Molecular Signatures of Idiopathic Pulmonary Fibrosis

Iain R. Konigsberg[1]*, Raphael Borie[2]*, Avram D. Walts[1], Jonathan Cardwell[1], Mauricio Rojas[3], Fabian Metzger[4], Stefanie M. Hauck[4], Tasha E. Fingerlin[5], Ivana V. Yang[1#], David A. Schwartz[#1]

1 Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora CO

2 Department of Medicine, Hôpital Bichat, Paris, France

3 Department of Medicine, University of Pittsburgh Medical Center, Pittsburgh PA

4 Research Unit Protein Science, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany

5 Department of Biomedical Informatics and Center for Genes, Environment and Health, National Jewish Health

* and # denote equal contributions of first and last authors, respectively

Correspondence should be addressed to:

Ivana V. Yang, PhD

12700 East 19th Avenue, 8611

Aurora, CO 80045

(303) 724-6449

Ivana.yang@cuanschutz.edu

AUTHOR CONTRIBUTIONS

TEF, IVY, and DAS conceived and designed the study. RB, ADW, FM and SMH collected the data. IRK, RB, JC, and FM analyzed the data. MR performed clinical phenotyping of the subjects. IRK, RB, IVY, and DAS wrote the manuscript. All authors edited and approved the manuscript.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

This article has an online data supplement, which is accessible from this issue's table of content online at www.atsjournals.org.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

ABSTRACT

Molecular patterns and pathways in idiopathic pulmonary fibrosis (IPF) have been extensively investigated but few studies have assimilated multi-omic platforms to provide an integrative understanding of molecular patterns that are relevant in IPF. Herein, we combine coding and non-coding transcriptome, DNA methylome, and proteome from IPF and healthy lung tissue to identify molecules and pathways associated with this disease. RNA sequencing, Illumina MethylationEPIC array, and liquid chromatography-mass spectrometry (LC-MS) proteomic data were collected on lung tissue from 24 IPF cases and 14 control subjects. Significant differential features were identified using linear models adjusting for age and sex, inflation and bias where appropriate. Data Integration Analysis for Biomarker discovery using a Latent component method for Omics studies (DIABLO) was used for integrative multi-omic analysis. We identified 4,643 differentially expressed transcripts aligning to 3,439 genes, 998 differentially abundant proteins, 2,500 differentially methylated regions (DMRs), and 1,269 differentially expressed lncRNAs that were significant after correcting for multiple tests (false discovery rate [FDR]<0.05). Unsupervised hierarchical clustering using 20 coding mRNA, protein, methylation, and lncRNA features with highest loadings on the top latent variable from the four datasets demonstrates perfect separation of IPF and control lungs. Our analysis confirmed previously validated molecules and pathways known to be dysregulated in disease, and implicated novel molecular features as potential drivers and modifiers of disease. For example, four proteins, 18 DMRs, and 10 lncRNAs were found to have strong correlations (|r|>0.8) with MMP7. Therefore, using a systems biology approach, we have identified novel molecular relationships in IPF.

KEYWORDS: systems biology, transcriptome, methylome, proteome, multi-omics

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

INTRODUCTION

Idiopathic pulmonary fibrosis (IPF) is a progressive and fatal disease of the aging lung (1, 2). Its prevalence is increasing (3) and it is likely underdiagnosed (4, 5). While cigarette smoke remains the most significant environmental risk factor for this complex disease (6), the gain-of-function *MUC5B* promoter variant is the strongest risk factor, genetic or otherwise, for the development of IPF. However, 13 other common variants and several rare variants including telomerase pathway genes also contribute to the risk of developing IPF (7-9). Although pirfenidone (10) and nintedanib (11) slow IPF progression, no treatment short of lung transplantation impacts survival. IPF is characterized by dysplastic bronchiolar metaplasia, alveolar epithelial injury and repair, proliferation of resident fibroblasts, formation of myofibroblastic foci, accumulation of extracellular matrix (ECM), and lung remodeling (12).

Genomic approaches have been used to characterize the molecular landscape of IPF. Gene expression studies have identified several thousand genes that are differentially regulated in the IPF lung (13-20), consistently reporting common genes and pathways (ECM organization & regulation, TGF-β signaling, endoplasmic reticulum stress, epithelial-mesenchymal transition (EMT), mitochondrial homeostasis, bronchial epithelial genes, fibroblast genes, smooth muscle markers, cytokines & chemokines, growth factors & receptors) that are differentially expressed in fibrotic lungs. A recent deep proteome profiling study has confirmed that many of these genomic differences result in differential protein abundance in the IPF lung with key genes such as MMP7 and MUC5B showing increased abundance (21). At the regulatory level, DNA methylation changes have been associated with many of the key transcriptional changes in IPF lung tissue (22-24) and hypermethylation of genes such as *CXCL10* (25), *PTGER2* (26), and *THY1* (27) have been shown to contribute to IPF pathogenesis. Genomic miRNA profiles have revealed several miRNAs that are known to affect fibroproliferation, epithelial-mesenchymal transition (EMT), and the TGF-β1 signaling pathway (28-32). Although studies of long noncoding RNAs in pulmonary

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

fibrosis are limited, there appears to be an anti-fibrotic role for FENDRR (33) and a pro-fibrotic role for DNM3OS (34). In addition, studies in peripheral blood have identified biomarkers of disease (35) and disease outcome (36, 37).

Despite the successful application of multiple single platform 'omic' technologies to characterize the molecular landscape of IPF, integrative approaches using system biology have not yet been applied to the IPF lung. Stimulated by a recent application of multi-omics to a small sample, big data study in newborns (38), we obtained DNA methylome, coding and non-coding transcriptome, and proteome results from 24 IPF and 14 control lungs. Leveraging supervised (39) and unsupervised (40) machine learning methods allowed us to identify integrated molecules and pathways across the multiple 'omic' platforms to more comprehensively characterize the complex molecular features of IPF.

METHODS

Ethics Statement

Human tissue was collected after appropriate ethical review for the protection of human subjects through the NHLBI-sponsored Lung Tissue Research Consortium (LTRC) and lung donor program at the University of Pittsburgh. De-identified data and samples were approved for use in this study at the University of Colorado (COMIRB # 15-1147).

Study Population

We selected 24 IPF subjects from the LTRC and 14 controls from the University of Pittsburgh Lung Core, all non-Hispanic whites. Details of the study population are provided in the Supplemental Methods.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

Sample Processing

DNA and RNA were isolated from the same sample of lung tissue using the AllPrep kit (Qiagen). Samples with RNA integrity number (RIN) and DNA integrity number (DIN) >5 were used. Genotyping for the *MUC5B* rs35705950 was performed utilizing a TaqMan assay (ThermoFisher). Sample preparation for proteomic analysis is described in Supplemental Methods.

Omic Data Collection

mRNA libraries were prepared from 500 ng total RNA with TruSeq stranded mRNA library preparation kits (illumina) and sequenced at the average depth of 80M reads on the Illumina NovaSeq 6000 (illumina). 4,011 unique proteins were detected using published MS spectrometry methods (41) and described in the Supplemental Methods. 1 ug of DNA was bisulfite treated using the Zymo EZ DNA Methylation kit, labeled and hybridized to Illumina Infinium Human MethylationEPIC BeadChip using standard protocols. rRNA-depleted libraries were prepared from 1 ug RNA with the Epidemiology Ribo-Zero Gold rRNA Removal Kit (Illumina) and sequenced at the average depth of 80M reads on the Illumina NovaSeq 6000 (Illumina). RNA-seq count level data and EPIC methylation data have been deposited to Gene Expression Omnibus (GEO) under accession GSE173357.

General Statistical Methods

All analyses were performed in R (3.6.2). Principal component analysis was used for quality control and no samples had to be excluded on this criterion. Principal component regression analysis (PCRA) was used to identify variables associated with top principal components, and strong batch effects were regressed out using ComBat (42). Differentially abundant features in each dataset were identified using linear models, adjusting for age and sex. In the mRNA-seq, lncRNA-seq, and DNA methylation datasets, p values were adjusted for

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

inflation and bias using Bacon (43). To control for multiple comparisons, p values were adjusted to a 5% false discovery rate (FDR) using the Benjamini-Hochberg procedure (44) in all datasets. Detailed methods for data processing and statistical analysis of individual datasets are provided in Supplemental Methods.

DIABLO

Data Integration Analysis for Biomarker discovery using Latent variable approaches for Omics studies (DIABLO) (39) was used to determine correlated 'omic' features associated with diagnosis. DIABLO is a supervised learning approach that builds on Regularized and Sparse Generalized Canonical Correlation Analysis (RGCCA), maximizing correlations between multiple datasets containing the same individuals and a classifier (diagnosis). DIABLO seeks for common information across different data types through the selection of a subset of molecular features, while discriminating between IPF and control lung tissue. Using simulations in DIABLO, we determined a single latent variable sufficiently captures most of the variation associated with diagnosis. For input into DIABLO, we used the four lists of differential features at FDR<0.05, with the mRNA set limited to 1,109 transcripts with |fold change|>4 to have a similar number of features as the remaining three datasets.

RESULTS

We selected IPF cases and controls to have similar demographic (age and sex) characteristics, all non-Hispanic white, and similar smoking histories (all ever or former smokers) (Table 1). All controls have GG *MUC5B* genotype and of cases 50% have the GG and 50% have the GT *MUC5B* genotype.

Coding RNA

116,503 transcripts were detected through polyA-enriched RNA-sequencing, 75,382 of which are annotated as protein coding or retained intron (alternatively spliced transcripts). 4,643

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

transcripts (protein coding and retained intron) aligning to 3,439 genes are differentially expressed in IPF compared to normal control lung tissue at FDR < 0.05 with stringent adjustment for inflation and bias (**Supplemental Figure S1A**). As an alternative to adjustment for inflation and bias, we performed cell deconvolution analysis with xCell (45) and adjusted for cell proportions in the statistical model but this method performed less well (**Supplemental Figure S2**). Of the 4,634 differentially expressed transcripts, 1,425 transcripts are upregulated in IPF lung tissue, while 3,218 RNAs are more abundant in control lung tissue (**Supplemental Table S1A**). The majority of differentially expressed transcripts are protein-coding (74%; **Figure 1A**) and the remaining 26% (**Figure 1B**) are alternatively spliced transcripts. Upregulated mRNAs are strongly enriched for protein products localizing to the mitochondria as well as species involved in oxidation. Downregulated mRNAs are enriched for focal adhesion and immune signaling pathways. Differentially expressed genes previously reported to be associated with IPF include matrix metalloproteinase 7 (MMP7), a gene that is the most established biomarker for IPF (17, 46, 47), and epidermal growth factor (EGF). Our recent analysis of transcriptional profiles of airway epithelial cells grown at air-liquid interface (ALI) at different time points identified an interaction of the epidermal growth factor receptor (EGFR) and the inducible transcriptional co-activator (YAP) as critical to the migratory phenotype of IPF cells (48). Additionally, we observe genes involved in other fibrotic lung disease, such as CUX1, a transcription factor which regulates COL1 expression and is upregulated in systemic sclerosis (49). CUX1 isoforms are localized within α-smooth muscle actin-positive cells in systemic sclerosis skin and IPF lung tissue sections, suggesting an important role for CUX1 in regulation of COL1 expression in fibrosis in multiple organs (50).

Protein

The LC-MS platform we used detected 22,198 peptides associated with 4,011 unique proteins/genes. 1,040 proteins were differentially abundant in IPF compared to control tissue at

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

FDR <0.05 (**Figure 1C, Supplemental Table S1B**). 522 proteins (including 27 core matrisome and 19 matrisome-associated proteins) are increased in IPF and 518 (including 24 core matrisome and 22 matrisome-associated proteins) are decreased. Differentially abundant proteins are significantly enriched for core matrisome and matrisome-associated proteins (Fisher's p = .001). We also detect multiple upregulated thioredoxin-related genes (P4HB, QSOX1, TXN2, TXNDC5, TXNL1) in IPF tissue. Thioredoxins are upregulated by ROS and reduce oxidative stress. TMEM231 shows the greatest increase in IPF. TMEM231 is a transmembrane protein present in the transition zone of cilia that prevents protein mis-localization by blocking protein diffusion across the ciliary membrane and is necessary for proper ciliogenesis. Our group has previously implicated ciliary dysfunction in IPF pathogenesis through patient clustering of gene expression microarray data (19).

Methylation

After stringent control for bias and inflation (**Supplemental Figure S1C**), we identified 2,500 DMRs overlapping or within 10 kb of 1,840 genes (**Figure 1D, Supplemental Table S1C**). As an alternative to adjustment for inflation and bias, we performed cell deconvolution analysis with RefFreeEWAS (51) and adjusted for cell confounding but this method performed less well (**Supplemental Figure S3**). Of the 2,568 DMR-gene relationships for the DMRs, 31% of DMRs are intronic to genes, 24% overlap an exonic region, and 11% are in promoters (defined as within 2kb upstream of the TSS). On average, significant DMRs contained four Illumina probes and span 335 bp. The absolute average difference in percent methylation of CpGs within significant DMRs in IPF versus control lung tissue is 9.6%. The greatest hypomethylated DMR shows a 30% decrease in methylation relative to control tissue. This 709 bp region contains six probes and overlaps the 3' UTR of VMP1 as well as most of the transcribed region of MiR-21, a microRNA shown to promote fibrogenesis through upregulation of TGF-β signaling, for which differential methylation has not been previously reported (29). Additional DMRs overlap genes shown to be

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

involved in lung development and fibrosis. We observe hypomethylation in FOXP1, a transcription factor involved in secretory epithelial cell fate determination in the lung (52), as well as differential methylation of genes previously shown to be involved in IPF such as CCL2 (53).

lncRNA

We identified 1,269 differentially expressed transcripts associated with 1,067 unique genes (FDR<0.05) (**Figure 1E, Supplemental Table S1D**), after controlling for bias and inflation (**Supplemental Figure S1D**). The majority of differentially expressed non-coding RNAs are lncRNAs (39%) and antisense RNAs (43%). As expected, most of the lncRNAs are those of unknown function. Among most differentially expressed lncRNAs with known function and upregulated in IPF are MUC5B-AS1, a noncoding RNA antisense to MUC5B, and LINP2, with multiple roles in cancer (54). lncRNAs of known function downregulated in IPF include long intergenic nonprotein coding RNA p53-induced transcript (LINC-PINT), which reduces lung cancer progression via sponging of miR-543 to induce tumor suppressor phosphatase and tensin homolog (PTEN)(55).

Protein-coding Transcriptome and Proteome Interactions

To begin to integrate datasets, we first performed pairwise comparisons of coding mRNA and protein data, initially focusing on transcripts and proteins with significant changes in both datasets. Comparing fold change of protein to mRNA, we demonstrated that most changes with large effect sizes (fourfold change in IPF vs control) are consistent directionally (**Figure 2A**). Protein activator of interferon induced protein kinase EIF2AK2 (PRKRA) is especially highly upregulated at the mRNA and protein level. This protein kinase is activated by double-stranded RNA and mediates the effects of interferon in response to viral infections, which are known risk factors of disease (56, 57). PRKRA promoter hypomethylation has been previously reported in

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

IPF lung tissue (23) and our novel observation of mRNA and protein upregulation further suggests a role for this gene in disease.

Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis revealed common pathways, specifically focal adhesion and adherens junctions, in mRNA and protein datasets (**Figure 2B**). Focal adhesion, adhesive contact between the cell and extracellular matrix through the interaction of integrin transmembrane proteins with their extracellular ligand, is strongly enriched in both mRNA and protein datasets. While the majority of transcripts in the focal adhesion pathway are downregulated in IPF, protein-level data demonstrate a mix of up and down-regulation, highlighting the importance of studying disease-related genes across 'omic datasets. We observe downregulation of integrin α1 and β5 subunits at the RNA level (ITGA1A and ITGA5B) and integrin α1 and β1 subunits at the protein level (ITGA1A and ITGA1B). Published findings have established a profibrotic role of integrins αvβ1 and αvβ6 integrins at the protein level by activation of TGF-β (58); more work is needed to understand the roles of integrins we identified in lung fibrosis.

Effect of DNA Methylation on Gene Expression

We next assessed the effect of DNA methylation on expression of nearest protein coding RNA (**Figure 3A**), alternatively spliced RNA **(Figure 3B)**, and protein **(Figure 3C)**. We observed only a few relationships within 10 kb among significantly differentially expressed genes, acknowledging that methylation marks do not always regulate the nearest gene (59). Inversely correlated methylation and expression were observed for genes of interest in IPF such as AGER (60), alpha catenin 2 (61), KRT17 (62), and CASZ1 (a gene we previously reported as regulated by methylation in IPF(24)). Increasing the distance of overlap to 100 kb reveals many more potential cis relationships between DNA methylation and gene expression. A potentially interesting novel finding in these data is regulation by methylation of COL17A1, a transmembrane protein that is a structural component of hemidesmosomes and has been reported to be regulated

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

by promoter methylation in epithelial cancers (63). Even with the limitation of only focusing on relationships with nearest genes, DNA methylation appears to be an important feature of gene regulation in IPF.

Multi-Omic Modeling

To fully integrate all four datasets (protein coding RNA, protein, DNA methylation, and noncoding RNA), we used the Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO) multi-omics integrative method. The DIABLO model differentiates IPF and control lungs using one latent variable (**Supplemental Figure S4A**), demonstrates strong correlations of individual features with the top latent variable (**Supplemental Figure S4B**) as well as strong correlations across features from different datasets (**Supplemental Figure S4C**). Contributions of individual dataset features on the top latent variable are shown in **Supplemental Figure S4D** and in **Supplemental Table S2A-D**. Unsupervised hierarchical clustering using 20 coding mRNA, protein, methylation, and lncRNA features with highest loadings on the top latent variable from the four datasets (80 features total; **Figure 4A**) demonstrates perfect separation of IPF and control lungs (**Figure 4B**). Among the top protein-coding mRNA features are MMP7, a key biomarker of pulmonary fibrosis (13, 47); PROM2, a gene expressed in basal cells that differentiates airway from alveolar transcriptional subtype of IPF (19); COL17A1, discussed above; and LAMC3, a focal adhesion gene. Among the top protein features are periostin, a protein that promotes myofibroblast differentiation and type 1 collagen production (64); palladin, a protein involved in cell adhesion; AGER, a gene polymorphic in IPF that encodes soluble RAGE decoy receptor (60); focal adhesion proteins LAMC2 and ITGA3; and PECAM1, a protein involved in leukocyte migration, angiogenesis, and integrin activation. Among the top DNA methylation features (all hypomethylated) are DMRs 5' to MIR21, a key profibrotic miRNA upregulated in IPF (29); the promoter of CCL2, a T-cell recruiting chemokine with an established role in IPF (65); the promoter of TNXB, a gene that has been reported hypomethylated and upregulated in IPF

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

fibroblasts (66); and in an intron of the LTBP1 gene that is upregulated in IPF, especially honeycomb cysts, and regulates the effects of TGF-β1 (67). Long non-coding RNA data are more difficult to dissect because of currently unknown functions of many of the lncRNAs. Of the top 20 lncRNAs, RARA-AS1 is promising as a potential regulator of RARA, a gene that has been shown downregulated in IPF fibroblasts (68). LINC01565 or GR6 is another potential candidate based on its expression patterns (highest in lung and bone marrow), but no studies have shown its role in fibrosis at this time. MIR34AHG is the host gene for miR-34 which has been shown to regulate cellular senescence in IPF type II alveolar epithelial cells (69). FENDRR, a lncRNA previously associated with IPF (33) is 28th on the extended list of features ranked by the strength of association with the top latent variable (**Supplemental Table S2D**) and we observe hypermethylation of a GeneHancer predicted FENDRR enhancer.

In general, we observed strong correlations among features from different 'omic' platforms that were prioritized by the DIABLO model, as would be expected. Many of the DNA methylation marks are negatively correlated to protein-coding RNAs as well as long noncoding RNAs (**Figure 4C**). This led us to construct a network of top 20 features from each of the individual datasets (**Figure 4D**). MMP7 RNA, for example, has strong correlations ($|r|>0.8$) with four proteins (ASH1L, BRAP, RHAG; all negative); 18 DMRs (all negative); negative correlations with three lncRNAs (AP001189.3, GATA2-AS1, and RARA-AS1) and positive correlations with seven lncRNAs (AC007552.2, AC007996.1, AC097478.1, LINC01480, MAST4-AS1, SMC5-AS1, and TMEM161B-AS1). MMP7 illustrates how this multi-omic approach may uncover novel relationships that will require additional computational (replication) and experimental (functional) validation.

Validation of the Multi-Omic Model

We used an unsupervised approach Multi-Omics Factor Analysis (MOFA) to independently identify the principal sources of variation in our multi-omics datasets. Results of

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

this confirmatory analysis are discussed and summarized in the Online Supplement (**Supplemental Figure S5** and **Supplemental Table S4**). Overall, a number of the same transcriptome and proteome features emerge as prioritized by both DIABLO and MOFA multi-omic methods while more work will be needed to further assess reproducibility of regulatory features of the transcriptome (DNA methylation and lncRNAs).

DISCUSSION

We present the first application of multi-omic integration modalities to IPF lung tissue, leveraging coding and non-coding RNA expression, proteomic, and DNA methylation data to construct a multi-omic network to gain insights into relevant pathogenic molecules and pathways in disease. Our analyses confirm previously validated molecules and pathways known to be dysregulated in disease, and implicate novel molecular features as potential drivers and modifiers of disease.

The multi-omic model provides a more complete characterization of molecular pathways in IPF and insight into the complex biology of IPF. It also provides better power for identification of such pathways by analysis of multiple datasets on the same tissue, as has been previously shown in other settings (38). Multi-omic analysis may be used to identify pathways that are dysregulated in IPF both at the transcript and protein level, such as focal adhesion, or regulators of genes/proteins already identified as important in IPF. For example, our multi-omic model indicated that 18 regions of differential methylation and 10 lncRNAs may be important in regulation of MMP7, a gene known with a known role in IPF. Interestingly, RARA-AS1, a potential regulator of RARA, a gene that has been shown downregulated in IPF fibroblasts (68), is highly negatively correlated with MMP7, suggesting that crosstalk between RARA and MMP7 may provide novel targets for IPF therapeutics. However, additional computational (replication) and experimental (functional) validation is needed.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

Given that IPF is a complex disease biologically (2), it is not surprising that narrowly targeted drugs have generally failed in IPF and currently approved drugs target multiple pathways, such as inhibition of several growth factor with Nintedanib (70). Analyzing proteomic, transcriptomic and regulatory (methylome and non-coding transcriptome) molecules at the same time provides a more complete picture of IPF pathophysiology and will raise the interest for new compounds in IPF. Data mining has become an important research direction in drug discovery that should take advantage of multi-omics analysis (71). Future work in IPF will likely apply these multi-omic methods to longitudinally collected data in peripheral blood, capturing earlier or preclinical stages of disease for development of treatments that can be used before the fibrotic process involves large portions of the lung and becomes irreversible (72).

By study design, cases were evenly distributed between patients homozygous for the major allele and patients heterozygous for the *MUC5B* promoter polymorphism rs3570950 to allow for examination of multi-omic signal in relation to the *MUC5B* variant. However, differential testing within datasets on the basis of the variant did not yield statistically significant results; much larger sample sizes will likely be needed to detect the effect of the variant in mixed tissue given that the small number of distal airway epithelial cells in which the variant exhibits the strongest effect (73, 74). Furthermore, while MUC5B transcript and protein are upregulated in IPF (7.7-fold at the transcript level and 2-fold at the protein level in our samples), MUC5B transcript and protein are not differentially abundant at FDR<0.05, due to heterogeneity in abundance among cases in whole lung tissue. This could also explain why MUC5B-AS1, one of the top differentially abundant (increased) antisense RNAs in our RNA-seq data, is not a top weighted feature in the DIABLO model. These results are not surprising due to our limited sample size. Further studies examining *MUC5B* genotype contributions to molecular signaling will have increased power if cell types of interest can be isolated through microdissection or cell selection/enrichment methodologies.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

We applied stringent corrections for bias and inflation using methodology specifically developed for transcriptome- and epigenome-wide association (TWAS and EWAS) studies. Methods developed for genome-wide association studies (GWAS) assume a Gaussian distribution of test statistics; this is a valid assumption in GWAS as the vast majority of (generally binary) variants are not expected to be associated with the trait of interest. In TWAS and EWAS of complex traits such as IPF, it is common to identify changes in hundreds to thousands of features, most of which are likely to be true associations but some may be spurious due to inherit inflation that has been documented in TWAS and EWAS (43). Because of this, we applied stringent corrections within our data for bias and inflation, using Bacon (43) to empirically derive a null testing distribution from our data, which takes into account the non-normal mean and variance of the data. This greatly reduced the number of differential features meeting significance within our datasets, compared to previous publications (19). However, some residual inflation remains in the DNA methylation dataset, an issue that is common in the field (75).

Future multi-omic research in IPF should attempt to increase power, as well as the genetic context of these molecular patterns. Larger cohort studies will provide the power to derive and then test and validate sparse multi-omic signatures for replication in independent samples. Larger numbers will also allow for clustering of IPF cases into potentially meaningful subgroups. The inclusion of genetic data, which explains a significant proportion of disease variability, will aid in patient clustering and recognition of distinct molecular subtypes. These improved integrative models hold promise to focus our attention on key molecules and pathways involved in the complex biology of lung fibrosis, and potentially identify critical checkpoints that can manipulated pharmacologically.

ACKNOWLEDGMENTS

We would like to thank the University of Colorado Genomics and Microarray Core Facility for collection of Illumina BeadChip and RNA sequencing data.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

FUNDING

Funded by the NIH-NHLBI P01-HL092870. IRK was supported by the University of Colorado Team Oriented Training across the Translational Sciences Spectrum (TOTTS) program (NIH-NCATS TL1-TR002533). RB was supported by the European Respiratory Society Long Term Research Fellowship.

COMPETING INTEREST

DAS is the founder and CSO of Eleven P15, a startup company that focuses on early detection and treatment of pulmonary fibrosis. IVY and TEF are consultants to Eleven P15. Eleven P15 did not have any scientific nor financial influence on this study.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

TABLES

Table 1. Clinical Characteristics of the subjects included in the +multi-omic profiling.

	IPF (n = 24)	**Controls (n = 14)**	**p**
Age	62 +- 5.9	64 +- 5.7	0.323*
Sex (M)	20 (83.3%)	10 (71.4%)	0.433†
Race (W)	24 (100%)	14 (100%)	1†
Ethnicity (NH)	24 (100%)	14 (100%)	1†
Smoking Status (Ever)	17 (70.1%)	9 (64.3%)	0.521†
MUC5B Genotype (GG)	13 (50%)	14 (100%)	NA^

*Assessed w/ student's t test

†Assessed w/ Fisher's Exact Test

^By design

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

FIGURE LEGENDS

Figure 1. Volcano plots depicting features statistically significant in IPF compared to control lung tissue at false discovery rate (FDR<0.05) (blue dots). Protein coding and alternatively spliced RNA were captured by polyA/mRNA-seq while lncRNAs were captured by ribosomal RNA (rRNA)-depleted sequencing. Alternative splicing includes retained exon annotations from Gencode. Noncoding RNA includes lincRNA, antisense RNA, miscellaneous RNA, sense intronic, snRNA, miRNA, snoRNA, sense overlapping, bidirectional promoter lncRNA, 3prime overlapping ncRNA, scaRNA, ribozyme, noncoding, macro lncRNA, scRNA, and vaultRNA Gencode annotations. All data other than the proteome dataset were adjusted for bias and inflation (43). Protein data were not adjusted for bias and inflation because of an inherent bias in the proteomics assay focusing on proteins/peptides known to be involved in IPF, therefore, inflation is expected in this dataset.

Figure 2. Comparison of protein coding mRNA and protein datasets. **(A)** Protein fold change plotted against fold change for the corresponding protein coding mRNA. Transcript/proteins with absolute fold change >4 (2 on the log2 scale) are highlighted. **(B)** KEGG pathway enrichment in mRNA (left) and protein (right) datasets. Boxes highlight pathways of interest in common to the two datasets. **(C)** mRNAs/proteins in the Focal Adhesion KEGG pathway are highly dysregulated in IPF lung tissue. Red represents upregulation and green represents downregulation.

Figure 3. The effect of DNA methylation on dysregulated gene expression in IPF lung tissue. **(A)** Protein coding mRNA, **(B)** alternatively spliced mRNA (retain intron), and **(C)** protein fold change plotted against % change in DNA methylation in DMRs assigned to the same genes. All fold changes are presented on the log2 scale. DNA methylation changes are presented as % methylation changes (on the scale 0-1).

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

Figure 4. DIABLO multi-omic model results for the top 20 features in each dataset (blue = coding RNA, green = protein, red = methylation, orange = noncoding RNA). IPF lung is represented in blue and control in orange. **(A)** Top 20 features from each dataset contributing to the top latent component. **(B)** Clustering of cases vs controls based on top 20 features from each dataset. **(C)** Circos plot of correlations ($|r|>0.8$) for all features contributing to the top latent component. Red lines represent positive correlations and blue lines represent negative correlations. **(D)** An interactome network of top features from each of the individual datasets. Red lines represent positive correlations and blue lines represent negative correlations.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

REFERENCES

1. Olson AL, Swigris JJ, Lezotte DC, Norris JM, Wilson CG, Brown KK. Mortality from pulmonary fibrosis increased in the united states from 1992 to 2003. *Am J Respir Crit Care Med* 2007;176:277-284.

2. Lederer DJ, Martinez FJ. Idiopathic pulmonary fibrosis. *N Engl J Med* 2018;378:1811-1823.

3. Navaratnam V, Fleming KM, West J, Smith CJ, Jenkins RG, Fogarty A, Hubbard RB. The rising incidence of idiopathic pulmonary fibrosis in the u.K. *Thorax* 2011;66:462-467.

4. Hunninghake GM, Hatabu H, Okajima Y, Gao W, Dupuis J, Latourelle JC, Nishino M, Araki T, Zazueta OE, Kurugol S, Ross JC, San Jose Estepar R, Murphy E, Steele MP, Loyd JE, Schwarz MI, Fingerlin TE, Rosas IO, Washko GR, O'Connor GT, Schwartz DA. Muc5b promoter polymorphism and interstitial lung abnormalities. *N Engl J Med* 2013;368:2192-2200.

5. Putman RK, Hatabu H, Araki T, Gudmundsson G, Gao W, Nishino M, Okajima Y, Dupuis J, Latourelle JC, Cho MH, El-Chemaly S, Coxson HO, Celli BR, Fernandez IE, Zazueta OE, Ross JC, Harmouche R, Estepar RS, Diaz AA, Sigurdsson S, Gudmundsson EF, Eiriksdottir G, Aspelund T, Budoff MJ, Kinney GL, Hokanson JE, Williams MC, Murchison JT, MacNee W, Hoffmann U, O'Donnell CJ, Launer LJ, Harrris TB, Gudnason V, Silverman EK, O'Connor GT, Washko GR, Rosas IO, Hunninghake GM. Association between interstitial lung abnormalities and all-cause mortality. *JAMA* 2016;315:672-681.

6. 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;155:242-248.

7. Allen RJ, Guillen-Guio B, Oldham JM, Ma SF, Dressen A, Paynton ML, Kraven LM, Obeidat M, Li X, Ng M, Braybrooke R, Molina-Molina M, Hobbs BD, Putman RK, Sakornsakolpat

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

P, Booth HL, Fahy WA, Hart SP, Hill MR, Hirani N, Hubbard RB, McAnulty RJ, Millar AB, Navaratnam V, Oballa E, Parfrey H, Saini G, Whyte MKB, Zhang Y, Kaminski N, Adegunsoye A, Strek ME, Neighbors M, Sheng XR, Gudmundsson G, Gudnason V, Hatabu H, Lederer DJ, Manichaikul A, Newell JD, Jr., O'Connor GT, Ortega VE, Xu H, Fingerlin TE, Bosse Y, Hao K, Joubert P, Nickle DC, Sin DD, Timens W, Furniss D, Morris AP, Zondervan KT, Hall IP, Sayers I, Tobin MD, Maher TM, Cho MH, Hunninghake GM, Schwartz DA, Yaspan BL, Molyneaux PL, Flores C, Noth I, Jenkins RG, Wain LV. Genome-wide association study of susceptibility to idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2020;201:564-574.

8. Moore C, Blumhagen RZ, Yang IV, Walts A, Powers J, Walker T, Bishop M, Russell P, Vestal B, Cardwell J, Markin CR, Mathai SK, Schwarz MI, Steele MP, Lee J, Brown KK, Loyd JE, Crapo JD, Silverman EK, Cho MH, James JA, Guthridge JM, Cogan JD, Kropski JA, Swigris JJ, Bair C, Kim DS, Ji W, Kim H, Song JW, Maier LA, Pacheco KA, Hirani N, Poon AS, Li F, Jenkins RG, Braybrooke R, Saini G, Maher TM, Molyneaux PL, Saunders P, Zhang Y, Gibson KF, Kass DJ, Rojas M, Sembrat J, Wolters PJ, Collard HR, Sundy JS, O'Riordan T, Strek ME, Noth I, Ma SF, Porteous MK, Kreider ME, Patel NB, Inoue Y, Hirose M, Arai T, Akagawa S, Eickelberg O, Fernandez IE, Behr J, Mogulkoc N, Corte TJ, Glaspole I, Tomassetti S, Ravaglia C, Poletti V, Crestani B, Borie R, Kannengiesser C, Parfrey H, Fiddler C, Rassl D, Molina-Molina M, Machahua C, Worboys AM, Gudmundsson G, Isaksson HJ, Lederer DJ, Podolanczuk AJ, Montesi SB, Bendstrup E, Danchel V, Selman M, Pardo A, Henry MT, Keane MP, Doran P, Vasakova M, Sterclova M, Ryerson CJ, Wilcox PG, Okamoto T, Furusawa H, Miyazaki Y, Laurent G, Baltic S, Prele C, Moodley Y, Shea BS, Ohta K, Suzukawa M, Narumoto O, Nathan SD, Venuto DC, Woldehanna ML, Kokturk N, de Andrade JA, Luckhardt T, Kulkarni T, Bonella F, Donnelly SC, McElroy A, Armstong ME, Aranda A, Carbone RG, Puppo F, Beckman KB, Nickerson DA, Fingerlin TE, Schwartz DA. Resequencing study confirms that host defense and cell senescence

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

gene variants contribute to the risk of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2019;200:199-208.

9. Mathai SK, Newton CA, Schwartz DA, Garcia CK. Pulmonary fibrosis in the era of stratified medicine. *Thorax* 2016;71:1154-1160.

10. King TE, Jr., Bradford WZ, Castro-Bernardini S, Fagan EA, Glaspole I, Glassberg MK, Gorina E, Hopkins PM, Kardatzke D, Lancaster L, Lederer DJ, Nathan SD, Pereira CA, Sahn SA, Sussman R, Swigris JJ, Noble PW, Group AS. A phase 3 trial of pirfenidone in patients with idiopathic pulmonary fibrosis. *N Engl J Med* 2014;370:2083-2092.

11. Richeldi L, du Bois RM, Raghu G, Azuma A, Brown KK, Costabel U, Cottin V, Flaherty KR, Hansell DM, Inoue Y, Kim DS, Kolb M, Nicholson AG, Noble PW, Selman M, Taniguchi H, Brun M, Le Maulf F, Girard M, Stowasser S, Schlenker-Herceg R, Disse B, Collard HR, Investigators IT. Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis. *N Engl J Med* 2014;370:2071-2082.

12. King TE, Jr., Pardo A, Selman M. Idiopathic pulmonary fibrosis. *Lancet* 2011;378:1949-1961.

13. Kaminski N. Microarray analysis of idiopathic pulmonary fibrosis. *Am J Respir Cell Mol Biol* 2003;29:S32-36.

14. Konishi K, Gibson KF, Lindell KO, Richards TJ, Zhang Y, Dhir R, Bisceglia M, Gilbert S, Yousem SA, Song JW, Kim DS, Kaminski N. Gene expression profiles of acute exacerbations of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2009;180:167-175.

15. Selman M, Carrillo G, Estrada A, Mejia M, Becerril C, Cisneros J, Gaxiola M, Perez-Padilla R, Navarro C, Richards T, Dauber J, King TE, Jr., Pardo A, Kaminski N. Accelerated variant of

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

idiopathic pulmonary fibrosis: Clinical behavior and gene expression pattern. *PLoS One* 2007;2:e482.

16. Selman M, Pardo A, Barrera L, Estrada A, Watson SR, Wilson K, Aziz N, Kaminski N, Zlotnik A. Gene expression profiles distinguish idiopathic pulmonary fibrosis from hypersensitivity pneumonitis. *Am J Respir Crit Care Med* 2006;173:188-198.

17. Zuo F, Kaminski N, Eugui E, Allard J, Yakhini Z, Ben-Dor A, Lollini L, Morris D, Kim Y, DeLustro B, Sheppard D, Pardo A, Selman M, Heller RA. Gene expression analysis reveals matrilysin as a key regulator of pulmonary fibrosis in mice and humans. *Proc Natl Acad Sci U S A* 2002;99:6292-6297.

18. Boon K, Bailey NW, Yang J, Steel MP, Groshong S, Kervitsky D, Brown KK, Schwarz MI, Schwartz DA. Molecular phenotypes distinguish patients with relatively stable from progressive idiopathic pulmonary fibrosis (ipf). *PLoS One* 2009;4:e5134.

19. Yang IV, Coldren CD, Leach SM, Seibold MA, Murphy E, Lin J, Rosen R, Neidermyer AJ, McKean DF, Groshong SD, Cool C, Cosgrove GP, Lynch DA, Brown KK, Schwarz MI, Fingerlin TE, Schwartz DA. Expression of cilium-associated genes defines novel molecular subtypes of idiopathic pulmonary fibrosis. *Thorax* 2013;68:1114-1121.

20. Yang IV, Burch LH, Steele MP, Savov JD, Hollingsworth JW, McElvania-Tekippe E, Berman KG, Speer MC, Sporn TA, Brown KK, Schwarz MI, Schwartz DA. Gene expression profiling of familial and sporadic interstitial pneumonia. *Am J Respir Crit Care Med* 2007;175:45-54.

21. Schiller HB, Mayr CH, Leuschner G, Strunz M, Staab-Weijnitz C, Preisendorfer S, Eckes B, Moinzadeh P, Krieg T, Schwartz DA, Hatz RA, Behr J, Mann M, Eickelberg O. Deep proteome profiling reveals common prevalence of mzb1-positive plasma b cells in human lung and skin fibrosis. *Am J Respir Crit Care Med* 2017;196:1298-1310.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

22. Sanders YY, Ambalavanan N, Halloran B, Zhang X, Liu H, Crossman DK, Bray M, Zhang K, Thannickal VJ, Hagood JS. Altered DNA methylation profile in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2012;186:525-535.

23. Rabinovich EI, Kapetanaki MG, Steinfeld I, Gibson KF, Pandit KV, Yu G, Yakhini Z, Kaminski N. Global methylation patterns in idiopathic pulmonary fibrosis. *PLoS One* 2012;7:e33770.

24. Yang IV, Pedersen BS, Rabinovich E, Hennessy CE, Davidson EJ, Murphy E, Guardela BJ, Tedrow JR, Zhang Y, Singh MK, Correll M, Schwarz MI, Geraci M, Sciurba FC, Quackenbush J, Spira A, Kaminski N, Schwartz DA. Relationship of DNA methylation and gene expression in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2014;190:1263-1272.

25. Tager AM, Kradin RL, LaCamera P, Bercury SD, Campanella GS, Leary CP, Polosukhin V, Zhao LH, Sakamoto H, Blackwell TS, Luster AD. Inhibition of pulmonary fibrosis by the chemokine ip-10/cxcl10. *Am J Respir Cell Mol Biol* 2004;31:395-404.

26. Huang SK, Fisher AS, Scruggs AM, White ES, Hogaboam CM, Richardson BC, Peters-Golden M. Hypermethylation of ptger2 confers prostaglandin e2 resistance in fibrotic fibroblasts from humans and mice. *Am J Pathol* 2010;177:2245-2255.

27. Sanders YY, Pardo A, Selman M, Nuovo GJ, Tollefsbol TO, Siegal GP, Hagood JS. Thy-1 promoter hypermethylation: A novel epigenetic pathogenic mechanism in pulmonary fibrosis. *Am J Respir Cell Mol Biol* 2008;39:610-618.

28. Cushing L, Kuang PP, Qian J, Shao F, Wu J, Little F, Thannickal VJ, Cardoso WV, Lu J. Mir-29 is a major regulator of genes associated with pulmonary fibrosis. *Am J Respir Cell Mol Biol* 2011;45:287-294.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

29. Liu G, Friggeri A, Yang Y, Milosevic J, Ding Q, Thannickal VJ, Kaminski N, Abraham E. Mir-21 mediates fibrogenic activation of pulmonary fibroblasts and lung fibrosis. *J Exp Med* 2010;207:1589-1597.

30. Oak SR, Murray L, Herath A, Sleeman M, Anderson I, Joshi AD, Coelho AL, Flaherty KR, Toews GB, Knight D, Martinez FJ, Hogaboam CM. A micro rna processing defect in rapidly progressing idiopathic pulmonary fibrosis. *PLoS One* 2011;6:e21253.

31. Pottier N, Maurin T, Chevalier B, Puissegur MP, Lebrigand K, Robbe-Sermesant K, Bertero T, Lino Cardenas CL, Courcot E, Rios G, Fourre S, Lo-Guidice JM, Marcet B, Cardinaud B, Barbry P, Mari B. Identification of keratinocyte growth factor as a target of microrna-155 in lung fibroblasts: Implication in epithelial-mesenchymal interactions. *PLoS One* 2009;4:e6718.

32. Pandit KV, Corcoran D, Yousef H, Yarlagadda M, Tzouvelekis A, Gibson KF, Konishi K, Yousem SA, Singh M, Handley D, Richards T, Selman M, Watkins SC, Pardo A, Ben-Yehudah A, Bouros D, Eickelberg O, Ray P, Benos PV, Kaminski N. Inhibition and role of let-7d in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2010;182:220-229.

33. Huang C, Liang Y, Zeng X, Yang X, Xu D, Gou X, Sathiaseelan R, Senavirathna LK, Wang P, Liu L. Lncrna fendrr exhibits anti-fibrotic activity in pulmonary fibrosis. *Am J Respir Cell Mol Biol* 2019.

34. Savary G, Dewaeles E, Diazzi S, Buscot M, Nottet N, Fassy J, Courcot E, Henaoui IS, Lemaire J, Martis N, Van der Hauwaert C, Pons N, Magnone V, Leroy S, Hofman V, Plantier L, Lebrigand K, Paquet A, Lino Cardenas CL, Vassaux G, Hofman P, Gunther A, Crestani B, Wallaert B, Rezzonico R, Brousseau T, Glowacki F, Bellusci S, Perrais M, Broly F, Barbry P, Marquette CH, Cauffiez C, Mari B, Pottier N. The long noncoding rna dnm3os is a reservoir of fibromirs with major functions in lung fibroblast response to tgf-beta and pulmonary fibrosis. *Am J Respir Crit Care Med* 2019;200:184-198.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

35. Yang IV, Luna LG, Cotter J, Talbert J, Leach SM, Kidd R, Turner J, Kummer N, Kervitsky D, Brown KK, Boon K, Schwarz MI, Schwartz DA, Steele MP. The peripheral blood transcriptome identifies the presence and extent of disease in idiopathic pulmonary fibrosis. *PLoS One* 2012;7:e37708.

36. Herazo-Maya JD, Sun J, Molyneaux PL, Li Q, Villalba JA, Tzouvelekis A, Lynn H, Juan-Guardela BM, Risquez C, Osorio JC, Yan X, Michel G, Aurelien N, Lindell KO, Klesen MJ, Moffatt MF, Cookson WO, Zhang Y, Garcia JGN, Noth I, Prasse A, Bar-Joseph Z, Gibson KF, Zhao H, Herzog EL, Rosas IO, Maher TM, Kaminski N. 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;5:857-868.

37. Herazo-Maya JD, Noth I, Duncan SR, Kim S, Ma SF, Tseng GC, Feingold E, Juan-Guardela BM, Richards TJ, Lussier Y, Huang Y, Vij R, Lindell KO, Xue J, Gibson KF, Shapiro SD, Garcia JG, Kaminski N. Peripheral blood mononuclear cell gene expression profiles predict poor outcome in idiopathic pulmonary fibrosis. *Sci Transl Med* 2013;5:205ra136.

38. Lee AH, Shannon CP, Amenyogbe N, Bennike TB, Diray-Arce J, Idoko OT, Gill EE, Ben-Othman R, Pomat WS, van Haren SD, Cao KL, Cox M, Darboe A, Falsafi R, Ferrari D, Harbeson DJ, He D, Bing C, Hinshaw SJ, Ndure J, Njie-Jobe J, Pettengill MA, Richmond PC, Ford R, Saleu G, Masiria G, Matlam JP, Kirarock W, Roberts E, Malek M, Sanchez-Schmitz G, Singh A, Angelidou A, Smolen KK, Consortium E, Brinkman RR, Ozonoff A, Hancock REW, van den Biggelaar AHJ, Steen H, Tebbutt SJ, Kampmann B, Levy O, Kollmann TR. Dynamic molecular changes during the first week of human life follow a robust developmental trajectory. *Nat Commun* 2019;10:1092.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

39. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, Le Cao KA. Diablo: An integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* 2019;35:3055-3062.

40. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* 2018;14:e8124.

41. Lepper MF, Ohmayer U, von Toerne C, Maison N, Ziegler AG, Hauck SM. Proteomic landscape of patient-derived cd4+ t cells in recent-onset type 1 diabetes. *J Proteome Res* 2018;17:618-634.

42. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 2012;28:882-883.

43. van Iterson M, van Zwet EW, Consortium B, Heijmans BT. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol* 2017;18:19.

44. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)* 1995;57:289-300.

45. Aran D, Hu Z, Butte AJ. Xcell: Digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol* 2017;18:220.

46. Richards TJ, Kaminski N, Baribaud F, Flavin S, Brodmerkel C, Horowitz D, Li K, Choi J, Vuga LJ, Lindell KO, Klesen M, Zhang Y, Gibson KF. Peripheral blood proteins predict mortality in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2012;185:67-76.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

47. Rosas IO, Richards TJ, Konishi K, Zhang Y, Gibson K, Lokshin AE, Lindell KO, Cisneros J, Macdonald SD, Pardo A, Sciurba F, Dauber J, Selman M, Gochuico BR, Kaminski N. Mmp1 and mmp7 as potential peripheral blood biomarkers in idiopathic pulmonary fibrosis. *PLoS Med* 2008;5:e93.

48. Ian T. Stancil, Jacob E. Michalski, Duncan Davis-Hall, Chelsea M. Magin, Ivana V. Yang, Evgenia Dobrinskikh, Bradford J. Smith, Schwartz DA. Aberrant airway epithelial function and signaling promote lung fibrosis. *Nature Communications* 2021:in revision.

49. Ikeda T, Fragiadaki M, Shi-Wen X, Ponticos M, Khan K, Denton C, Garcia P, Bou-Gharios G, Yamakawa A, Morimoto C, Abraham D. Transforming growth factor-beta-induced cux1 isoforms are associated with fibrosis in systemic sclerosis lung fibroblasts. *Biochem Biophys Rep* 2016;7:246-252.

50. Ikeda T, Fragiadaki M, Shi-Wen X, Ponticos M, Khan K, Denton C, Garcia P, Bou-Gharios G, Yamakawa A, Morimoto C, Abraham D. Data on cux1 isoforms in idiopathic pulmonary fibrosis lung and systemic sclerosis skin tissue sections. *Data Brief* 2016;8:1377-1380.

51. Houseman EA, Kile ML, Christiani DC, Ince TA, Kelsey KT, Marsit CJ. Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics* 2016;17:259.

52. Li S, Wang Y, Zhang Y, Lu MM, DeMayo FJ, Dekker JD, Tucker PW, Morrisey EE. Foxp1/4 control epithelial cell fate during lung development and regeneration through regulation of anterior gradient 2. *Development* 2012;139:2500-2509.

53. Osafo-Addo AD, Herzog EL. Ccl2 and t cells in pulmonary fibrosis: An old player gets a new role. *Thorax* 2017;72:967-968.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

54. Zhang Y, He Q, Hu Z, Feng Y, Fan L, Tang Z, Yuan J, Shan W, Li C, Hu X, Tanyi JL, Fan Y, Huang Q, Montone K, Dang CV, Zhang L. Long noncoding rna linp1 regulates repair of DNA double-strand breaks in triple-negative breast cancer. *Nat Struct Mol Biol* 2016;23:522-530.

55. Wang S, Jiang W, Zhang X, Lu Z, Geng Q, Wang W, Li N, Cai X. Linc-pint alleviates lung cancer progression via sponging mir-543 and inducing pten. *Cancer Med* 2020.

56. Moore BB, Moore TA. Viruses in idiopathic pulmonary fibrosis. Etiology and exacerbation. *Ann Am Thorac Soc* 2015;12 Suppl 2:S186-192.

57. Sheng G, Chen P, Wei Y, Yue H, Chu J, Zhao J, Wang Y, Zhang W, Zhang H-L. Viral infection increases the risk of idiopathic pulmonary fibrosis. *Chest* 2019.

58. Friedman SL, Sheppard D, Duffield JS, Violette S. Therapy for fibrotic diseases: Nearing the starting line. *Sci Transl Med* 2013;5:167sr161.

59. Schoenfelder S, Fraser P. Long-range enhancer-promoter contacts in gene expression control. *Nat Rev Genet* 2019;20:437-455.

60. Yamaguchi K, Iwamoto H, Horimasu Y, Ohshimo S, Fujitaka K, Hamada H, Mazur W, Kohno N, Hattori N. Agergene polymorphisms and soluble receptor for advanced glycation end product in patients with idiopathic pulmonary fibrosis. *Respirology* 2017;22:965-971.

61. Lovgren AK, Kovacs JJ, Xie T, Potts EN, Li Y, Foster WM, Liang J, Meltzer EB, Jiang D, Lefkowitz RJ, Noble PW. Beta-arrestin deficiency protects against pulmonary fibrosis in mice and prevents fibroblast invasion of extracellular matrix. *Sci Transl Med* 2011;3:74ra23.

62. Habermann AC, Gutierrez AJ, Bui LT, Yahn SL, Winters NI, Calvi CL, Peter L, Chung M-I, Taylor CJ, Jetter C, Raju L, Roberson J, Ding G, Wood L, Sucre JM, Richmond BW, Serezani AP, McDonnell WJ, Mallal SB, Bacchetta MJ, Loyd JE, Shaver CM, Ware LB, Bremner R, Walia

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

R, Blackwell TS, Banovich NE, Kropski JA. Single-cell rna-sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *bioRxiv* 2019.

63. Thangavelu PU, Krenacs T, Dray E, Duijf PH. In epithelial cancers, aberrant col17a1 promoter methylation predicts its misexpression and increased invasion. *Clin Epigenetics* 2016;8:120.

64. O'Dwyer DN, Moore BB. The role of periostin in lung fibrosis and airway remodeling. *Cell Mol Life Sci* 2017;74:4305-4314.

65. Milger K, Yu Y, Brudy E, Irmler M, Skapenko A, Mayinger M, Lehmann M, Beckers J, Reichenberger F, Behr J, Eickelberg O, Konigshoff M, Krauss-Etschmann S. Pulmonary ccr2(+)cd4(+) t cells are immune regulatory and attenuate lung fibrosis development. *Thorax* 2017;72:1007-1020.

66. Garner IM, Evans IC, Barnes JL, Maher TM, Renzoni EA, Denton CP, Scotton CJ, Abraham DJ, McAnulty RJ. Hypomethylation of the tnxb gene contributes to increased expression and deposition of tenascin-x in idiopathic pulmonary fibrosis. *American Journal of Respiratory and Critical Care Medicine* 2014;189:A3378.

67. Khalil N. Regulation of the effects of tgf-beta 1 by activation of latent tgf-beta 1 and differential expression of tgf-beta receptors (tbeta r-i and tbeta r-ii) in idiopathic pulmonary fibrosis. 2001;56:907-915.

68. Emblom-Callahan MC, Chhina MK, Shlobin OA, Ahmad S, Reese ES, Iyer EP, Cox DN, Brenner R, Burton NA, Grant GM, Nathan SD. Genomic phenotype of non-cultured pulmonary fibroblasts in idiopathic pulmonary fibrosis. *Genomics* 2010;96:134-145.

69. Disayabutr S, Kim EK, Cha SI, Green G, Naikawadi RP, Jones KD, Golden JA, Schroeder A, Matthay MA, Kukreja J, Erle DJ, Collard HR, Wolters PJ. Mir-34 mirnas regulate cellular

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

senescence in type ii alveolar epithelial cells of patients with idiopathic pulmonary fibrosis. *PLoS One* 2016;11:e0158367.

70. Spagnolo P, Tzouvelekis A, Bonella F. The management of patients with idiopathic pulmonary fibrosis. *Front Med (Lausanne)* 2018;5:148.

71. Agatonovic-Kustrin S MD. Data mining in drug discovery and design. In: Puri M PY, Sutariya VK, Tipparaju S, Moreno W, editor. Artif neural netw drug des deliv dispos. Boston, MA: Academic Press; 2016. p. 181–193.

72. Salisbury ML, Hewlett JC, Ding G, Markin CR, Douglas K, Mason W, Guttentag A, Phillips JA, 3rd, Cogan JD, Reiss S, Mitchell DB, Wu P, Young LR, Lancaster LH, Loyd JE, Humphries SM, Lynch DA, Kropski JA, Blackwell TS. Development and progression of radiologic abnormalities in individuals at risk for familial interstitial lung disease. *Am J Respir Crit Care Med* 2020;201:1230-1239.

73. Nakano Y, Yang IV, Walts AD, Watson AM, Helling BA, Fletcher AA, Lara AR, Schwarz MI, Evans CM, Schwartz DA. Muc5b promoter variant rs35705950 affects muc5b expression in the distal airways in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2016;193:464-466.

74. Helling BA, Gerber AN, Kadiyala V, Sasse SK, Pedersen BS, Sparks L, Nakano Y, Okamoto T, Evans CM, Yang IV, Schwartz DA. Regulation of muc5b expression in idiopathic pulmonary fibrosis. *Am J Respir Cell Mol Biol* 2017;57:91-99.

75. Mansell G, Gorrie-Stone TJ, Bao Y, Kumari M, Schalkwyk LS, Mill J, Hannon E. Guidance for DNA methylation studies: Statistical insights from the illumina epic array. *BMC Genomics* 2019;20:366.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

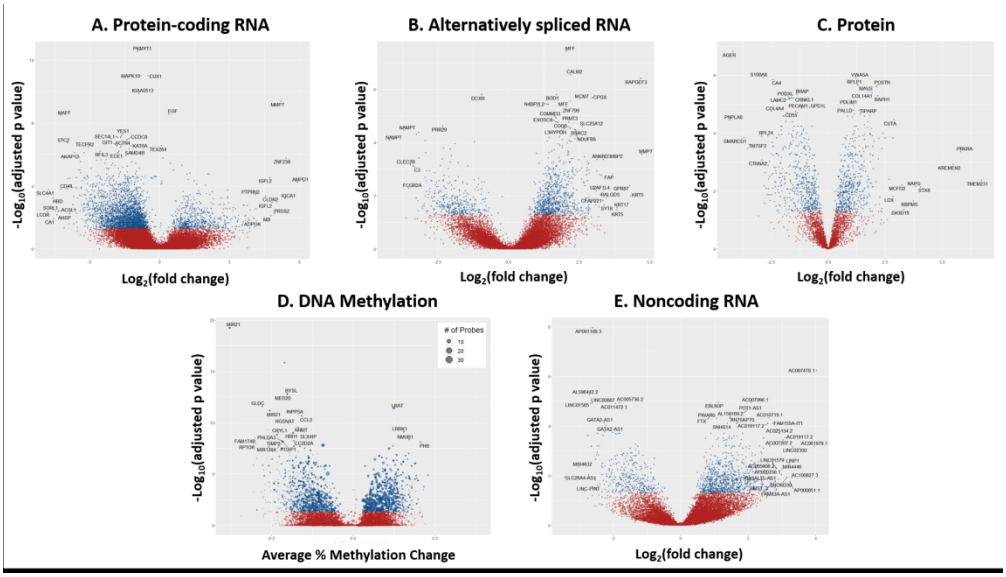


Figure 1

328x186mm (150 x 150 DPI)

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

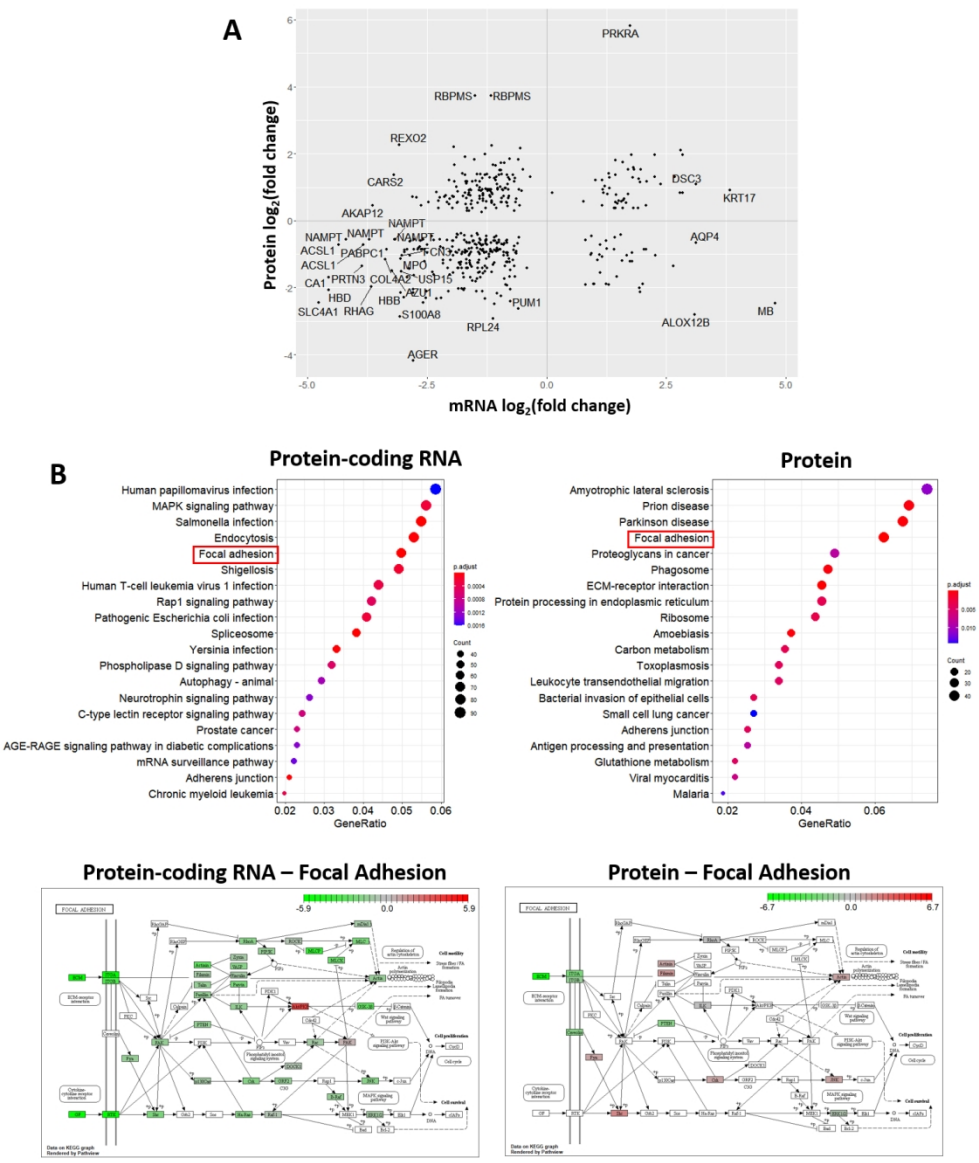


Figure 2

267x311mm (150 x 150 DPI)

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

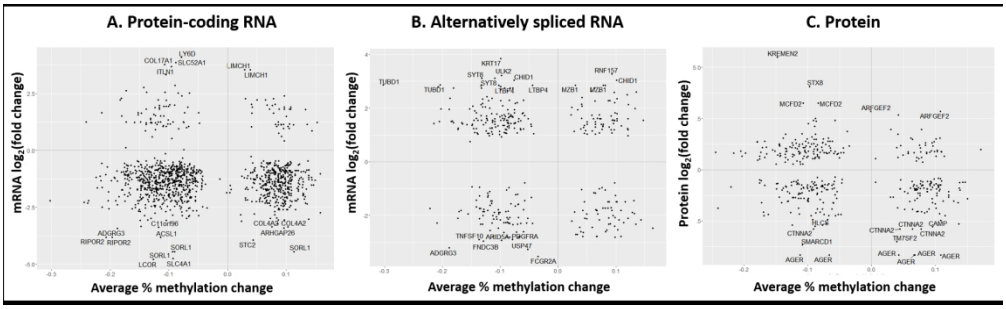


Figure 3

344x103mm (150 x 150 DPI)

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

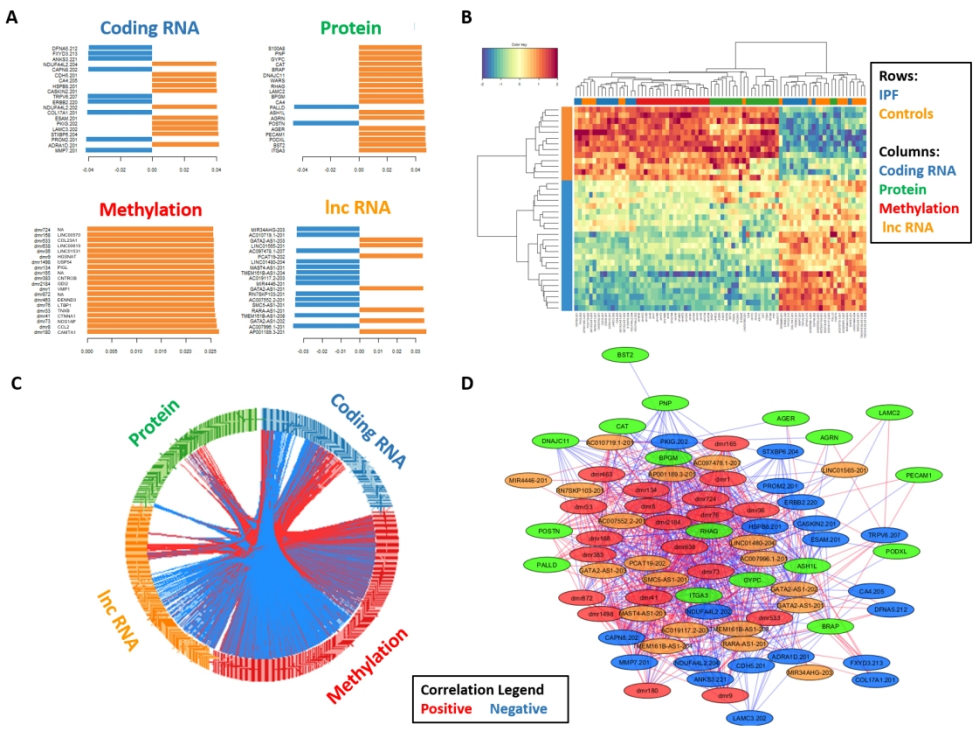


Figure 4

326x243mm (144 x 144 DPI)

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

Molecular Signatures of Idiopathic Pulmonary Fibrosis

SUPPLEMENTAL METHODS AND RESULTS

Iain R. Konigsberg, Raphael Borie, Avram D. Walts, Jonathan Cardwell, Mauricio Rojas, Fabian Metzger, Stefanie M. Hauck, Tasha E. Fingerlin, Ivana V. Yang, David A. Schwartz

METHODS

Ethics Statement

Human tissue was collected after appropriate ethical review for the protection of human subjects through the NHLBI-sponsored Lung Tissue Research Consortium (LTRC) and lung donor program at the University of Pittsburgh. De-identified data and samples were approved for use in this study at the University of Colorado by the Colorado Multiple Institutional Review Board (COMIRB # 15-1147).

Study Population

LTRC enrolled donor subjects prior to lung surgery, collected blood and extensive phenotypic data from the prospective donors, and then processed surgical waste tissue for research use. Clinical data include symptoms, radiologic, and pathological diagnoses, chest CT images, pulmonary function tests (spirometry, D_LCO, and ABG), exposure (including cigarette smoking history) and symptom questionnaires (including Borg dyspnea scale), and family history of lung disease. IPF was diagnosed in LTRC using American Thoracic Society/European respiratory Society (ATS/ERS) criteria (1) and final LTRC diagnosis was based on an integrated clinical, radiologic, and pathologic diagnoses. University of Pittsburgh Lung Donor Core lung tissue was collected from subjects that failed regional criteria for transplants using the same protocols followed for IPF lungs, ensuring technical comparability across IPF and control lung tissue. Potential donor lungs are carefully screened to ensure that there is no evidence of acute

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

or chronic lung disease, that gas exchange is preserved and lung mechanics are normal. Accepted donor lungs are processed with 24 hours of resection. We selected 24 IPF subjects from the LTRC and 14 controls from the University of Pittsburgh Lung Core, all non-Hispanic whites.

Sample Processing

DNA and RNA were isolated from the same sample of lung tissue using the AllPrep kit (Qiagen). Samples with RNA integrity number (RIN) and DNA integrity number (DIN) >5 were used. Genotyping for the *MUC5B* rs35705950 was performed utilizing a TaqMan assay (ThermoFisher). RNA and DNA were evaluated by the University of Colorado Genomics Core. For proteomic analysis, adjacent pieces of lung tissue samples were processed. These samples were washed with cold phosphate-buffered saline (PBS) and transferred to 0.5 ml tubes (Precellys CK14 lysing kit- including 2 x 2.8 mm Zirconium oxide beads, Bertin), containing 200 ul 8M Urea. Lysis was performed in Precellys 24 (Bertin) homogenizer with 20 second cycles at 5000 RPM twice. Total protein content was analyzed by Bradford assay and 10 ug total protein for each sample were proteolyzed by filter-aided sample preparation as previously described (2).

General Statistical Methods

All analyses were performed in R (3.6.2). Principal component analysis was used for quality control and no samples had to be excluded on this criterion. Principal component regression analysis (PCRA) was used to identify variables associated with top principal components, and strong batch effects were regressed out using ComBat (3). Differentially abundant features in each dataset were identified using linear models, adjusting for age and sex. In the mRNA-seq, lncRNA-seq, and DNA methylation datasets, p values were adjusted for inflation and bias using Bacon (4). To control for multiple comparisons, p values were adjusted to a 5% false discovery rate (FDR) using the Benjamini-Hochberg procedure (5) in all datasets.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

Coding Transcriptome

mRNA libraries were prepared from 500 ng total RNA with TruSeq stranded mRNA library preparation kits (illumina) and sequenced at the average depth of 80M reads on the Illumina NovaSeq 6000 (illumina). RNA paired-end reads were aligned at the transcript level to Ensembl GrCh38 using Kallisto (6). 116,503 transcripts were detected in the mRNA dataset using Gencode v27. Differential expression was assessed using DESeq2 (7). 41,121 (35%) transcripts were not included in differential expression testing based on DESq2 cutoff for low expression (defined as 1.9 counts in this dataset based on Cook's distance).

Proteome

22,198 unique peptides associated with 4,011 unique proteins/genes were detected in our samples. Peptides were analyzed on a Q Exactive HF mass spectrometer (ThermoFisher) coupled to an RSLC system (Ultimate 3000) in data-independent acquisition (DIA) mode as described (8). Protein identification was performed with Spectronaut Pulsar by peptide mapping to an in-house spectral library at a precursor Q value cut-off of 0.01 and using the match-between run option at a 0.25 percentile threshold. Label-free quantification was performed on the intensities of summed MS2 fragment spectra. Raw intensity data were normalized via a local retention time-dependent method and log transformation, given the skewness of the data. Differential abundance was tested in Limma (9).

DNA Methylome

1 ug of DNA was bisulfite treated using the Zymo EZ DNA Methylation kit, labeled and hybridized to Illumina Infinium Human MethylationEPIC BeadChip using standard protocols. Illumina idat signal intensity files were processed using seSAMe and Minfi. Probes containing a SNP site (minor allele frequency [MAF] > 1% in the general population) as well as probes with non-unique mapping and off-target hybridization were removed. Additionally, probes with an

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

average detection P value >= .05 across samples and sex chromosome probes were removed prior to analysis. This resulted in 743,256 probes that passed quality control and were tested for association with diagnosis. Array position appeared as a batch effect, and was regressed out of the data using ComBat (3). Differentially methylated positions were tested using M-values in Limma (9). Bacon-adjusted p-values were used to seed DMRs with a window size of 300 bases and a corrected P value <0.05 in comb-p (10) that were subsequently annotated to nearest and overlapping genes in Ensembl GrCh38.

Non-coding Transcriptome (lncRNA)

rRNA-depleted libraries were prepared from 1 ug RNA with the Epidemiolofy Ribo-Zero Gold rRNA Removal Kit (Illumina) and sequenced at the average depth of 80M reads on the Illumina NovaSeq 6000 (Illumina). RNA paired-end reads were aligned at the transcript level to Ensembl GrCh38 using Kallisto (6). Non-coding RNA's were selected for differential testing on the basis of Gencode v27 transcript biotype (11). Transcripts designated as lncRNA (8714), antisense_RNA (8556), misc_RNA (1310), sense_intronic (871), snRNA (795), miRNA (674), snoRNA (480), sense_overlapping (234), bidirectional_promoter_lncRNA (32), 3prime_overlapping_ncRNA (28), scaRNA (27), ribozyme (4), non_coding (3), macro_lncRNA (1), scRNA (1), and valutRNA (1) were included for differential testing. 21,733 transcripts aligning to 14,956 Ensembl non-coding transcript genes identified. Differential expression was assessed using DESeq2 (7).

Enrichment

Differentially abundant features were tested for enrichment of relevant biological gene lists in MSiGDB (12). Specifically, features were tested for enrichment with MSigDB's hallmark gene lists, canonical pathways, and Gene Ontology terms, as well as TISSUE and BTM.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

DIABLO

Data Integration Analysis for Biomarker discovery using Latent variable approaches for Omics studies (DIABLO) (13) was used to determine correlated 'omic' features associated with diagnosis. DIABLO is a supervised learning approach that builds on Regularized and Sparse Generalized Canonical Correlation Analysis (RGCCA), maximizing correlations between multiple datasets containing the same individuals and a classifier (DX). Significantly differentially abundant features from previous analyses were normalized (variance stabilizing transformation for both RNA datasets, protein log2abundance, mean M values for DMR). Using simulations in DIABLO, we determined a single latent variable sufficiently captures most of the variation associated with diagnosis.

MOFA

We also analyzed our multi-omic data with an unsupervised approach Multi-Omics Factor Analysis (MOFA) (14). MOFA learns latent factors that best explain the variance in and between datasets. As such, these factors may represent sources of variation shared between datasets as well as dataset-specific variation.

RESULTS

Validation of the Multi-Omic Model

We used an unsupervised approach Multi-Omics Factor Analysis (MOFA) to independently identify the principal sources of variation in our multi-omics datasets. Unsurprisingly, the top two latent factors identified by MOFA are able to distinguish IPF from control lung tissue (**Supplemental Figure S3**). The first MOFA latent factor (LF1) captures most of the variation in protein-coding RNA, protein, and lncRNA data while latent factor 2 (LF2) captures the majority of variation in the DNA methylation dataset (**Supplemental Figure S3**). To assess reproducibility of the DIABLO model findings in MOFA latent factors, we examined MOFA

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

loadings on LF1 (protein-coding RNA, protein, and lncRNA) and on LF2 (DNA methylation) of top 100 DIABLO features from each dataset. This analysis revealed good replication of protein-coding RNA and protein data; MMP7, PROM2, COL17A1, LAMC3, AGER, and ITGA3 are among the top features (**Supplemental Table S4A and B**). We observed less replication overall in the DNA methylation and lncRNA datasets but observed that MOFA prioritized DMR1/MIR21 promoter and FENDRR among the features with strongest loadings (**Supplemental Table S4C and S4D**). Taken together, it appears that a number of the same transcriptome and proteome features emerge as prioritized by both DIABLO and MOFA multi-omic methods while more work will be needed to further assess reproducibility of regulatory features of the transcriptome (DNA methylation and lncRNAs).

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

SUPPLEMENTAL TABLES (PROVIDED SEPARATELY)

Supplemental Table S1. Differential features (bacon-adjusted FDR-adjusted p<0.05) in individual datasets: **(A)** polyA RNA-seq, **(B)** proteomics (no bacon adjustment), **(C)** DNA methylation, and **(D)** rRNA-depleted noncoding RNA-seq.

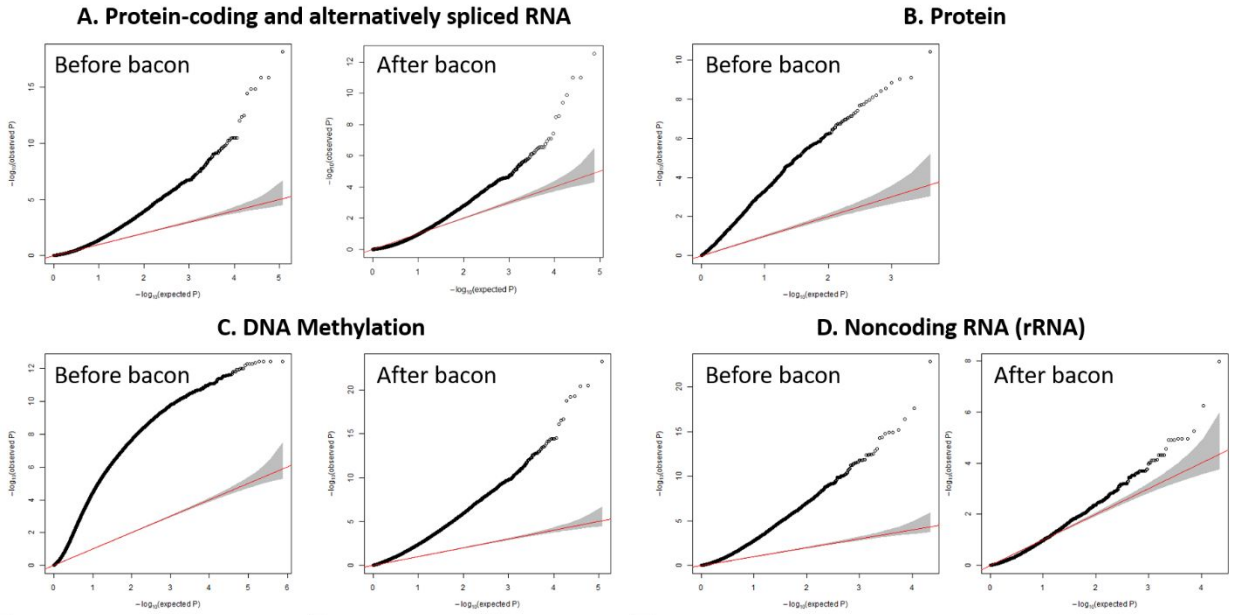
Supplemental Table S2. DIABLO model loadings on the top latent variable. **(A)** polyA RNA-seq (protein coding and retained intron), **(B)** proteomics, **(C)** DNA methylation, and **(D)** rRNA-depleted noncoding RNA-seq.

Supplemental Table S3. MOFA model loadings on the top two latent variables. **(A)** polyA RNA-seq (protein coding and retained intron), **(B)** proteomics, **(C)** DNA methylation, and **(D)** rRNA-depleted noncoding RNA-seq.

Supplemental Table S4. MOFA model loading rank for the top 20 DIABLO model features by dataset. **(A)** polyA RNA-seq (protein coding and retained intron), **(B)** proteomics, **(C)** DNA methylation, and **(D)** rRNA-depleted noncoding RNA-seq.

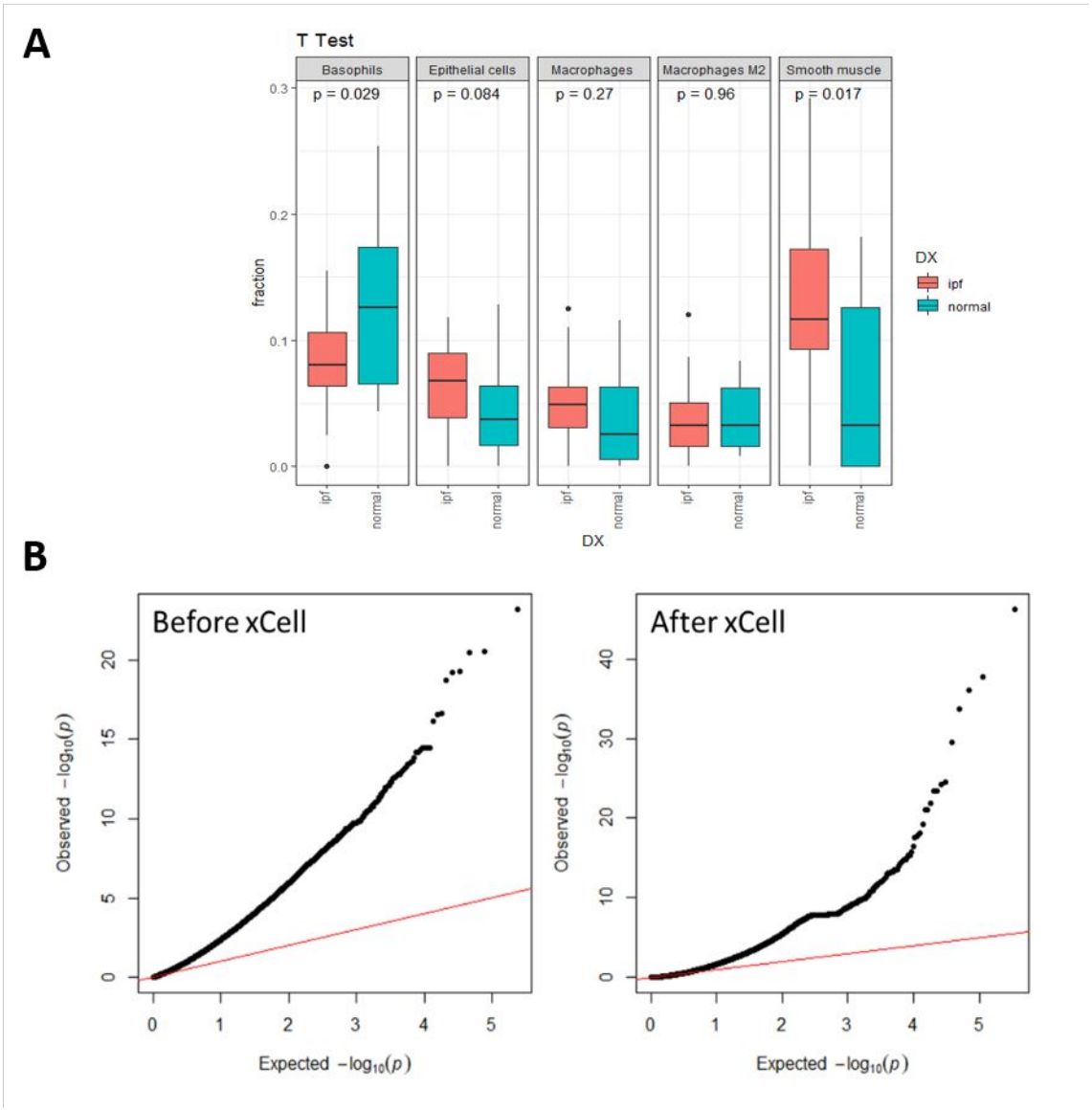
AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

SUPPLEMENTAL FIGURES



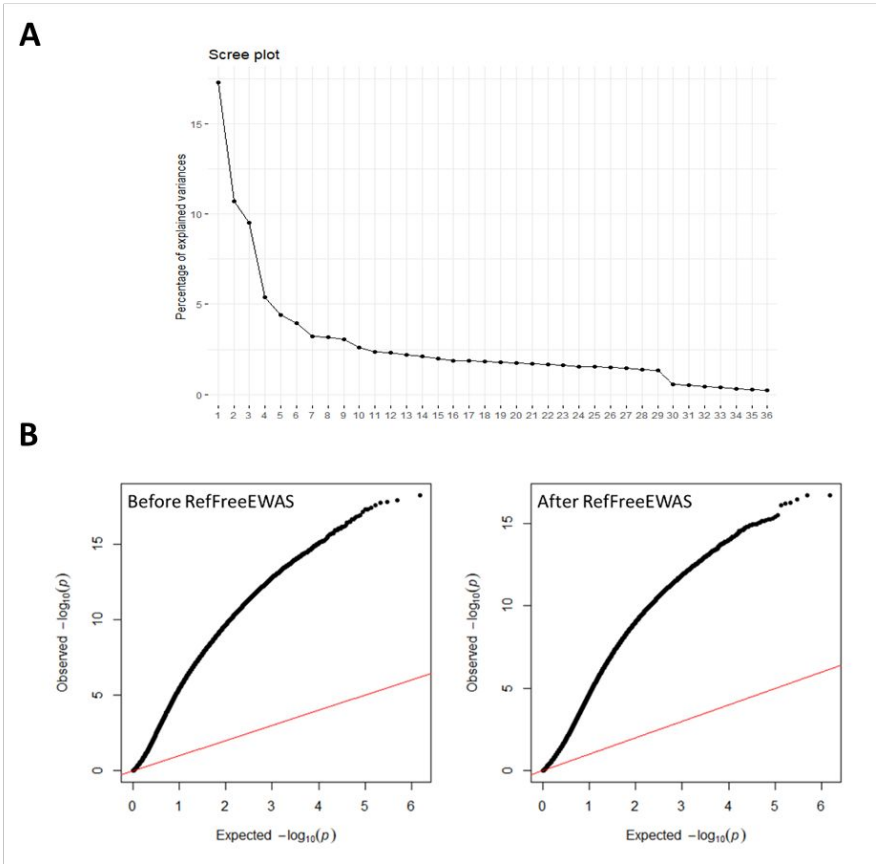
Supplemental Figure S1. Quantile-quantile (Q-Q) plots of observed compared to expected FDR-adjusted p values for each dataset. In each panel the plot on the left is pre-bacon and the plot on the right is post-bacon model fitting to adjust for bias and inflation. Protein data were not fitted to a bacon model because of an inherent bias in the proteomics assay focusing on proteins/peptides known to be involved in IPF, therefore, inflation is expected in this dataset.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC



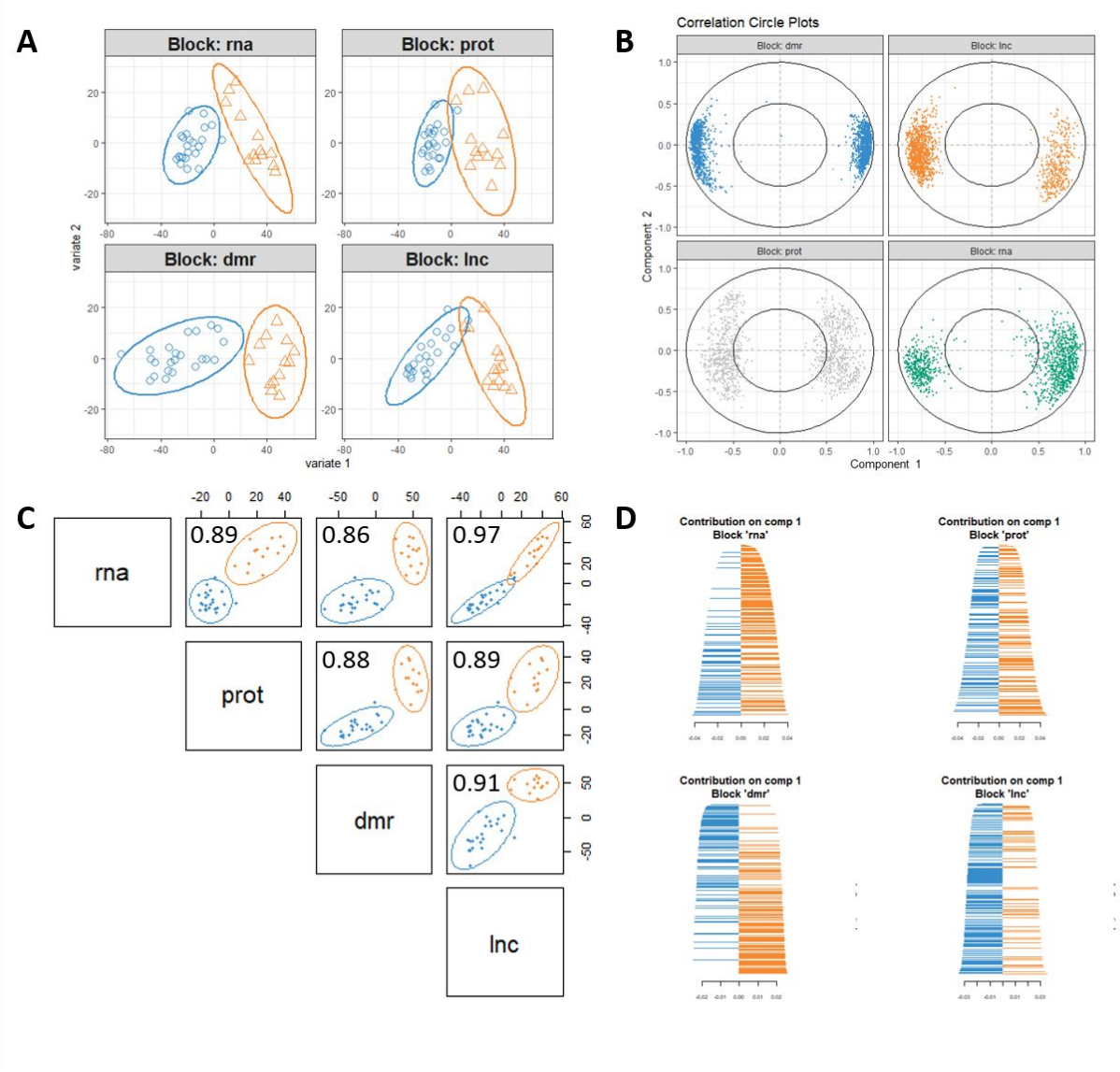
Supplemental Figure S2. xCell adjustment for cell proportions in DNA methylation analysis. **(A)** Estimated proportions of the fives cell types that are present at significant numbers in >75% of our samples **(B)** Quantile-quantile (Q-Q) plots of observed compared to expected FDR-adjusted p values before and after adjustment for the proportions of five cell types. Inclusion of the proportions of the five cell types in the model reduce inflation in the data, but to lesser extent than using Bacon, suggesting that adjustment for bias and inflation performs better (refer to Supplemental Figure S1).

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC



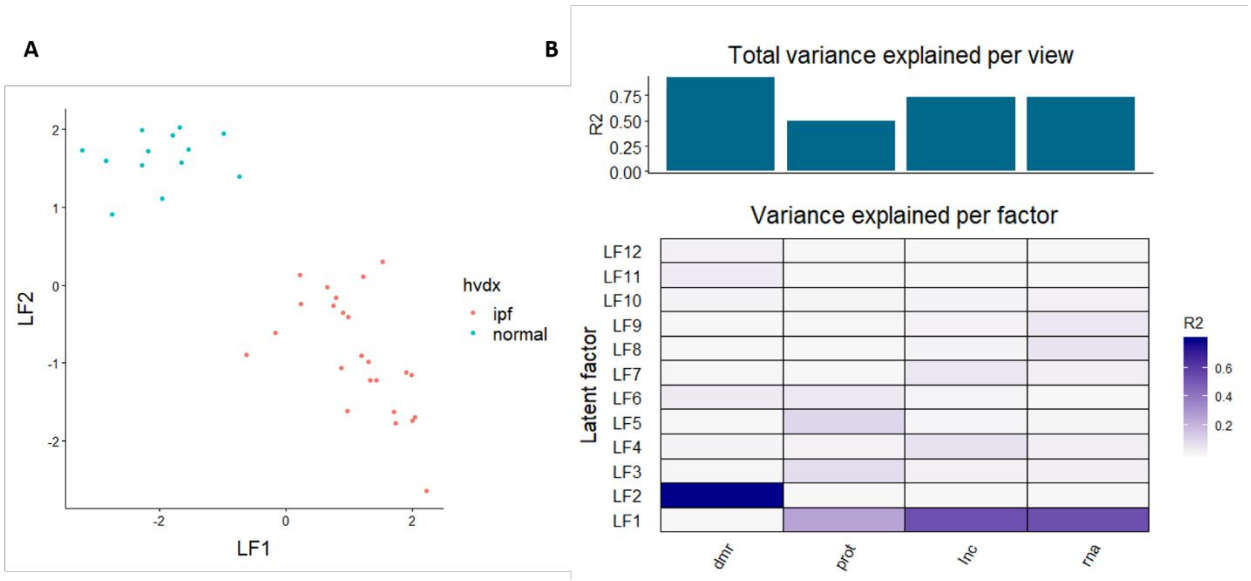
Supplemental Figure S3. RefFreeEWAS adjustment for cell proportions in DNA methylation analysis. **(A)** Scree plot depicting the amount of variance accounted for the different numbers of calculated cell components. **(B)** Quantile-quantile (Q-Q) plots of observed compared to expected FDR-adjusted p values before and after adjustment for 7 cell components identified by RefFreeEWAS. Inclusion of these seven components in the model did not reduce inflation in the data, suggesting that adjustment for bias and inflation performs better (refer to Supplemental Figure S1).

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC



Supplemental Figure S4. DIABLO multi-omic model. **(A)** Separation of cases and controls on the top two latent variables, **(B)** Correlation circle plots of individual features with the top two latent variables, **(C)** Pairwise correlations of datasets on top two latent variables, and **(D)** All features from each dataset contributing to the top latent component.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC



Supplemental Figure S5. MOFA multi-omic model. **(A)** Separation of cases and controls on the top two latent factors (LF), **(B)** Total variance explained by individual datasets (top) and variance explained by individual datasets per latent factor (bottom).

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

REFERENCES

1. Raghu G, Remy-Jardin M, Myers JL, Richeldi L, Ryerson CJ, Lederer DJ, Behr J, Cottin V, Danoff SK, Morell F, Flaherty KR, Wells A, Martinez FJ, Azuma A, Bice TJ, Bouros D, Brown KK, Collard HR, Duggal A, Galvin L, Inoue Y, Jenkins RG, Johkoh T, Kazerooni EA, Kitaichi M, Knight SL, Mansour G, Nicholson AG, Pipavath SNJ, Buendia-Roldan I, Selman M, Travis WD, Walsh S, Wilson KC, American Thoracic Society ERSJRS, Latin American Thoracic S. Diagnosis of idiopathic pulmonary fibrosis. An official ats/ers/jrs/alat clinical practice guideline. *Am J Respir Crit Care Med* 2018;198:e44-e68.

2. Grosche A, Hauser A, Lepper MF, Mayo R, von Toerne C, Merl-Pham J, Hauck SM. The proteome of native adult muller glial cells from murine retina. *Mol Cell Proteomics* 2016;15:462-480.

3. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 2012;28:882-883.

4. van Iterson M, van Zwet EW, Consortium B, Heijmans BT. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol* 2017;18:19.

5. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)* 1995;57:289-300.

6. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic rna-seq quantification. *Nat Biotechnol* 2016;34:525-527.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

7. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol* 2014;15:550.

8. Lepper MF, Ohmayer U, von Toerne C, Maison N, Ziegler AG, Hauck SM. Proteomic landscape of patient-derived cd4+ t cells in recent-onset type 1 diabetes. *J Proteome Res* 2018;17:618-634.

9. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res* 2015;43:e47.

10. Pedersen BS, Schwartz DA, Yang IV, Kechris KJ. Comb-p: Software for combining, analyzing, grouping and correcting spatially correlated p-values. *Bioinformatics* 2012;28:2986-2988.

11. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigo R, Hubbard TJ. Gencode: The reference human genome annotation for the encode project. *Genome Res* 2012;22:1760-1774.

12. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;102:15545-15550.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC

13. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, Le Cao KA. Diablo: An integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* 2019;35:3055-3062.

14. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* 2018;14:e8124.

AJRCMB Articles in Press. Published May 26, 2021 as 10.1165/rcmb.2020-0546OC