RESEARCH ARTICLE

A robust data-driven genomic signature for idiopathic pulmonary fibrosis with applications for translational model selection

Ron Ammar[1]*, **Pitchumani Sivakumar**[2], **Gabor Jarai**[2], **John Ryan Thompson**[1]

1 Translational Bioinformatics, Translational Medicine, Bristol-Myers Squibb, Princeton, NJ, United States of America, **2** Fibrosis, Translational Research & Development, Bristol-Myers Squibb, Princeton, NJ, United States of America

* ron.ammar@bms.com

OPEN ACCESS

Citation: Ammar R, Sivakumar P, Jarai G, Thompson JR (2019) A robust data-driven genomic signature for idiopathic pulmonary fibrosis with applications for translational model selection. PLoS ONE 14(4): e0215565. https://doi.org/10.1371/journal.pone.0215565

Editor: Antje Prasse, Medizinische Hochschule Hannover, GERMANY

Received: June 4, 2018

Accepted: April 4, 2019

Published: April 18, 2019

Data Availability Statement: All relevant data have been published previously and are available in the NCBI Gene Expression Omnibus. Accessions are provided in the manuscript (see Table 1).

Funding: The funder, Bristol Myers Squibb Co., provided support in the form of salaries for authors R.A., P.S., G.J. and J.R.T., but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Abstract

Idiopathic pulmonary fibrosis (IPF) is a chronic and progressive lung disease affecting ~5 million people globally. We have constructed an accurate model of IPF disease status using elastic net regularized regression on clinical gene expression data. Leveraging whole transcriptome microarray data from 230 IPF and 89 control samples from *Yang et al*. (2013), sourced from the Lung Tissue Research Consortium (LTRC) and National Jewish Health (NJH) cohorts, we identify an IPF gene expression signature. We performed optimal feature selection to reduce the number of transcripts required by our model to a parsimonious set of 15. This signature enables our model to accurately separate IPF patients from controls. Our model outperforms existing published models when tested with multiple independent clinical cohorts. Our study underscores the utility of elastic nets for gene signature/panel selection which can be used for the construction of a multianalyte biomarker of disease. We also filter the gene sets used for model input to construct a model reliant on secreted proteins. Using this approach, we identify the preclinical bleomycin rat model that is most congruent with human disease at day 21 post-bleomycin administration, contrasting with earlier timepoints suggested by other studies.

Introduction

Idiopathic Pulmonary Fibrosis (IPF) is a fatal disease of unknown etiology characterized by scarring of the lung parenchyma resulting in progressive loss of lung function and eventual death [1]. Although two recently approved drugs, pirfenidone and nintedanib, reduce lung function decline in IPF, their efficacy is limited and mechanism of action poorly understood [2–4]. Even though meta analyses of large clinical trials suggest that pirfenidone reduces risk of mortality [5], lung transplant still remains the only option to significantly prolong survival in IPF, suggesting a dire need for new therapies. Development of new drugs for IPF is extremely challenging due to complicated diagnosis, limited disease understanding, lack of robust pre-clinical models predictive of human disease as well as biomarkers of disease progression and drug treatment. Current diagnosis of IPF requires careful integration of

Competing interests: All authors are employed by Bristol Myers Squibb Co. This commercial affiliation does not alter our adherence to PLOS ONE policies on sharing data and materials.

radiographic findings (honeycombing and presence of fibroblast foci), lung function (FVC, FEV1 and 6-minute walk test) and clinical data and the rational exclusion of other potentially similar interstitial lung diseases [6]. Often, the disease is diagnosed at an advanced stage when it is refractory to treatment. Therefore, there is a pressing need to develop newer, less-invasive and robust methods to efficiently diagnose IPF and enable early intervention strategies. Transcriptomic and proteomic disease signatures generated from clinically-relevant human samples including tissue and plasma, combined with robust in silico modeling can enable translational disease understanding, diagnosis and stratification of patients for effective drug treatments. Several studies have utilized microarray profiling of IPF-patient derived lung tissue to define genes and/or pathways that are differentially-regulated in comparison to healthy controls or patients with other lung diseases [4,7–9] and define signatures for disease classification. Peripheral blood profiling across small cohorts of patients have also identified potential biomarkers of disease such as MMP1 and MMP7 [10–12].

Comparative gene expression profiles of preclinical models of fibrosis with human tissue derived profiles have provided useful information on the utility of the models as well as insights into pathways or mechanisms that are altered during the induction, progression and resolution of fibrosis [13,14]. In many of these studies, gene/protein expression profiles have been correlated to clinical diagnosis, disease severity and measures of lung function [15].

In the most commonly studied preclinical model of IPF, the chemotherapeutic antibiotic bleomycin is intratracheally injected into rodents to induce an inflammatory response in the lung, damaging the epithelium, activating fibroblasts and ultimately leading to a fibrotic phase of increased collagen deposition and loss of alveolar structures [16]. The induced fibrosis manifests over the course of 7–14 days post-bleomycin treatment, and several studies have suggested different time points where congruence between the model and IPF are highest [16]. *Chaudhary et al.* (2006) measured profibrotic gene expression including pro-collagen I, TGF-β1, fibronectin and collagen deposition, determining the fibrotic phase to begin between days 9 and 14 post bleomycin treatment. *Bauer et al.* (2015) found the most congruent rat bleomycin model (day post-treatment) by first extracting a differential-expression signature from the rat and subsequently using that gene set to construct a translational signature from IPF samples. Day 7 was identified as having the highest similarity to IPF based on gene expression measurements [13]. The authors suggest that day 7 is the time point to administer antifibrotic compounds in order to best assess potential clinical outcomes.

Here, we have leveraged microarray profiling data from an extensive cohort of IPF and control samples within the Lung Tissue Research Consortium (LTRC) to develop an unbiased statistical model that defines a parsimonious 15-gene disease signature for IPF. The model has been trained and validated to accurately predict disease status across several IPF data sets. In addition, we identified a 29-gene secreted protein plasma signature for IPF and show that the

Table 1. Clinical samples.

Study	GEO Accession	Microarray platform	# IPF	# Normal
LTRC (Yang et al. (2013))	GSE32537	Affymetrix 1.0 ST	119	50
NJH (Yang et al. (2013))	NA	Affymetrix 1.0 ST	111	39
LGRC	GSE47460	Agilent-014850; Agilent-028004	160	108
Konishi et al. (2009)	GSE10667	Agilent-014850	23	15
Melzter et al. (2011)	GSE24206	Affymetrix U133	11	6
DePianto et al. (2015)	GSE53845	Agilent-014850	40	8

Models were trained and tested using these public cohorts of expression data for IPF and normal healthy patient lung samples. Sample counts are from the original studies.

https://doi.org/10.1371/journal.pone.0215565.t001

bleomycin model of lung fibrosis at 21 days shows the largest congruence to the disease signature. Our work defines a robust genetic signature for IPF providing a potential multi-analyte biomarker panel for validation, as well as enables the identification of preclinical models that most closely resemble human IPF.

Materials and methods

Clinical data

Expression data for IPF and normal healthy patient lung samples were derived from 6 distinct cohorts (Table 1). The bulk of these expression data was available via the NCBI Gene Expression Omnibus (GEO). The clinical expression data include the Lung Tissue Research Consortium (LTRC; GSE32537) cohort [9,17], the Lung Genomics Research Consortium (LGRC; GSE47460) [17,18], the National Jewish Health (NJH) cohort [9] (data via personal communication, Ivana Yang) and several smaller cohorts, GSE10667 [7], GSE24206 [19] and GSE53845 [20]. Transcript abundances were measured on both Affymetrix and Agilent microarray platforms. The LGRC and LTRC share samples, and these were excluded appropriately during model testing. We also excluded non-IPF or normal patient samples such as non-IPF interstitial lung diseases and Chronic Obstructive Pulmonary Disease (COPD) (these can be found in the LGRC). We note that due to insufficient annotation information across all studies, we did not correct for cellular composition or type in lung tissue samples.

In order to predict disease status of patients in the test cohorts we had to map Agilent expression measurements to the Affymetrix measurement space. This is similar to the scaling approach used by *Meltzer et al.* (2011) when mapping GSE24206 Affymetrix training features to the GSE10667 Agilent features. Conveniently, due to the common source of LTRC lung tissue used to generate both GSE32537 and GSE47460, there exist 85 common patient samples with both Affymetrix and Agilent data, allowing us to directly map expression signal across platforms. We generated gene-level scaling factors, which were possible because the ratios of Affymetrix/Agilent for each gene had very low variance. Genes included in the model were present on both Affymetrix and Agilent platforms.

Preclinical model data

Bleomycin preclinical rat model data was publicly available (GSE48455) [13]. *Bauer et al.* (2013) intratracheally administered Sprague Dawley rats with a single instillation of saline or bleomycin and sacrified the animals along a time course of 3, 7, 14, 21, 28, 42, and 56 days post-treatment (Table 2). Rat-Human orthologs were mapped using NCBI HomoloGene [21]. For simplicity in interpretation, only orthologs with a one-to-one mapping were included (excluding one-to-many mappings).

Identifying secreted proteins

Secreted genes were annotated using Gene Ontology (GO) cellular component annotations [22,23]. Genes were included if identified as existing in the extracellular space (GO:0005615)

Table 2. Bleomycin preclinical rat model samples.

	3	7	14	21	28	42	56
Bleomycin	5	5	5	5	5	5	5
Vehicle	5	5	4	5	5	5	5

Sample breakdown of bleomycin preclinical rat model (GSE48455) [13]. The time course experiment contains samples from 3, 7, 14, 21, 28, 42, and 56 days post-treatment.

https://doi.org/10.1371/journal.pone.0215565.t002

and not on the cell surface (GO:0009986). Our motivation was to exclude genes found on the cellular surface which were annotated as secreted.

Computational and statistical processing

R version 3.4.1 and Bioconductor were used for expression data retrieval from GEO, normalization, filtering and scaling [24,25]. When present, batch/microarray platform effects were removed using the sva package [26]. Differential expression contrasts were computed using the limma package [27]. Regularized regression using elastic nets was computed using the glmnet package [28]. Balanced and repeated cross-validation was executed using the caret package [29].

All code and data required to execute the analysis described in this manuscript have been deposited in GitHub (https://github.com/ronammar/ipf_signature_elastic_net).

Model construction and optimization

Disease status (IPF or normal) was used as a categorial response with two possible outcomes in logistic regression, which models the probability of response using a binomial link function to define a model of disease. However, logistic regression can be unreliable when $n \approx p$ or $p > n$. By linearly combining both l_1 and l_2 penalties of the lasso and ridge regression methods, respectively, *elastic net* regularization improves model performance and simultaneously selects features [28,30–32] (Appendix A1). Regularized regression techniques shrink coefficient estimates towards zero, and the use of the l_2 penalty in our model forces some coefficient estimates to be equal to exactly zero. Coefficient estimates that are non-zero are selected for inclusion in the model [32].

All models were trained on the LTRC lung tissue expression data. Only gene expression data were used for modeling, and clinical or demographic data were not included as these covariates are not always available and of uniform quality. Disease classification was accomplished using an elastic net regularized regression model [28]. Elastic net training requires the selection of both a lasso and ridge mixing parameter, α, and a penalty strength parameter, λ (Appendix A1). To identify the optimal combination with the highest performance, we conducted 10-fold balanced cross-validation for each α,λ pair in a grid search on the LTRC training data (S1 Fig).

The grid search appears to indicate no significant performance associated with α, which controls the number of features included in the model. This means we can increase α to make the model more lasso-like, while maintaining high performance by adjusting λ accordingly. We chose $\alpha = 0.95$ based on the suggestion in the glmnet documentation to set $\alpha = 1 - \epsilon$ for some small $\epsilon > 0$ [31]. The rationale is to improve numerical stability and reduce the degeneracies cause by high correlations between covariates.

Once we set α, we performed 1000 repeats of 10-fold cross-validation in caret to select the λ that yielded the highest performing model (lowest misclassification error) on the LTRC training data. This generated the final model and set of selected features (genes). For completeness, we computed the inclusion frequencies for each feature (S1 Supporting Information) [33,34]. We do not calculate significance of features in our model, as this is a relatively new and active area of statistics research [35]. Due to the challenges in computing appropriate estimates of the degrees of freedom, this significance test is currently in development for elastic nets.

This same approach was used to construct a model for each subset of genes including all, secreted genes, genes differentially-expressed in the bleomycin rat model and a combination of secreted and differentially-expressed genes. Four models were constructed in total, and these are available as serialized R objects in our code respository.

Results

Feature selection and model construction

We chose the LTRC and NJH cohorts for model training and initial testing, respectively, because they represented distinct patient populations, but were processed on the same expression platform (Affymetrix) by the same authors [9]. These two cohorts also each contained a relatively large number of samples, which is ideal for training and testing statistical models.

Before training, we compared patients to one another in an unbiased manner with t-Distributed Stochastic Neighbor Embedding (t-SNE), a nonlinear dimensionality reduction method capable of reducing the entire transcriptome signal into just two (or three) dimensions for visualization [36]. With transcription data for all IPF and normal patient samples in the LTRC and NJH cohorts, we observed distinct grouping of patient samples by disease status with no clear trend indicating a grouping by cohort (Fig 1, S2 and S5 Figs). A few outliers were identified with this method, but were not excluded from the subsequent work.

The LTRC lung tissue expression data was used to train all models (see Materials and Methods). Four models were constructed in total based on different gene subsets as input. These include models built with all genes ($\mathcal{M}$, 13896 initial features), secreted genes ($\mathcal{M}_{secreted}$, 910 initial features), genes differentially-expressed in the bleomycin rat model ($\mathcal{M}_{bleomycin}$, 1677 initial features) and the intersection of secreted and differentially-expressed genes ($\mathcal{M}_{secreted \cap bleomycin}$, 210 initial features) (S1–S3 Tables). 15 gene features were selected by $\mathcal{M}$ (Table 3).

When the expression of these 15 genes is hierarchically-clustered, we observe a very clear separation between IPF and normal patient samples (Fig 2 and S6–S9 Figs). While clustering is not used for disease status classification, the use of this orthogonal data-driven approach independently demonstrates that the 15 gene panel can be used to effectively discriminate between IPF and normal using transcript abundance alone.

Model validation on independent clinical cohorts

We first validated all models ($\mathcal{M}$, $\mathcal{M}_{secreted}$, $\mathcal{M}_{bleomycin}$, $\mathcal{M}_{secreted \cap bleomycin}$) on the NJH cohort. Due to the identical platform and processing, the NJH cohort provided a novel patient sample set while reducing variance from technical factors. All models were also tested on four other independent cohorts (Fig 3). As expected, the most unbiased model (no subsetting of genes before regularization), $\mathcal{M}$, performed the best, while reducing the number of genes for subsequent regularization generally reduced performance. Based on the area under the curve (AUC) metrics, $\mathcal{M}$ is the most performant published model of IPF disease status [13,19,37].

Identifying the most congruent rat bleomycin model

The IPF signatures derived from each of our four models could be used to identify the preclinical rat bleomycin model with the highest congruence to IPF. Before comparing ortholog expression across species, we attempted to normalize species-specific expression by comparing ratios from rat to human, computed as $log_2(bleomycin/saline)$ for rat samples and $log_2(IPF/control)$ for human samples. During initial comparisons between rat and IPF using the $\mathcal{M}$ feature set, we noticed that many rat genes were not differentially-expressed ($|log_2(bleomycin/saline)| \approx 0$), introducing noise when computing similarity between rat and IPF expression. Therefore, when comparing rat to IPF, we used the $\mathcal{M}_{bleomycin}$ feature set (30 gene features), which includes only genes that were differentially-expressed at any of the days in the bleomycin time course. Similarity was computed using Pearson correlation between each day post-

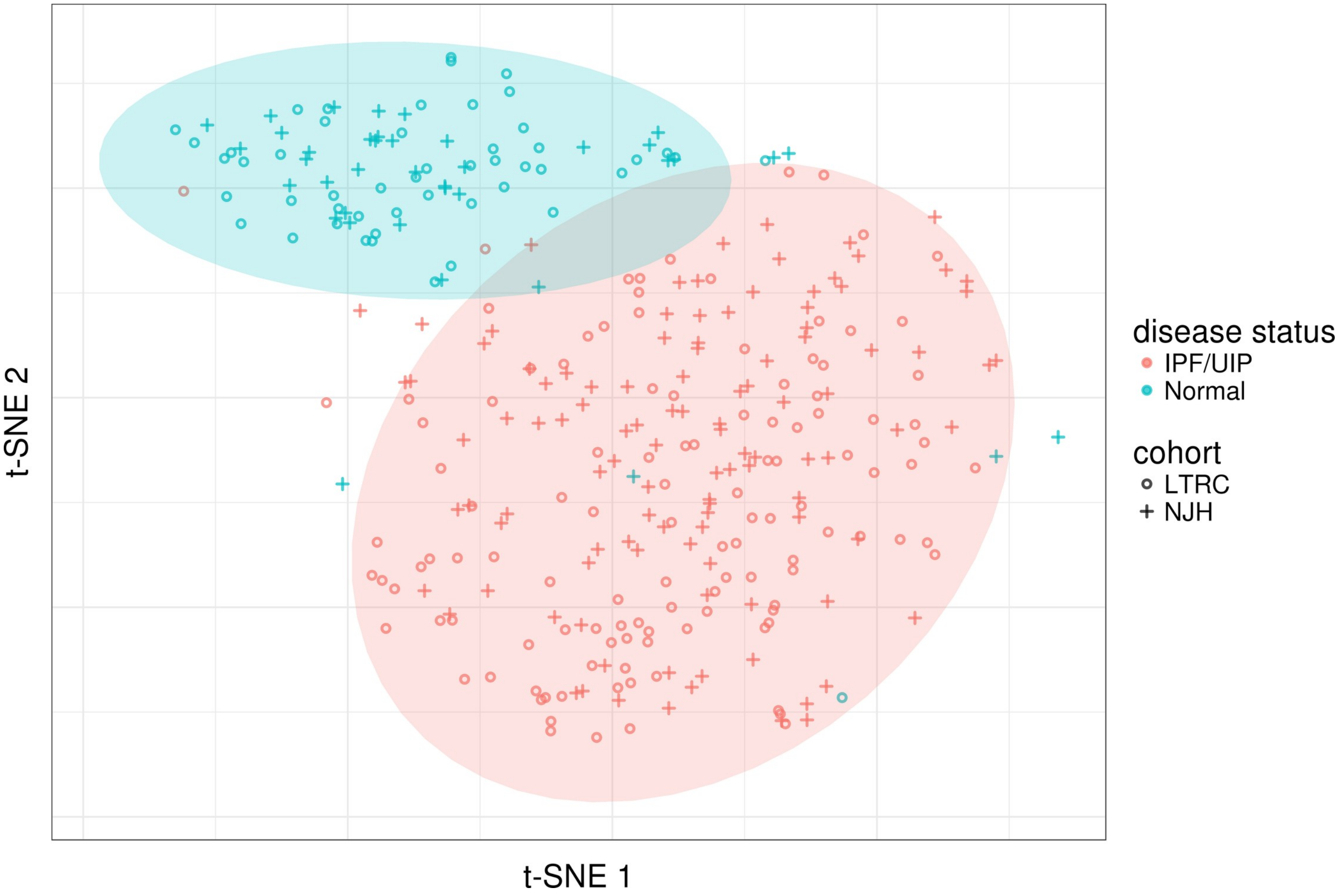

Fig 1. t-SNE models each high-dimensional observation into just two dimensions such that similar observations are modeled by nearby points and dissimilar objects are modeled by distant points. Applying t-SNE to our clinical samples from the LTRC and NJH, We observe distinct grouping of IPF and normal samples with a few outliers. There does not appear to be any grouping of patients by cohort.

https://doi.org/10.1371/journal.pone.0215565.g001

bleomycin treatment and IPF samples (using the LGRC samples, to compare our results more directly with those previously published [13]). We found model-IPF congruence increased from days 3 to 14 with maximum similarity between the model and IPF at day 21 (S3 Fig). It is important to note that multiple other murine models of pulmonary fibrosis exist [39], and we have only chosen the rat bleomycin model to compare to the disease, but other model comparisons may be the subject of future work.

Discussion

Given the challenges associated with the diagnosis of IPF and the inaccuracy of clinical prediction tools, it is imperative to explore new methods for diagnosis, classification and patient stratification. We have effectively leveraged microarray data from a large cohort of IPF patients within the LTRC to generate a new computational classifier of IPF disease. Although IPF disease signatures have been described before [9,13,19,20], the strength of our approach is the number of samples used, the unbiased computational model developed to define the signature and the extensive validation across multiple IPF cohorts. Our model outperforms several other previous models based on the near 100% prediction of disease status across multiple validation

Table 3. 15-gene signature for IPF.

Coefficient	Accession	Symbol	Description
-0.3644650	54829	ASPN	Asporin
1.0505567	875	CBS	Cystathionine-beta-synthase
0.3145191	1131	CHRM3	cholinergic receptor muscarinic 3
0.0221471	114805	GALNT13	Polypeptide N-acetylgalactosaminyltransferase 13
0.2791872	374378	GALNT18	polypeptide N-acetylgalactosaminyltransferase 18
0.0460736	2878	GPX3	Glutathione peroxidase 3
0.0214608	4047	LSS	lanosterol synthase (2,3-oxidosqualene-lanosterol cyclase)
0.0028236	56922	MCCC1	methylcrotonoyl-CoA carboxylase 1
0.0747452	8972	MGAM	Maltase-glucoamylase
-0.0382877	4316	MMP7	matrix metallopeptidase 7
0.0546881	5028	P2RY1	purinergic receptor P2Y1
-0.1286685	6423	SFRP2	secreted frizzled related protein 2
0.3537117	25777	SUN2	Sad1 and UNC84 domain containing 2
0.0437323	10579	TACC2	transforming acidic coiled-coil containing protein 2
-0.0245404	64393	ZMAT3	zinc finger matrin-type 3
-10.1203104	(Intercept)	NA	NA

Gene features selected by elastic nets defining the IPF gene signature when no genes have been filtered, using $\mathcal{M}$. The coefficients are extracted from $\mathcal{M}$. Accessions are Entrez gene identifiers.

https://doi.org/10.1371/journal.pone.0215565.t003

cohorts. *Bauer et al.* (2015) described a 12-gene signature identified from about 100 IPF samples compared with control lungs and established the commonality of this signature with that derived from the rat model of bleomycin induced fibrosis at the 7-day time point. Our study complements and extends these findings by developing alternate signatures and establishing congruence with the rat model of bleomycin induced fibrosis. Tissue and peripheral gene/protein expression signatures provide complex information that could be poorly or incompletely understood in the absence of effective computational modeling. Our study identifies a novel 15-gene signature that accurately predicts IPF disease status (Table 3). The signature contains several genes previously not associated with IPF as well as genes such as MMP7 which is a known biomarker for IPF [10,11] and sFRP2, a Wnt-signaling molecule described as a prospective therapeutic target [40]. Notably, MMP7 knockout mice do not develop fibrosis in response to bleomycin treatment [41]. Also, active MMP7 has been detected in IPF lungs but not healthy lungs and has been implicated as a profibrotic metalloprotease [42,43]. Glutathione Peroxidase-3 (GPX3) identified in our signature has been shown to be present in the epithelial lining fluid in the bleomycin-induced fibrosis model and upregulated in IPF [44].

Peripheral blood-derived biomarkers and expression signatures are more clinically translatable and developable as diagnostic tools as opposed to tissue-derived signatures, especially in diseases like IPF where tissues are hard to obtain and gene expression patterns are spatially restricted within the tissue. Profiling of plasma proteome in IPF has identified minimal protein signatures of IPF, as well as potential biomarkers [10,11,45,46] of disease progression including MMP1, MMP7, and surfactant protein-D. In a recent study [47], a 52-gene signature was developed from gene expression profiling of peripheral blood mononuclear cells from a cohort of IPF patients and validated for outcome prediction across two additional cohorts. Many of the identified genes were involved in defense response, wound healing and protein phosphorylation. In our study, we generated a 29-gene secreted protein signature from the tissue microarray data. This signature is enriched for genes in immune response and cell-matrix interaction pathways. Additionally, several extracellular matrix genes such as COMP,

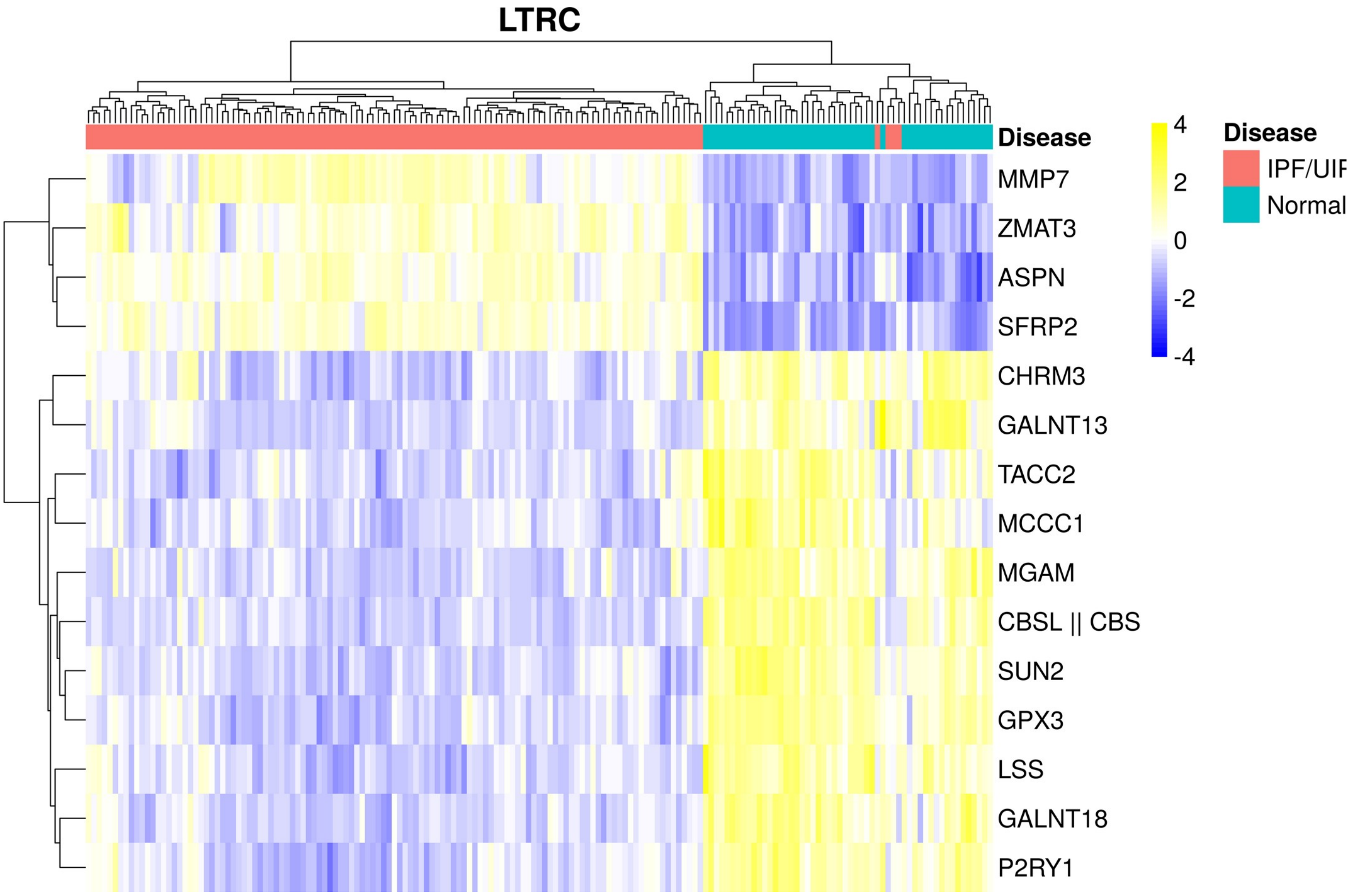

Fig 2. Hierarchical clustering of 15 gene signature used by model *M* to classify disease status. A clear separation is observed between IPF and normal patient samples. Note that this is just for visualization purposes, and $\mathcal{M}$ uses logistic regression to classify samples, not clustering. Row-scaled log intensity units are plotted. We use the complete linkage method for hierarchical clustering with a Euclidean distance measure.

https://doi.org/10.1371/journal.pone.0215565.g002

SPOCK1, Laminin C1 and ECM2 were identified as signature genes in our study. A secreted protein signature from tissue derived expression data could represent a robust and specific reflection of disease status. Future studies should validate the protein-level expression of these genes in serum/plasma.

In our study, we also show that the rat bleomycin model at day 21 has the highest congruence to the human IPF signature. This contrasts with the results of *Bauer et al.* (2015) wherein the rat model of fibrosis day 7 was determined to be the most similar to human disease. This is likely due to our similarity being assessed only using the IPF-derived gene signatures and not larger sets of genes (S3 and S4 Figs). We determined that using 30 genes to define similarity is more informative than using the entire set of genes that are differentially expressed in IPF and mapped to rat. After day 21, similarity is reduced, but remains relatively high, suggesting a persistent fibrotic state.

In future work, we propose to predict disease progression or severity of IPF with the inclusion of FVC or DLCO lung function measures. This would be analogous to the PROFILE study where *Maher et al.* (2017) showed that a 4 serum biomarker panel could be used to predict mortality and distinguish between stable and progressive IPF [48]. We also note that endpoint gene expression measurements represent a functional vignette of a biological system. Having access to gene expression changes over time along with protein abundances among

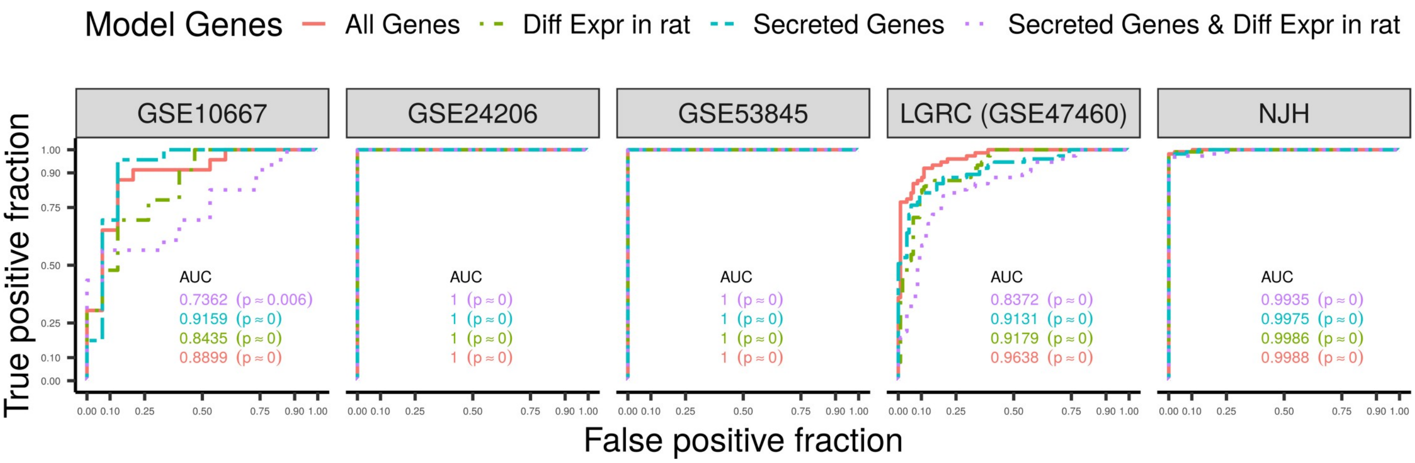

Fig 3. Model performance is assessed on the test data and evaluated with Receiver Operating Characteristic (ROC) curves. We compute the area under the curve (AUC) for each model and each cohort, where a perfect classifier has an $AUC = 1$ and a random classifier has an $AUC = 0.5$ (the diagonal line). The ROC curves and AUCs were calculated by passing the true positive fraction (the probability of a test positive among the diseased population) and false positive fraction (the probability of a test positive among the normal population) to the plotROC package (S5 Table) [38]. We performed 1000 bootstraps of the data for each cohort to establish a null distribution yielding a mean AUC of approximately 0.5, as expected. Based on the empirical p-values from these bootstrap AUCs, we find that all our reported AUCs are statistically significant, confirming the performance of our models across all test cohorts (S4 Table).

https://doi.org/10.1371/journal.pone.0215565.g003

other measures would shed more light on mechanisms behind IPF, and this is the subject of future work.

We have discriminated effectively between IPF and control lung tissues which is relevant in translational models of disease, but the complexity in diagnosing IPF manifests largely in distinguising it from other idiopathic interstitial pneumonias (IIPs) [49]. While the LGRC contains gene expression from lung samples of control and IPF patients, it also contains COPD and other IIP samples. However, given the paucity of similar data sets, it is challenging to validate a model trained to discriminate between COPD and respiratory bronchiolitis-associated interstitial lung disease or desquamative interstitial pneumonia. Promising modeling efforts are underway, but are limited by the number and diversity of available patient samples (eg. 115 samples across 14 pathology diagnoses, with most diagnoses matching very few patients) [50]. For future work, given the appropriate training and test data, we propose to construct models similar to our own with the ability to distinguish between IIPs.

Appendix

A1. Elastic nets: Logistic regression with a binomial distribution

From the glmnet elastic nets R package [28,31], we define the following:

The response variable takes a value in $\mathcal{G} = 1, 2$. Denote $y_i = I(g_i = 1)$.

We model $\Pr(G = 2|X = x) + \frac{e^{\beta_0+\beta^T x}}{1+e^{\beta_0+\beta^T x}}$,

With the log-odds transformation $\log\frac{\Pr(G=2|X=x)}{\Pr(G=1|X=x)} = \beta_0 + \beta^T x$.

The objective function for the penalized logistic regression uses the negative binomial log-likelihood $\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} -\left[\frac{1}{N}\sum_{i=1}^{N} y_i \cdot (\beta_0 + x_i^T\beta) - \log(1 + e^{(\beta_0+x_i^T\beta)})\right] + \lambda\left[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1\right]$, where the elastic net penalty is controlled by the mixing parameter α combining both lasso (l_1, $\alpha = 1$) and ridge (l_2, $\alpha = 0$) penalties. The tuning parameter λ corresponds to the strength of the penalty.

We note that due to the presence of multicollinearity in high-dimensional expression data, where $p \gg n$, regularized regression may be used to construct an accurate disease classifier based on transcript abundance, but each model represents one of many possible models [32]. We establish confidence in an individual model (and set of features) by validating/testing it on multiple independent cohorts.

Supporting information

S1 Fig. Elastic net grid search performance. We iterated over a grid of possible paired α and λ parameters for the elastic net module to determine optimal performance while reducing the number of features to create a minimal gene signature. Minimum classification error can be achieved at any value of α given an optimization for λ. The number of features included is annotated for each pair of parameters. The large red block represents a 0 gene feature model (only including an intercept β0).
(TIF)

S2 Fig. PCA dimensionality reduction. The proportions of variance accounted for by each of the first two principal components are indicated in parentheses. In this instance, t-SNE was more informative than Principal Components Analysis (PCA) because PCA yields $n-1$ principal components for an observation matrix of $n \times p$ where $p \geq n$ (n is the number of observations and p is the number of variables), where the variance is non-uniformly distributed across these eigenvectors. Instead the variance is typically spread across more than the first two or three eigenvectors yielding poorer separation between disease and control patients when only taking these eigenvectors into account.
(TIF)

S3 Fig. Congruence between bleomycin model and IPF. For the 30 gene expresion signature from $\mathcal{M}_{bleomycin}$, the similarity between the rat and IPF expression increased from days 3 to 14 post-bleomycin treatment with maximum similarity at day 21. After day 21, similarity is reduced, but remains relatively high, suggesting a possible fibrotic state. r = Pearson correlation coefficient where $-1 \leq r \leq 1$, with 1 meaning perfectly correlated and -1 perfectly anticorrelated.
(TIF)

S4 Fig. Congruence between bleomycin model and IPF using all differentially-expressed genes. If we examine only those genes that are differentially-expressed in IPF relative to controls ($\frac{IPF}{control} \leq 1.5$ and $FDR < 0.1$), and identify the orthologs in the rat, we do not observe increased similarity at any time point post-bleomycin treatment to suggest maximal congruence with IPF. This motivates the use of a smaller gene expression signature to extract only IPF-relevant gene expression.
(TIF)

S5 Fig. t-SNE dimensionality reduction for all test cohorts.
(TIF)

S6 Fig. Hierarchical clustering of 15 gene signature used by model $\mathcal{M}$ to classify disease status for all test cohorts. We use the complete linkage method for hierarchical clustering with a Euclidean distance measure.
(TIF)

S7 Fig. Hierarchical clustering of gene signature used by model $\mathcal{M}_{secreted}$ to classify disease status for all test cohorts. We use the complete linkage method for hierarchical clustering with a Euclidean distance measure.
(TIF)

S8 Fig. Hierarchical clustering of gene signature used by model $\mathcal{M}_{bleomycin}$ to classify disease status for all test cohorts. We use the complete linkage method for hierarchical clustering with a Euclidean distance measure.
(TIF)

S9 Fig. Hierarchical clustering of gene signature used by model $\mathcal{M}_{secreted \cap bleomycin}$ to classify disease status for all test cohorts. We use the complete linkage method for hierarchical clustering with a Euclidean distance measure.
(TIF)

S1 Table. Gene features selected by elastic nets defining the IPF gene signature when only secreted genes are included in $\mathcal{M}_{secreted}$. The coefficients are extracted from $\mathcal{M}_{secreted}$.
(CSV)

S2 Table. Gene features selected by elastic nets defining the IPF gene signature when only differentially-expressed genes from the bleomycin model are included in $\mathcal{M}_{bleomycin}$. The coefficients are extracted from $\mathcal{M}_{bleomycin}$.
(CSV)

S3 Table. Gene features selected by elastic nets defining the IPF gene signature when secreted and differentially-expressed genes from the bleomycin model are included in $\mathcal{M}_{secreted \cap bleomycin}$. The coefficients are extracted from $\mathcal{M}_{secreted \cap bleomycin}$.
(CSV)

S4 Table. Mean AUC from 1000 bootstraps of the test cohort data.
(CSV)

S5 Table. True positive fraction (the probability of a test positive among the diseased population) and false positive fraction (the probability of a test positive among the normal population) passed to the plotROC package to plot Fig 3.
(CSV)

S1 Supporting Information. For each model, we report inclusion frequencies of each gene feature using the method of Meinshausen & Bühlmann [33].
(XLSX)

Acknowledgments

A. Apfel, S. Chasalow and A. Greenfield provided critical statistical feedback. A. Greenfield provided valuable comments on early drafts of this manuscript. C.A. Tilford provided updated probe mapping for the Affymetrix Human Gene 1.0 ST Array and identified secreted proteins from GO annotations. We also thank I. Yang (University of Colorado Denver) for providing data for the NJH cohort and T. Hastie (Stanford University) for bug fixes to the glmnet package during development.

Author Contributions

Conceptualization: Ron Ammar, John Ryan Thompson.

Data curation: Ron Ammar.

Formal analysis: Ron Ammar.

Methodology: Ron Ammar, Pitchumani Sivakumar, Gabor Jarai, John Ryan Thompson.

Writing – original draft: Ron Ammar, Pitchumani Sivakumar, Gabor Jarai, John Ryan Thompson.

Writing – review & editing: Ron Ammar, Pitchumani Sivakumar, Gabor Jarai, John Ryan Thompson.

References

1. Mora AL, Rojas M, Pardo A, Selman M. Emerging therapies for idiopathic pulmonary fibrosis, a progressive age-related disease. Nat Rev Drug Discov. Division of Pulmonary, Allergy; Critical Care Medicine, Department of Medicine, E1246 BST, 200 Lothrop Street, University of Pittsburgh, Pittsburgh, Pennsylvania 15213, USA. 2017; 16: 755–772. https://doi.org/10.1038/nrd.2017.170 PMID: 28983101
2. George PM, Wells AU. Pirfenidone for the treatment of idiopathic pulmonary fibrosis. Expert Rev Clin Pharmacol. a Department of Respiratory Medicine, Royal Brompton Hospital, Interstitial Lung Disease Unit, London, SW3 6NP, UK. 2017; 10: 483–491. https://doi.org/10.1080/17512433.2017.1295846 PMID: 28266906
3. Rogliani P, Calzetta L, Cavalli F, Matera MG, Cazzola M. Pirfenidone, nintedanib and n-acetylcysteine for the treatment of idiopathic pulmonary fibrosis: A systematic review and meta-analysis. Pulm Pharmacol Ther. University of Rome Tor Vergata, Department of Systems Medicine, Unit of Respiratory Clinical Pharmacology, Rome, Italy; University of Rome Tor Vergata, Department of Systems Medicine, Chair of Respiratory Medicine, Rome, Italy. 2016; 40: 95–103. https://doi.org/10.1016/j.pupt.2016.07.009 PMID: 27481628
4. Tomioka H, Takada H. Treatment with nintedanib for acute exacerbation of idiopathic pulmonary fibrosis. Respirol Case Rep. Department of Respiratory Medicine Kobe City Medical Center West Hospital Kobe Japan. 2017; 5: e00215. https://doi.org/10.1002/rcr2.215 PMID: 28096998
5. Nathan SD, Albera C, Bradford WZ, Costabel U, Glaspole I, Glassberg MK, et al. Effect of pirfenidone on mortality: Pooled analyses and meta-analyses of clinical trials in idiopathic pulmonary fibrosis. Lancet Respir Med. Inova Fairfax Hospital, Falls Church, VA, USA. Electronic address: steven.nathan@inova.org. 2017; 5: 33–41. https://doi.org/10.1016/S2213-2600(16)30326-5 PMID: 27876247
6. Martinez FJ, Collard HR, Pardo A, Raghu G, Richeldi L, Selman M, et al. Idiopathic pulmonary fibrosis. Nat Rev Dis Primers. Joan; Sanford I. Weill Department of Medicine, Weill Cornell Medical College, New York-Presbyterian Hospital/Weill Cornell Medical Center, 1305 York Avenue, Box 96, Room Y-1059, New York, New York 10021, USA. 2017; 3: 17074. https://doi.org/10.1038/nrdp.2017.74 PMID: 29052582
7. Konishi K, Gibson KF, Lindell KO, Richards TJ, Zhang Y, Dhir R, et al. Gene expression profiles of acute exacerbations of idiopathic pulmonary fibrosis. Am J Respir Crit Care Med. Division of Pulmonary, Allergy; Critical Care Medicine, Dorothy P.; Richard P. Simmons Center for Interstitial Lung Diseases, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. 2009; 180: 167–175. https://doi.org/10.1164/rccm.200810-1596OC PMID: 19363140
8. Kusko RL, Brothers JF 2nd, Tedrow J, Pandit K, Huleihel L, Perdomo C, et al. Integrated genomics reveals convergent transcriptomic networks underlying chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis. Am J Respir Crit Care Med. 1 Computational Biomedicine, Boston University School of Medicine, Boston, Massachusetts. 2016; 194: 948–960. https://doi.org/10.1164/rccm.201510-2026OC PMID: 27104832
9. Yang IV, Coldren CD, Leach SM, Seibold MA, Murphy E, Lin J, et al. Expression of cilium-associated genes defines novel molecular subtypes of idiopathic pulmonary fibrosis. Thorax. Department of Medicine, University of Colorado School of Medicine, Aurora, Colorado, USA. 2013; 68: 1114–1121. https://doi.org/10.1136/thoraxjnl-2012-202943 PMID: 23783374
10. Bauer Y, White ES, de Bernard S, Cornelisse P, Leconte I, Morganti A, et al. MMP-7 is a predictive biomarker of disease progression in patients with idiopathic pulmonary fibrosis. ERJ Open Res. Actelion Pharmaceuticals Ltd, Allschwil, Switzerland. 2017; 3. https://doi.org/10.1183/23120541.00074–2016
11. Rosas IO, Richards TJ, Konishi K, Zhang Y, Gibson K, Lokshin AE, et al. MMP1 and mmp7 as potential peripheral blood biomarkers in idiopathic pulmonary fibrosis. PLoS Med. Dorothy P.; Richard P. Simmons Center for Interstitial Lung Diseases, Division of Pulmonary, Allergy; Critical Care Medicine, University of Pittsburgh School ofMedicine, Pittsburgh, Pennsylvania, United States of America. 2008; 5: e93. https://doi.org/10.1371/journal.pmed.0050093 PMID: 18447576
12. Yang IV, Luna LG, Cotter J, Talbert J, Leach SM, Kidd R, et al. The peripheral blood transcriptome identifies the presence and extent of disease in idiopathic pulmonary fibrosis. PLoS One. Center for Genes, Environment; Health, National Jewish Health, Denver, Colorado, United States of America. 2012; 7: e37708. https://doi.org/10.1371/journal.pone.0037708 PMID: 22761659

13. Bauer Y, Tedrow J, de Bernard S, Birker-Robaczewska M, Gibson KF, Guardela BJ, et al. A novel genomic signature with translational significance for human idiopathic pulmonary fibrosis. Am J Respir Cell Mol Biol. 1 Actelion Pharmaceuticals Ltd., Allschwil, Switzerland. 2015; 52: 217–231. https://doi.org/10.1165/rcmb.2013-0310OC PMID: 25029475

14. Cabrera S, Selman M, Lonzano-Bolanos A, Konishi K, Richards TJ, Kaminski N, et al. Gene expression profiles reveal molecular mechanisms involved in the progression and resolution of bleomycin-induced lung fibrosis. Am J Physiol Lung Cell Mol Physiol. Facultad de Ciencias, Universidad Nacional Autonoma de Mexico, Mexico DF, Mexico. 2013; 304: L593–601. https://doi.org/10.1152/ajplung.00320.2012 PMID: 23457188

15. Steele MP, Luna LG, Coldren CD, Murphy E, Hennessy CE, Heinz D, et al. Relationship between gene expression and lung function in idiopathic interstitial pneumonias. BMC Genomics. Department of Medicine, Vanderbilt University, Nashville, TN, USA. 2015; 16: 869. https://doi.org/10.1186/s12864-015-2102-3 PMID: 26503507

16. Chaudhary NI, Schnapp A, Park JE. Pharmacologic differentiation of inflammation and fibrosis in the rat bleomycin model. Am J Respir Crit Care Med. Department of Pulmonary Research, Boehringer Ingelheim Pharma GmbH & Co., KG, Biberach an der Riss, Germany. 2006; 173: 769–776. https://doi.org/10.1164/rccm.200505-717OC PMID: 16415276

17. Holmes DR III, Bartholmai BJ, Karwoski RA, Zavaletta V, Robb RA. The lung tissue research consortium: An extensive open database containing histological, clinical, and radiological data to study chronic lung disease. The Insight Journal—2006 MICCAI Open Science Workshop. 2006;

18. Kim S, Herazo-Maya JD, Kang DD, Juan-Guardela BM, Tedrow J, Martinez FJ, et al. Integrative phenotyping framework (iPF): Integrative clustering of multiple omics data identifies novel lung disease subphenotypes. BMC Genomics. Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA, 15261, USA. swiss747@gmail.com. 2015; 16: 924. https://doi.org/10.1186/s12864-015-2170-4 PMID: 26560100

19. Meltzer EB, Barry WT, D'Amico TA, Davis RD, Lin SS, Onaitis MW, et al. Bayesian probit regression model for the diagnosis of pulmonary fibrosis: Proof-of-principle. BMC Med Genomics. Department of Medicine, Division of Pulmonary, Allergy; Critical Care Medicine, Duke University Medical Center, Durham, North Carolina, USA. 2011; 4: 70. https://doi.org/10.1186/1755-8794-4-70 PMID: 21974901

20. DePianto DJ, Chandriani S, Abbas AR, Jia G, N'Diaye EN, Caplazi P, et al. Heterogeneous gene expression signatures correspond to distinct lung pathologies and biomarkers of disease severity in idiopathic pulmonary fibrosis. Thorax. Genentech Research; Early Development, South San Francisco, California, USA. 2015; 70: 48–56. https://doi.org/10.1136/thoraxjnl-2013-204596 PMID: 25217476

21. NCBI Resource C. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2013; 41: D8–D20. https://doi.org/10.1093/nar/gks1189 PMID: 23193264

22. The Gene Ontology C. Expansion of the gene ontology knowledgebase and resources. Nucleic Acids Res. 2017; 45: D331–D338. https://doi.org/10.1093/nar/gkw1108 PMID: 27899567

23. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: Tool for the unification of biology. the gene ontology consortium. Nat Genet. Department of Genetics, Stanford University School of Medicine, California, USA. cherry@stanford.edu; 2000; 25: 25–29. https://doi.org/10.1038/75556 PMID: 10802651

24. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2018. Available: https://www.R-project.org/

25. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods. 2015; 12: 115–121. Available: http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html https://doi.org/10.1038/nmeth.3252 PMID: 25633503

26. Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Storey JD, et al. Sva: Surrogate variable analysis. 2019.

27. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015; 43: e47. https://doi.org/10.1093/nar/gkv007 PMID: 25605792

28. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software. 2010; 33: 1–22. Available: http://www.jstatsoft.org/v33/i01/ PMID: 20808728

29. Jed Wing MKC from, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, et al. Caret: Classification and regression training [Internet]. 2018. Available: https://CRAN.R-project.org/package=caret

30. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). Wiley/Blackwell (10.1111); 2005; 67: 301–320.

31. Hastie T, Qian J. Glmnet vignette [Internet]. 2016 [cited 16 Nov 2017]. Available: https://web.stanford.edu/~hastie/glmnet/glmnet_beta.html

32. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Springer; 2013.

33. Meinshausen N, Bühlmann P. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology). Wiley Online Library; 2010; 72: 417–473.

34. Shah RD, Samworth RJ. Variable selection with error control: Another look at stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology). Wiley Online Library; 2013; 75: 55–80.

35. Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. A significance test for the lasso. Ann Stat. Department of Statistics; Actuarial Science, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada. 2014; 42: 413–468. https://doi.org/10.1214/13-AOS1175 PMID: 25574062

36. Maaten L van der, Hinton G. Visualizing data using t-sne. Journal of machine learning research. 2008; 9: 2579–2605.

37. Bauer Y, Nayler O, Kaminski N. Reply: The bleomycin model: In pursuit of relevant biomakers. Am J Respir Cell Mol Biol. 1 Actelion Pharmaceuticals Ltd. Allschwil, Switzerland. 2015; 53: 748–749. https://doi.org/10.1165/rcmb.2015-0196LE PMID: 26517754

38. Sachs MC. plotROC: A tool for plotting roc curves. Journal of Statistical Software, Code Snippets. 2017; 79: 1–19. https://doi.org/10.18637/jss.v079.c02 PMID: 30686944

39. Degryse AL, Lawson WE. Progress toward improving animal models for idiopathic pulmonary fibrosis. Am J Med Sci. Division of Allergy, Pulmonary; Critical Care Medicine, Vanderbilt University School of Medicine, Nashville, TN 37232–2650, USA. amber.degryse@vanderbilt.edu; 2011; 341: 444–449. https://doi.org/10.1097/MAJ.0b013e31821aa000 PMID: 21613932

40. Mastri M, Shah Z, Hsieh K, Wang X, Wooldridge B, Martin S, et al. Secreted frizzled-related protein 2 as a target in antifibrotic therapeutic intervention. Am J Physiol Cell Physiol. Department of Biochemistry; Department of Biomedical Engineering, Center for Research in Cardiovascular Medicine, University at Buffalo, Buffalo, New York. 2014; 306: C531–9. https://doi.org/10.1152/ajpcell.00238.2013 PMID: 24336656

41. Zuo F, Kaminski N, Eugui E, Allard J, Yakhini Z, Ben-Dor A, et al. Gene expression analysis reveals matrilysin as a key regulator of pulmonary fibrosis in mice and humans. Proc Natl Acad Sci U S A. Roche Bioscience, Palo Alto, CA 94304; Functional Genomics,; Institute of Respiratory Medicine, Sheba Medical Center, Tel Hashomer, 52621 Israel. 2002; 99: 6292–6297. https://doi.org/10.1073/pnas.092134099 PMID: 11983918

42. Fujishima S, Shiomi T, Yamashita S, Yogo Y, Nakano Y, Inoue T, et al. Production and activation of matrix metalloproteinase 7 (matrilysin 1) in the lungs of patients with idiopathic pulmonary fibrosis. Arch Pathol Lab Med. Department of Emergency; Critical Care Medicine, School of Medicine, Keio University, Shinjuku-ku, Tokyo, Japan. 2010; 134: 1136–1142. https://doi.org/10.1043/2009-0144-OA.1 PMID: 20670133

43. Pardo A, Cabrera S, Maldonado M, Selman M. Role of matrix metalloproteinases in the pathogenesis of idiopathic pulmonary fibrosis. Respir Res. Facultad de Ciencias, Universidad Nacional Autonoma de Mexico, Mexico, DF, Mexico. apardos@unam.mx. 2016; 17: 23. https://doi.org/10.1186/s12931-016-0343-6 PMID: 26944412

44. Schamberger AC, Schiller HB, Fernandez IE, Sterclova M, Heinzelmann K, Hennen E, et al. Glutathione peroxidase 3 localizes to the epithelial lining fluid and the extracellular matrix in interstitial lung disease. Sci Rep. Comprehensive Pneumology Center, Helmholtz Zentrum Munchen, Member of the German Center of Lung Research (DZL), Munich, Germany. 2016; 6: 29952. https://doi.org/10.1038/srep29952 PMID: 27435875

45. O'Dwyer DN, Norman KC, Xia M, Huang Y, Gurczynski SJ, Ashley SL, et al. The peripheral blood proteome signature of idiopathic pulmonary fibrosis is distinct from normal and is associated with novel immunological processes. Sci Rep. Division of Pulmonary; Critical Care Medicine, Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA. 2017; 7: 46560. https://doi.org/10.1038/srep46560 PMID: 28440314

46. Richards TJ, Kaminski N, Baribaud F, Flavin S, Brodmerkel C, Horowitz D, et al. Peripheral blood proteins predict mortality in idiopathic pulmonary fibrosis. Am J Respir Crit Care Med. The Dorothy P. & Richard P. Simmons Center for Interstitial Lung Disease, Department of Medicine, University of Pittsburgh Medical Center, NW 628 MUH, 3459 5th Avenue, Pittsburgh, PA 15261, USA. 2012; 185: 67–76. https://doi.org/10.1164/rccm.201101-0058OC PMID: 22016448

47. Herazo-Maya JD, Sun J, Molyneaux PL, Li Q, Villalba JA, Tzouvelekis A, 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. Section of Pulmonary, Critical Care,; Sleep Medicine, Department of Medicine, Yale School of Medicine, Yale University, New Haven, CT, USA; Section of

Pulmonary, Critical Care; Sleep Medicine, Department of Medicine, NCH Healthcare System; Mayo Clinic School of Medicine, Naples, FL, USA. Electronic address: jose.herazo-maya@yale.edu. 2017; 5: 857–868. https://doi.org/10.1016/S2213-2600(17)30349-1 PMID: 28942086

48. Maher TM, Oballa E, Simpson JK, Porte J, Habgood A, Fahy WA, et al. An epithelial biomarker signature for idiopathic pulmonary fibrosis: An analysis from the multicentre profile cohort study. Lancet Respir Med. NIHR Respiratory Biomedical Research Unit, Royal Brompton Hospital, London, UK; Fibrosis Research Group, National Heart; Lung Institute, Imperial College, London, UK. 2017; 5: 946–955. https://doi.org/10.1016/S2213-2600(17)30430-7 PMID: 29150411

49. Zhang Y, Kaminski N. Biomarkers in idiopathic pulmonary fibrosis. Curr Opin Pulm Med. Division of Pulmonary, Allergy; Critical Care Medicine, Dorothy P.; Richard P. Simmons Center for Interstitial Lung Disease, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, USA. 2012; 18: 441–446. https://doi.org/10.1097/MCP.0b013e328356d03c PMID: 22847105

50. Kim SY, Diggans J, Pankratz D, Huang J, Pagan M, Sindy N, et al. Classification of usual interstitial pneumonia in patients with interstitial lung disease: Assessment of a machine learning approach using high-dimensional transcriptional data. Lancet Respir Med. Veracyte, South San Francisco, CA, USA. 2015; 3: 473–482. https://doi.org/10.1016/S2213-2600(15)00140-X PMID: 26003389