MD; A. James Mamary, MD; Nathaniel Marchetti, DO; Aditi Satti, MD; Kartik Shenoy, MD; Robert M. Steiner, MD; Alex Swift, MD; Irene Swift, MD; Maria Elena Vega-Sanchez, MD.

University of Alabama, Birmingham, AL: Mark Dransfield, MD; William Bailey, MD; Surya P. Bhatt, MD; Anand Iyer, MD; Hrudaya Nath, MD; J. Michael Wells, MD.

University of California, San Diego, CA: Douglas Conrad, MD; Xavier Soler, MD, PhD; Andrew Yen, MD.

University of Iowa, Iowa City, IA: Alejandro P. Comellas, MD; Karin F. Hoth, PhD; John Newell, Jr., MD; Brad Thompson, MD.

University of Michigan, Ann Arbor, MI: MeiLan K. Han, MD MS; Ella Kazerooni, MD MS; Wassim Labaki, MD MS; Craig Galban, PhD; Dharshan Vummidi, MD.

University of Minnesota, Minneapolis, MN: Joanne Billings, MD; Abbie Begnaud, MD; Tadashi Allen, MD.

University of Pittsburgh, Pittsburgh, PA: Frank Sciurba, MD; Jessica Bon, MD; Divay Chandra, MD, MSc; Joel Weissfeld, MD, MPH.

University of Texas Health, San Antonio, San Antonio, TX: Antonio Anzueto, MD; Sandra Adams, MD; Diego Maselli-Caceres, MD; Mario E. Ruiz, MD; Harjinder Singh.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12931-022-02077-8.

Additional file 1. Supplementary appendix.

Acknowledgements

We thank the patients in the COPDGene and ECLIPSE studies for their commitment to helping others and contributing their valuable time and blood to these studies.

Author contributions

Study Design: MM, BDH, GMH, MHC. Acquisition, analysis, or interpretation of the data: MM, AG, DDS, PJC, BDH, EKS, JQ, HH, TH, AH, MN, GMH, MHC. Critical revision of the manuscript for important intellectual content. Statistical analysis: MM, BDH, AH, GMH, MHC. Obtained funding: EKS, GMH, MHC. All authors read and approved the final manuscript.

Funding

MM and AJG are supported by T32HL007427 and K08HL159318.
BDH is supported by NIH K08HL136928 and U01 HL089856.
CPH is supported by NIH R01HL130512, R01HL157879 and P01HL114501.
PJC is supported by NIH R01HL124233 and R01HL147326.
EKS is supported by NIH R01 HL137927, R01 HL147148, U01 HL089856, R01 HL133135, R01 HL152728, P01 HL132825, and P01 HL114501.
HH is supported by NIH R01CA203636, 5U01CA209414, NIH/NHLBI 2R01HL111024, NIH R01HL135142, and NIH/NHLBI 1R01HL130974.
MN is supported by R01CA203636, U01CA209414, R01HL111024.
GMH is supported by NIH R01 HL111024, R01 HL130974, and R0135142.
MHC is supported by NIH R01HL137927, R01HL135142, HL147148, and HL089856.
The COPDGene project described was supported by Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health. COPDGene is also supported by the COPD Foundation through contributions made to an Industry Advisory Board that has included AstraZeneca, Bayer Pharmaceuticals, Boehringer Ingelheim, Genentech, GlaxoSmithKline, Novartis, Pfizer, and Sunovion. The ECLIPSE study (NCT00292552; GSK code SCO104960) was funded by GlaxoSmithKline.

Availability of data and materials

Data are available in the Gene Expression Omnibus accession numbers GSE158699 (COPDGene) and GSE76705 (ECLIPSE).

Declarations

Ethics approval and consent to participate

All COPDGene and ECLIPSE study participants provided written informed consent. Each study center obtained institutional review board (IRB) approval (Mass General Brigham IRB protocol 2007P000554).

Consent for publication

Not applicable.

Competing interests

MM received grant support from Bayer. EKS received grant support from GlaxoSmithKline and Bayer. CPH reports grant support from Boehringer-Ingelheim, Novartis, Bayer and Vertex, outside of this study. MN has received research grants to the institution from AstraZeneca, Daiichi Sankyo, Canon Medical Systems, and has served as a consultant to AstraZeneca, and Daiichi Sankyo. PJC has received grant support from GlaxoSmithKline and Bayer and consulting fees from GlaxoSmithKline and Novartis. RTS was employed by GlaxoSmithKline until 2019 and is a shareholder of GlaxoSmithKline stock; RTS received consulting fees from Immunomet, ENA Respiratory, Teva, and Vocalis Health; she is currently a Board Member of ENA Respiratory on behalf of the COPD Foundation. JQ is on the scientific advisory boards of Caris Life Sciences and Renalytix. HH received grants from Canon Medical Systems Inc and Konica-Minolta Inc. and served as consultant for Mitsubishi Chemical Co and Canon Medical Systems Inc outside of this study. GMH has performed consulting work for Boehringer-Ingelheim, and the Gerson Lehrman Group within the last 3 years. MHC has received grant support from GlaxoSmithKline and Bayer, consulting fees from Genentech and AstraZeneca, and speaking fees from Illumina.

Author details
[1]Channing Division for Network Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA. [2]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA. [3]Harvard Medical School, Boston, MA 02115, USA. [4]Department of Radiology, Center for Pulmonary Functional Imaging, Brigham and Women's Hospital, Boston, MA 02115, USA. [5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA. [6]Division of General Internal Medicine and Primary Care, Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, Canada. [7]Centre for Heart Lung Innovation, St. Paul's Hospital, and Department of Medicine (Respiratory Division), University of British Columbia, Vancouver, BC, Canada. [8]COPD Foundation, Washington, D.C., USA.

Received: 31 January 2022 Accepted: 3 June 2022
Published online: 17 June 2022

References

1. Washko GR, Hunninghake GM, Fernandez IE, et al. Lung volumes and emphysema in smokers with interstitial lung abnormalities. N Engl J Med. 2011;364:897–906.
2. Hunninghake GM, Hatabu H, Okajima Y, et al. MUC5B promoter polymorphism and interstitial lung abnormalities. N Engl J Med. 2013;368:2192–200.
3. Hatabu H, Hunninghake GM, Richeldi L, et al. Interstitial lung abnormalities detected incidentally on CT: a Position Paper from the Fleischner Society. Lancet Respir Med. 2020;8:726–37.
4. Putman RK, Gudmundsson G, Axelsson GT, et al. Imaging Patterns Are Associated with Interstitial Lung Abnormality Progression and Mortality. Am J Respir Crit Care Med. 2019;200:175–83.
5. Araki T, Putman RK, Hatabu H, et al. Development and Progression of Interstitial Lung Abnormalities in the Framingham Heart Study. Am J Respir Crit Care Med. 2016;194:1514–22.
6. Doyle TJ, Washko GR, Fernandez IE, et al. Interstitial lung abnormalities and reduced exercise capacity. Am J Respir Crit Care Med. 2012;185:756–62.
7. Lederer DJ, Enright PL, Kawut SM, et al. Cigarette smoking is associated with subclinical parenchymal lung disease: the Multi-Ethnic Study of Atherosclerosis (MESA)-lung study. Am J Respir Crit Care Med. 2009;180:407–14.
8. Tsushima K, Sone S, Yoshikawa S, Yokoyama T, Suzuki T, Kubo K. The radiological patterns of interstitial change at an early phase: over a 4-year follow-up. Respir Med. 2010;104:1712–21.
9. Putman RK, Hatabu H, Araki T, et al. Association between interstitial lung abnormalities and all-cause mortality. JAMA. 2016;315:672–81.
10. Hobbs BD, Putman RK, Araki T, et al. Overlap of genetic risk between interstitial lung abnormalities and idiopathic pulmonary fibrosis. Am J Respir Crit Care Med. 2019;200:1402–13.
11. Putman RK, Gudmundsson G, Araki T, et al. The MUC5B promoter polymorphism is associated with specific interstitial lung abnormality subtypes. Eur Respir J. 2017. https://doi.org/10.1183/13993003.00537-2017.
12. Baumgartner KB, Samet JM, Stidley CA, Colby TV, Waldron JA. Cigarette smoking: a risk factor for idiopathic pulmonary fibrosis. Am J Respir Crit Care Med. 1997;155:242–8.
13. Parker MM, Chase RP, Lamb A, et al. RNA sequencing identifies novel non-coding RNA and exon-specific effects associated with cigarette smoking. BMC Med Genomics. 2017;10:58.
14. Herazo-Maya JD, Noth I, Duncan SR, et al. Peripheral blood mononuclear cell gene expression profiles predict poor outcome in idiopathic pulmonary fibrosis. Sci Transl Med. 2013;5:205ra136.
15. Herazo-Maya JD, Sun J, Molyneaux PL, 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. 2017;5:857–68.
16. Regan EA, Hokanson JE, Murphy JR, et al. Genetic Epidemiology of COPD (COPDGene) Study Design. COPD J Chronic Obstr Pulm Dis. 2011;7:32–43.
17. Vestbo J, Anderson W, Coxson HO, et al. Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE). Eur Respir J. 2008;31:869–73.
18. Hatabu H, Hunninghake GM, Lynch DA. Interstitial lung abnormality: recognition and perspectives. Radiology. 2019;291:1–3.
19. Hata A, Schiebler ML, Lynch DA, Hatabu H. Interstitial lung abnormalities: state of the art. Radiology. 2021;34:204367.
20. Hino T, Lee KS, Han J, Hata A, Ishigami K, Hatabu H. Spectrum of pulmonary fibrosis from interstitial lung abnormality to usual interstitial pneumonia: importance of identification and quantification of traction bronchiectasis in patient management. Korean J Radiol. 2021;22:811–28.
21. Jiang H, Lei R, Ding S-W, Zhu S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics. 2014;15:182.
22. DeLuca DS, Levin JZ, Sivachenko A, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics. 2012;28:1530–2.
23. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. Curr Protoc Bioinforma. 2015;51:11141–111419.
24. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43: e47.
25. Obeidat M, Nie Y, Chen V, et al. Network-based analysis reveals novel gene signatures in peripheral blood of patients with chronic obstructive pulmonary disease. Respir Res. 2017;18:72.
26. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. Bioinformatics. 2010;26:2363–7.
27. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4:249–64.
28. Hochreiter S, Clevert D-A, Obermayer K. A new summarization method for Affymetrix probe level data. Bioinformatics. 2006;22:943–9.
29. Cox DR. Regression models and life tables. J R Stat Soc Ser B. 1972. https://doi.org/10.1079/IVPt200454()IN.
30. T T. A Package for Survival Analysis in R. R package version 3.1–12. https://CRANR-project.org/package=survival 2020.
31. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361–87.
32. Gan S. meta: An R package for meta-analysis. R News. 2007;7:40–5.
33. Johan S, Tom L, Beatrijs MSV. medflex: An R Package for Flexible Mediation Analysis using Natural Effect Models. J Stat Softw. 2017;76:1–46.
34. Lange T, Vansteelandt S, Bekaert M. A simple unified approach for estimating natural direct and indirect effects. Am J Epidemiol. 2012;176:190–5.
35. Vansteelandt S, Bekaert M, Lange T. Imputation strategies for the estimation of natural direct and indirect effects. Epidemiol Method. 2012;1:7.
36. Foroushani ABK, Brinkman FSL, Lynn DJ. Pathway-GPS and SIGORA: identifying relevant pathways based on the over-representation of their gene-pair signatures. PeerJ. 2013;1: e229.
37. Kaminski N, Allard JD, Pittet JF, et al. Global analysis of gene expression in pulmonary fibrosis reveals distinct programs regulating lung inflammation and fibrosis. Proc Natl Acad Sci U S A. 2000;97:1778–83.
38. Huang Y, Oldham JM, Ma S-F, et al. Blood transcriptomic predicts progression of pulmonary fibrosis and associates natural killer cells. Am J Respir Crit Care Med. 2021. https://doi.org/10.1164/rccm.202008-3093OC.
39. DePianto DJ, Chandriani S, Abbas AR, et al. Heterogeneous gene expression signatures correspond to distinct lung pathologies and biomarkers of disease severity in idiopathic pulmonary fibrosis. Thorax. 2015;70:48–56.
40. Reyfman PA, Walter JM, Joshi N, et al. Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. Am J Respir Crit Care Med. 2019;199:1517–36.
41. Huang M, Sharma S, Zhu LX, et al. IL-7 inhibits fibroblast TGF-beta production and signaling in pulmonary fibrosis. J Clin Invest. 2002;109:931–7.
42. Morgan AJ, Guillen C, Symon FA, et al. Expression of CXCR6 and its ligand CXCL16 in the lung in health and disease. Clin Exp Allergy. 2005;35:1572–80.
43. Ma Z, Ma C, Zhang Q, et al. Role of CXCL16 in BLM-induced epithelial-mesenchymal transition in human A549 cells. Respir Res. 2021;22:42.
44. Xie T, Wang Y, Deng N, et al. Single-cell deconvolution of fibroblast heterogeneity in mouse pulmonary fibrosis. Cell Rep. 2018;22:3625–40.
45. Miles T, Hoyne GF, Knight DA, Fear MW, Mutsaers SE, Prêle CM. The contribution of animal models to understanding the role of the immune

system in human idiopathic pulmonary fibrosis. Clin Transl Immunol. 2020;9: e1153.
46. Kass DJ, Yu G, Loh KS, et al. Cytokine-like factor 1 gene expression is enriched in idiopathic pulmonary fibrosis and drives the accumulation of CD4+ T cells in murine lungs: evidence for an antifibrotic role in bleomycin injury. Am J Pathol. 2012;180:1963–78.
47. Qiu D, Chu X, Hua L, et al. Gpr174-deficient regulatory T cells decrease cytokine storm in septic mice. Cell Death Dis. 2019;10:233.
48. Attia M, Rachez C, De Pauw A, Avner P, Rogner UC. Nap1l2 promotes histone acetylation activity during neuronal differentiation. Mol Cell Biol. 2007;27:6093–102.
49. Plantier L, Renaud H, Respaud R, Marchand-Adam S, Crestani B. Transcriptome of cultured lung fibroblasts in idiopathic pulmonary fibrosis: meta-analysis of publically available microarray datasets reveals repression of inflammation and immunity pathways. Int J Mol Sci. 2016. https://doi.org/10.3390/ijms17122091.
50. Ninou I, Magkrioti C, Aidinis V. Autotaxin in pathophysiology and pulmonary fibrosis. Front Med. 2018;5:180.
51. Wald NJ, Morris JK, Rish S. The efficacy of combining several risk factors as a screening test. J Med Screen. 2005;12:197–201.
52. Wald NJ, Old R. The illusion of polygenic disease risk prediction. Genet Med. 2019;21:1705–7.

Publisher's Note