Host–Microbial Interactions in Idiopathic Pulmonary Fibrosis

Philip L. Molyneaux[1,2], Saffron A. G. Willis-Owen[1], Michael J. Cox[1], Phillip James[1], Steven Cowman[1,2], Michael Loebinger[1,2], Andrew Blanchard[3], Lindsay M. Edwards[3], Carmel Stock[1,2], Cécile Daccord[1,2], Elisabetta A. Renzoni[1,2], Athol U. Wells[2], Miriam F. Moffatt[1]*, William O. C. Cookson[1,2]*, and Toby M. Maher[1,2]*

[1]National Heart and Lung Institute, Imperial College London, London, United Kingdom; [2]Royal Brompton Hospital, London, United Kingdom; and [3]Fibrosis Discovery Performance Unit, GlaxoSmithKline R&D, GlaxoSmithKline Medicines Research Centre, Stevenage, United Kingdom

Abstract

Rationale: Changes in the respiratory microbiome are associated with disease progression in idiopathic pulmonary fibrosis (IPF). The role of the host response to the respiratory microbiome remains unknown.

Objectives: To explore the host–microbial interactions in IPF.

Methods: Sixty patients diagnosed with IPF were prospectively enrolled together with 20 matched control subjects. Subjects underwent bronchoalveolar lavage (BAL), and peripheral whole blood was collected into PAXgene tubes for all subjects at baseline. For subjects with IPF, additional samples were taken at 1, 3, and 6 months and (if alive) 1 year. Gene expression profiles were generated using Affymetrix Human Gene 1.1 ST arrays.

Measurements and Main Results: By network analysis of gene expression data, we identified two gene modules that strongly associated with a diagnosis of IPF, BAL bacterial burden (determined by 16S quantitative polymerase chain reaction), and specific microbial operational taxonomic units, as well as with lavage and peripheral blood neutrophilia. Genes within these modules that are involved in the host defense response include *NLRC4*, *PGLYRP1*, *MMP9*, and *DEFA4*. The modules also contain two genes encoding specific antimicrobial peptides (*SLPI* and *CAMP*). Many of these particular transcripts were associated with survival and showed longitudinal overexpression in subjects experiencing disease progression, further strengthening the relationship of the transcripts with disease.

Conclusions: Integrated analysis of the host transcriptome and microbial signatures demonstrated an apparent host response to the presence of an altered or more abundant microbiome. These responses remained elevated in longitudinal follow-up, suggesting that the bacterial communities of the lower airways may act as persistent stimuli for repetitive alveolar injury in IPF.

Keywords: usual interstitial pneumonia; acute lung injury; microbiome; idiopathic pulmonary fibrosis; expression

(Received in original form July 11, 2016; accepted in final form January 11, 2017)

*These three senior authors contributed equally to the study.

P.L.M. was an Asmarley Clinical Research Fellow. Funding was received from the Asmarley Charitable Trust and the Wellcome Trust. T.M.M. is supported by a National Institute for Health Research (NIHR) clinician scientist fellowship (reference CS-2013-13-017). The project was supported by the NIHR Respiratory Disease Biomedical Research Unit at the Royal Brompton and Harefield NHS Foundation Trust, the Academic Health Science Centre Biomedical Research Centre at Imperial College London, the National Institutes of Health (grants HL097163 and HL092870), and an unrestricted academic industry research grant from GlaxoSmithKline. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions: P.L.M. planned the project, designed and performed experiments, analyzed the data, and wrote the manuscript; S.A.G.W.-O., M.J.C., and P.J. performed statistical analyses, interpreted the results, and helped write the manuscript; E.A.R., C.S., A.U.W., S.C., and M.L. recruited patients and healthy control subjects; C.S. and C.D. performed genotyping; A.B. and L.M.E. helped analyze the data; T.M.M., M.F.M., and W.O.C.C. conceived of the studies of IPF, planned the project, designed experiments, analyzed the data, and wrote the manuscript; and all authors reviewed and revised the manuscript and approved it for submission.

Correspondence and requests for reprints should be addressed to Philip L. Molyneaux, M.B. B.S., Ph.D., National Heart and Lung Institute, Imperial College London, Guy Scadding Building, Royal Brompton Campus, London SW3 6LY, UK. E-mail: p.molyneaux@imperial.ac.uk

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

Am J Respir Crit Care Med Vol 195, Iss 12, pp 1640–1650, Jun 15, 2017

Originally Published in Press as DOI: 10.1164/rccm.201607-1408OC on January 13, 2017
Internet address: www.atsjournals.org

At a Glance Commentary

Scientific Knowledge on the Subject: Idiopathic pulmonary fibrosis (IPF) is a progressive and fatal disease of unknown cause. Changes in the respiratory microbiome and bacterial burden have been associated with disease progression in IPF. The role of the host response to the respiratory microbiome remains unknown.

What This Study Adds to the Field: Integrated analysis of the host transcriptome and microbial signatures demonstrates an interaction between host and environment in IPF. The response to an altered and more abundant microbiome remains during longitudinal follow-up, suggesting that the bacterial communities of the lower airways may act as persistent stimuli for repetitive alveolar injury in IPF.

Idiopathic pulmonary fibrosis (IPF) is a progressive disease of unknown etiology with a 5-year survival rate of only 20% (1). Current evidence suggests that IPF develops in genetically susceptible individuals with dysfunctional alveolar epithelial repair mechanisms after repeated episodes of alveolar injury (2). Although understanding of both the underlying genetics and potential environmental stimuli causing alveolar injury have progressed over recent years, the link between the two remains unclear (3).

Active infection in IPF is known to carry high morbidity and mortality rates (4). In individuals with IPF, immunosuppression is clearly deleterious (5), whereas treatment-adherent subjects in a large trial in which researchers assessed prophylactic co-trimoxazole in IPF experienced a reduction in overt infections and mortality (6). Results of recent transcriptomic studies have hinted at the role of disordered host defense, and thus susceptibility to infection, as an important contributor to disease progression in IPF (7–9). Indeed, polymorphisms within two genes—Toll-interacting protein (*TOLLIP*) and mucin 5B (*MUC5B*)—have both recently been associated with IPF susceptibility and linked to alterations in the lung immune response (10, 11).

To date, the most significant genetic association with IPF is with *MUC5B* polymorphisms, which have been linked to higher IPF risk but, paradoxically, slower disease progression (8). In mice, the role of *Muc5b* appears essential for normal macrophage function and effective mucociliary clearance of bacteria (12), and evidence is building to suggest a similar role in humans. Impaired mucociliary clearance would allow bacteria to persist in the lower airways, potentially acting as a trigger for alveolar injury. Indeed, the recent characterization of the respiratory microbiome in IPF has suggested that an increased bacterial burden and the presence of specific organisms could drive disease progression (13, 14). The *TOLLIP* gene encodes an adaptor protein, an important regulator of innate immune responses mediated through pattern recognition Toll-like receptors. Polymorphisms in the *TOLLIP* genes have now been linked to both IPF susceptibility and mortality (11).

Although these observations strengthen the epidemiological argument that infective environmental factors may be integral to the pathogenesis of IPF in genetically susceptible individuals, to date there has been no assessment of the host–microbial interaction. We therefore set out to explore in individuals with IPF the relationship between the peripheral whole-blood transcriptome, *MUC5B* and *TOLLIP* genotypes, and the respiratory microbiome. We used unbiased network analysis to cluster similarly expressed genes into modules, allowing us to dissect large transcriptomic datasets into easy-to-interpret functional clusters. Some of these data were previously presented in abstract form (15).

Methods

Study Design

Patients were prospectively recruited from the Interstitial Lung Disease Unit at the Royal Brompton Hospital, London, United Kingdom, between November 2010 and January 2013. Diagnoses of IPF were made according to international guidelines (16) after multidisciplinary team discussion. Healthy control subjects, including nonsmokers and smokers, were recruited using the same protocols. Subjects were excluded if they had a history of self-reported upper or lower respiratory tract infection, antibiotic use in the prior 3 months, acute IPF exacerbation, or other respiratory disorders. Written informed consent was obtained from all subjects, and the study was approved by the local research ethics committee (reference numbers 10/H0720/12 and 12/LO/1034).

After recruitment, patients were reassessed in the clinic at 1, 3, 6, and 12 months. At baseline and at each subsequent visit, peripheral blood samples were collected into PAXgene RNA tubes (PreAnalytiX, Hombrechtikon, Switzerland). Pulmonary function testing was performed at baseline and at 6 and 12 months. At baseline, subjects underwent fiberoptic bronchoscopy with bronchoalveolar lavage (BAL) as previously described (17). Genomic DNA was extracted, and the V3–V5 region of the bacterial *16S* ribosomal RNA (rRNA) gene was amplified using the 357F forward primer and the 926R reverse primer for 16S quantitative polymerase chain reaction as previously described (17).

Genotyping

Genotypes of the *MUC5B* SNP rs35705950 and *TOLLIP* SNPs rs3750920 and rs5743890 were determined using TaqMan assays (Life Technologies, Carlsbad, CA). Reactions were performed in 384-well plates, and fluorescence was read using a ViiA 7 Sequence Detection System (Applied Biosystems, Foster City, CA).

RNA Extraction, Quality Assessment, and Expression

The PAXgene Blood RNA Kit (PreAnalytiX) was used to isolate RNA according to the manufacturer's protocol. Total RNA was quantified using the NanoDrop ND 1000 UV-Vis spectrophotometer (Thermo Scientific, Wilmington, DE), and the quality and integrity were assessed using the 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) by ratio comparison of the 18S and 28S rRNA bands.

Thirty nanograms of each RNA sample was used to synthesize double-stranded complementary DNA (dscDNA) using the Ovation Pico WTA System V2 Kit (NuGEN, San Carlos, CA). Exogenous poly(A)-positive control subjects were added to monitor the efficiency of the synthesis of the dscDNA and target-labeling process. The Encore Biotin Module Kit (NuGEN) was used to fragment 2.8 μg of the purified cDNA template, which was then hybridized, washed, and scanned on the GeneTitan system (Affymetrix, Santa Clara, CA) using Human Gene 1.1 ST 16- or 24-sample array plates (Affymetrix). The complete data sets are available in the Gene Expression Omnibus database (www.ncbi.nlm.nih.gov/bioproject/PRJNA361279).

Host Transcriptome Analysis

Raw expression data were background adjusted, quantile normalized, and median polished using the robust multiarray average algorithm as implemented in the Affymetrix Power Tools software suite (version 1.12.0). To identify differentially expressed genes between each pair of sample groups, we used the linear models for microarray data (limma) package and applied a single contrast between samples at different time points and the zero time point. Significance analysis of microarrays was used to test the association between microarray gene expression and survival in patients with IPF. *P* values were adjusted for multiple testing using the Benjamini-Hochberg method for control of the expected false discovery rate (FDR). Genes with significant differential expression were then selected by use of a cutoff of an FDR-adjusted *P* value and fold change (FC) in the level of expression.

WGCNA (Weighted Gene Co-expression Network Analysis) was used to discover correlation patterns among differentially expressed genes (18). Groups of transcript modules exhibiting high topological overlap were identified. A minimum module size of 40 was specified. A representative variable (module eigengene) was calculated for each module as the first principal component (19). WGCNA was conducted using R version 3.0.2 software (Bioconductor; www.bioconductor.org).

Module eigengenes generated from WGCNA were correlated with phenotypic and microbial traits of interest. Genes in each cluster were analyzed using the Database for Annotation, Visualization and Integrated Discovery (20) with the stringency level set to medium to allow functional annotation clustering and Gene Ontology (GO) term enrichment analysis. Cytoscape 3.2.158 was used to visualize the network with a prefuse force-directed layout (21). Survival analysis was performed with a Cox proportional hazards model to assess the association between continuous explanatory variables and overall survival. The statistical significance of association of variables with a diagnosis of IPF was assessed using stepwise backward logistic regression to select the most parsimonious model from among potential covariates.

Results

Subjects and Sampling

Sixty patients with IPF and 20 control subjects (Table 1) were enrolled into the study. The subjects with IPF were predominantly male (65%), their mean age was 67.8 years, and they had moderately severe disease (diffusing capacity of the lung for carbon monoxide [$D_{L_{CO}}$], 40.9% predicted; FVC, 73.4% predicted). Twenty-four subjects with IPF died during follow-up, and a further 13 experienced a decline in FVC less than or equal to 10% and/or a decline in $D_{L_{CO}}$ less than or equal to 15% over a 6-month period (*see* Figure E1 in the online supplement). None of the subjects with IPF were receiving immunosuppressive therapy. The 20 control subjects were matched for age, sex, and smoking history, and there were no significant differences between cohorts. Half of the subjects with IPF were sampled longitudinally, with 30 subjects sampled at time point 1 (1 mo), 29 subsequently sampled at time point 2 (3 mo), 30 sampled at time point 3 (6 mo), and 21 of the 30 sampled at time point 4 (1-year sample).

Respiratory Microbiome

All of the subjects underwent BAL, and we previously reported the differences between the microbiota of subjects with IPF with those of age-, sex-, and smoking status–matched control subjects (13). We found a twofold increase in bacterial burden in the lavage of subjects with IPF and significant differences in a number of bacterial operational taxonomic units (OTUs) compared with control subjects. There were 464 bacterial OTUs identified across the IPF and control subjects. Subjects with IPF had sequence reads of four OTUs—a *Haemophilus* sp., a *Neisseria* sp., a *Streptococcus* sp., and a *Veillonella* sp.—significantly higher than in control subjects. The impact of these OTUs on the host transcriptome was therefore investigated further.

Baseline Gene Expression

At baseline, 1,358 transcript clusters were found to be differentially expressed between subjects with IPF and control subjects (1% FDR). GO analysis revealed that the top GO biological processes most enriched within the transcript clusters were related to host defense and stress (Table E1).

The five top differentially expressed genes were thioredoxin (*TXN*; FC, 2.2; $P = 2.75 \times 10^{-9}$), cystatin A (*CSTA*; FC, 2.1; $P = 5.81 \times 10^{-9}$), chemokine-like factor superfamily member 2 (*CMTM2*; FC, 1.7; $P = 1.78 \times 10^{-8}$), S100 calcium binding protein A12 (*S100A12*; FC, 1.94; $P = 8.75 \times 10^{-8}$), and retinol binding protein 7 (*RBP7*; FC, 1.86; $P = 1.45 \times 10^{-6}$) (Table E2). The largest FC observed within the differentially expressed genes was 3.62 for *ORM1*. Two other notable genes with large FCs in expression were specific antimicrobial peptides: secretory leukocyte peptidase inhibitor (*SLPI*; FC, 2.29; $P = 7.05 \times 10^{-5}$) and cathelicidin antimicrobial peptide

Table 1. Baseline Characteristics of the Subjects

	IPF (*n = 60*)	Control Subjects (*n = 20*)
Age, yr	67.8 ± 8	66.0 ± 10.0
Female sex, n (%)	39 (65)	12 (60)
Smoker (ever), n (%)	41 (68)	12 (60)
FVC, % predicted	73.4 ± 21	NR
FEV_1, % predicted	74.3 ± 19	NR
Ratio of FEV_1 to FVC	81.2 ± 7	NR
$D_{L_{CO}}$, %	40.9 ± 16	NR
O_2 saturation, %	95 ± 2	97 ± 4
6-min-walk distance, m	321 ± 134	NR

Definition of abbreviations: $D_{L_{CO}}$ = diffusing capacity of the lung for carbon monoxide; IPF = idiopathic pulmonary fibrosis; NR = not recorded.

Details are provided for subjects with IPF and healthy control subjects. Data are mean ± SD unless otherwise indicated.

(*CAMP*; FC, 2.11; $P = 3.0 \times 10^{-4}$). Up-regulation of two transcripts previously associated with IPF—*MMP9* and defensin alpha 4 (*DEFA4*)—was also seen (Table 2).

Next the 1,358 differentially expressed transcript clusters identified were included in a signed WGCNA (18). A total of five modules were identified and assigned a unique color-coded identifier: turquoise (containing 690 members), blue (289 members), brown (186 members), yellow (131 members), or green (54 members). Eight transcript clusters were unassigned.

Transcriptome and Microbial Association Analysis

To explore the specific clinical contribution of each module, correlations were sought between the module eigengenes (the first principal component of gene expression profiles for each module) (Figure E2), microbial OTUs of interest, and the phenotypic data (Figure 1). Using this approach, the brown, blue, and green modules showed the highest positive correlation with the IPF phenotype, whereas the turquoise and yellow modules were negatively correlated with the presence of disease.

Network modules often derive from specific cell types (18), providing further insights into pathogenesis. The blue and green network modules were strongly associated with peripheral blood neutrophil counts ($P < 0.0001$), whereas the turquoise module related to lymphocyte counts (Figure 1), suggesting neutrophil and lymphocyte origins, respectively, of these networks. The brown module was more weakly associated with neutrophil counts ($P = 0.05$) and marginally associated with platelet numbers ($P = 0.06$) The WGCNA userListEnrichment function (22) confirmed these findings, demonstrating the blue module to be highly enriched for known neutrophil markers ($P < 0.001$), whereas the turquoise module was enriched for markers related to lymphocytes ($P < 0.01$), and the brown module was enriched for platelet markers ($P < 0.01$).

The modules showed individual correlations to distinctive features of the disease, particularly the BAL bacterial burden (determined by 16S quantitative polymerase chain reaction), the abundance of *Neisseria and Veillonella* OTUs (both elevated in cases of IPF), and BAL and blood neutrophilia (Figure 1). There were no associations between host gene expression or *MUC5B* or *TOLLIP* genotype (Figure 1 and Figure E3).

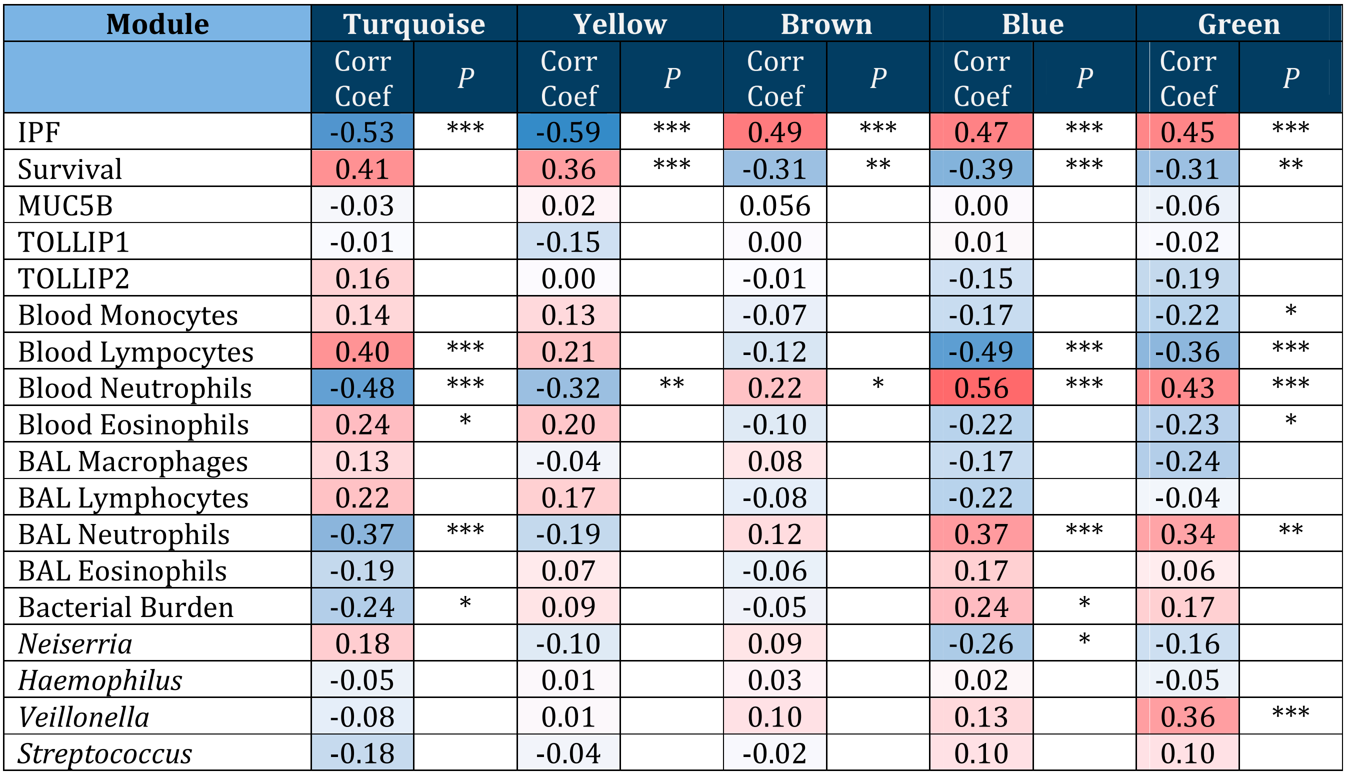
The blue and the green modules correlated with measurements of the bacterial microbiome. The blue module showed the strongest negative correlation with survival ($P = 5 \times 10^{-4}$), accompanied by a strong positive correlation with declines in FVC ($P = 0.001$), $D_{L_{CO}}$ ($P = 0.02$), and Composite Physiological Index ($P = 5 \times 10^{-4}$). The module was positively correlated with increased neutrophilia in BAL ($P = 9 \times 10^{-4}$) and blood ($P = 2 \times 10^{-7}$) as well as a higher BAL bacterial burden ($P = 0.04$). It was negatively correlated with the abundance of the *Neisseria* OTU ($P = 0.02$). There was no difference in abundance of the *Neisseria* OTU between those with stable disease and those with progressive disease. The blue module remained significantly associated with a diagnosis of IPF ($P < 0.001$) when the peripheral neutrophil counts were included as a covariate in a logistic regression, indicating that it was not acting as a simple surrogate for cell counts.

The top three GO biological processes enriched within the blue module were GO:0006952 (defense response; $P = 3.5 \times 10^{-4}$), GO:0009617 (response to bacterium; $P = 1.68 \times 10^{-6}$), and GO:0006955 (immune response; $P = 0.003$). Consistent with these GO enrichments, highly connected genes (hubs) in the blue module included *ALOX5* ($P = 2.37 \times 10^{-31}$), *NLRC4* ($P = 5.97 \times 10^{-26}$), *IL1R1* ($P = 7.03 \times 10^{-25}$), *PGLYRP1* ($P = 8.32 \times 10^{-25}$), and *TGFA* ($P = 8.40 \times 10^{-27}$) (Figure 2). The module also contained *SLPI* ($P = 1.06 \times 10^{-14}$),

Table 2. The Top 20 Transcript Clusters Significant at a 1% False Discovery Rate Ordered by Fold Change

Gene Name	Avg Expr	*t* Statistic	*P* Value	B-H–Adjusted *P* Value	Absolute Fold Change
ORM1	5.56	6.68	2.79×10^{-9}	3.61×10^{-6}	3.62
DEFA4	6.62	3.69	0.0004	0.0051	3.04
CD177	5.66	4.63	1.39×10^{-5}	0.0006	2.52
ARG1	5.09	4.02	0.0001	0.0027	2.29
SLPI	7.69	5.56	3.41×10^{-7}	7.05×10^{-5}	2.29
MMP9	7.75	4.79	7.42×10^{-6}	0.0004	2.28
RNASE3	6.49	4.17	7.61×10^{-5}	0.0019	2.26
TXN	7.80	8.80	1.96×10^{-13}	2.75×10^{-9}	2.22
BCL2A1	6.50	6.03	4.58×10^{-8}	2.01×10^{-5}	2.19
TNFAIP6	6.85	4.93	4.37×10^{-6}	0.0003	2.15
SNORD64	4.68	−4.86	5.57×10^{-6}	0.0003	−2.11
ANXA3	7.23	5.38	7.12×10^{-7}	0.0001	2.11
CAMP	7.24	4.78	7.82×10^{-6}	0.0004	2.11
CSTA	8.41	8.48	8.27×10^{-13}	5.81×10^{-9}	2.06
HP	5.51	4.11	9.29×10^{-5}	0.0021	2.05
CLEC4D	5.87	5.18	1.58×10^{-6}	0.0001	2.02
SUB1	7.32	5.37	7.23×10^{-7}	0.0001	2.01
OLFM4	4.80	2.88	0.005	0.0301	2.00
PGLYRP1	7.64	5.72	1.71×10^{-7}	4.54×10^{-5}	1.99
RPL26	8.92	5.64	2.42×10^{-7}	5.66×10^{-5}	1.97

Definition of abbreviations: Avg Expr = average $\log_2$-adjusted expression level for that gene across all the arrays; B-H = Benjamini-Hochberg.
The highest fold change in complete set of differentially expressed genes (n = 1,358) was 3.62.



Module	Turquoise		Yellow		Brown		Blue		Green	
	Corr Coef	*P*	Corr Coef	*P*	Corr Coef	*P*	Corr Coef	*P*	Corr Coef	*P*
IPF	-0.53	***	-0.59	***	0.49	***	0.47	***	0.45	***
Survival	0.41		0.36	***	-0.31	**	-0.39	***	-0.31	**
MUC5B	-0.03		0.02		0.056		0.00		-0.06	
TOLLIP1	-0.01		-0.15		0.00		0.01		-0.02	
TOLLIP2	0.16		0.00		-0.01		-0.15		-0.19	
Blood Monocytes	0.14		0.13		-0.07		-0.17		-0.22	*
Blood Lympocytes	0.40	***	0.21		-0.12		-0.49	***	-0.36	***
Blood Neutrophils	-0.48	***	-0.32	**	0.22	*	0.56	***	0.43	***
Blood Eosinophils	0.24	*	0.20		-0.10		-0.22		-0.23	*
BAL Macrophages	0.13		-0.04		0.08		-0.17		-0.24	
BAL Lymphocytes	0.22		0.17		-0.08		-0.22		-0.04	
BAL Neutrophils	-0.37	***	-0.19		0.12		0.37	***	0.34	**
BAL Eosinophils	-0.19		0.07		-0.06		0.17		0.06	
Bacterial Burden	-0.24	*	0.09		-0.05		0.24	*	0.17	
Neiserria	0.18		-0.10		0.09		-0.26	*	-0.16	
Haemophilus	-0.05		0.01		0.03		0.02		-0.05	
Veillonella	-0.08		0.01		0.10		0.13		0.36	***
Streptococcus	-0.18		-0.04		-0.02		0.10		0.10	

Figure 1. Associations between WGCNA (Weighted Gene Co-expression Network Analysis) modules and phenotypic traits. Positive correlations are shown in *red*, and negative correlations are shown in *blue*. $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$. BAL = bronchoalveolar lavage; Corr Coef = correlation coefficient; IPF = idiopathic pulmonary fibrosis; MUC5B = mucin 5B; TOLLIP1 = Toll-interacting protein 1; TOLLIP2 = Toll-interacting protein 2.

CAMP ($P = 1.65 \times 10^{-10}$), and *ORM1* ($P = 2.20 \times 10^{-16}$) (Table E4). *NLRC4*, *SLPI*, and *ORM1* all remained significantly associated with a diagnosis of IPF ($P = 0.001$, $P = 0.05$, and $P < 0.001$, respectively) when we adjusted for covariates, including bacterial burden, in a logistic regression.

The green module was significantly associated with BAL and peripheral blood neutrophilia, but, in contrast to the blue module, it showed a strong association with higher proportions of the *Veillonella* OTU within the BAL microbiota ($P = 0.001$). The top GO biological process associated with the module was response to bacterium (GO:0009617). Highly connected nodes within the green module included *ZNF267* ($P = 6.76 \times 10^{-30}$), the antigen-presenting cell surface marker *CD58* ($P = 8.11 \times 10^{-30}$), *NMI* ($P = 3.17 \times 10^{-28}$), *BCL2A1* ($P = 3.69 \times 10^{-28}$), and LY96 ($P = 5.08 \times 10^{-23}$).

The brown module demonstrated strong associations with decline in lung function, survival, and death, but it did not associate with BAL neutrophilia or any features of the microbiome (Figure E3 and Table 1). This module contained many genes from complex I of the mitochondrial respiratory chain (*NDUFA1*, *NDUFA6*, *ATP5I*, and *ATP5E*), as well as the cytochrome c oxidase subunit *COX7A2* and two genes involved in modulating nuclear factor (NF)-κB activation (*COMMD6* and *TXNDC17*) (Table E4).

In addition to its relationship with the IPF phenotype, the turquoise module was found to have a strong positive correlation ($P = 2 \times 10^{-4}$) with survival and a negative correlation with decline in FVC ($P = 3 \times 10^{-4}$), D_{LCO} ($P = 0.03$), and death ($P = 0.001$) (Table 1 and Figure E3). The module was associated with significantly lower BAL bacterial burden ($P = 0.04$) and neutrophil count in both BAL ($P = 9 \times 10^{-4}$) and blood ($P = 1 \times 10^{-5}$). This finding, given the strong positive association of this module with peripheral blood lymphocytosis ($P = 3 \times 10^{-5}$), most likely reflects relatively low numbers of peripheral lymphocytes in patients with IPF compared with the control subjects. The module remained significantly associated with a diagnosis of IPF ($P = 0.023$), however, when the peripheral cell counts were included as a covariate in a logistic regression. Significant hub genes within the turquoise module were *CMTM1* ($P = 1.24 \times 10^{-52}$), *LCK* ($P = 8.46 \times 10^{-51}$), *EVL* ($P = 1.29 \times 10^{-50}$), *RNF130* ($P = 1.34 \times 10^{-50}$), and *NMI* ($P = 3.17 \times 10^{-28}$) (Table E4). The yellow module showed similar negative associations with IPF disease and with blood neutrophilia, and it contained many genes related to RNA processing (Table E4).

Survival Analysis

Next, the expression profiles of subjects with IPF were examined for differences based upon survival to investigate for any overlap with those identified by network analysis. The significance analysis of microarrays algorithm (23) was used to generate a Cox score for each gene (Figure E4). A positive Cox score indicates that higher expression correlates with higher risk; in this case, increased expression corresponds to shorter survival, with the reverse being true for negative scores. Increased expression of five

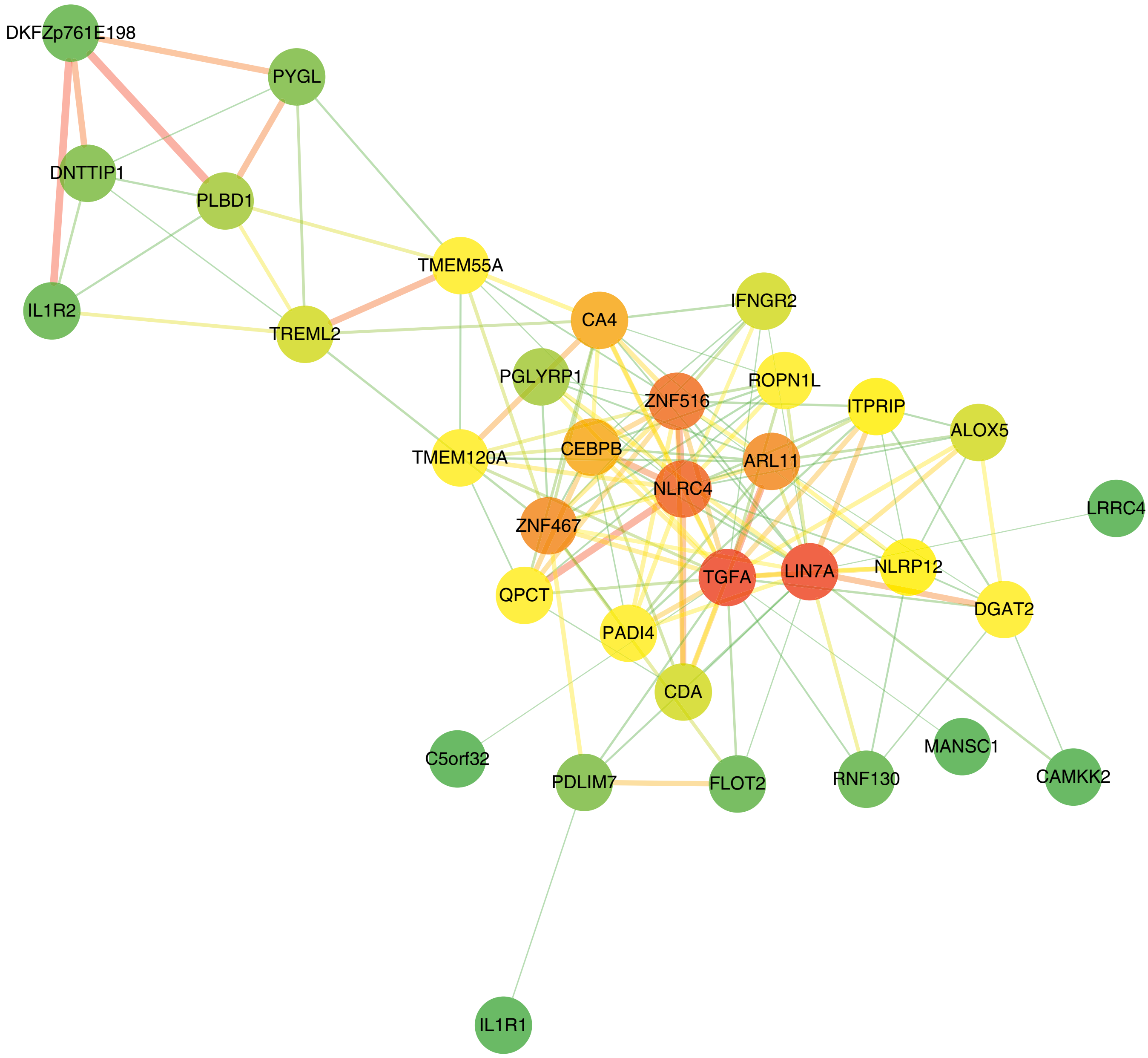


Figure 2. Visualization of the most highly connected nodes in the blue module. For improved clarity, only those nodes with the highest module memberships are considered (up to a maximum of 40), and only those connections with a topological overlap greater than 0.025 are shown. Cytoscape 3.2.158 was used to visualize the network with a prefuse force-directed layout. Genes are represented by nodes of different colors, with colors corresponding to degree (*green* to *red* indicating low to high). Edge thickness is proportional to the strength of the association.

genes (genes with a Cox score ⩾2.5) and decreased expression of 29 genes (genes with a Cox score less than or equal to −2.5) were found to correlate with survival at an FDR less than 1% (Table E3). In the IPF cohort, higher expression of *TST*, *SLPI*, *GALK1* (all members of the WGCNA blue module), *PVALB*, and *GALNT14* all correlated with decreased survival, whereas higher expression of 29 genes, including *CD247* (T-cell receptor zeta), *PRKCH*, and *SNORD78*, correlated with a longer survival (*see* online supplement).

Longitudinal Changes

Longitudinal expression profiles over 12 months were then constructed for the most significant genes and all hub genes within each module. Having identified a number of gene transcripts that associate with the IPF phenotype and correlate with bacterial burden and a neutrophilic lavage signal, we used these longitudinal constructs to establish if there was any change in the level of expression of these transcripts over time (Figure 3). No significant changes in expression values from baseline to 1 month were seen for any of the transcripts studied. The rate of gene expression change (Δ) over 12 months, however, demonstrated that expression of *MMP9*, *CAMP*, *DEFA4*, *NLRC4*, and *TXN* all increased in the IPF cohort as a whole. The most significant

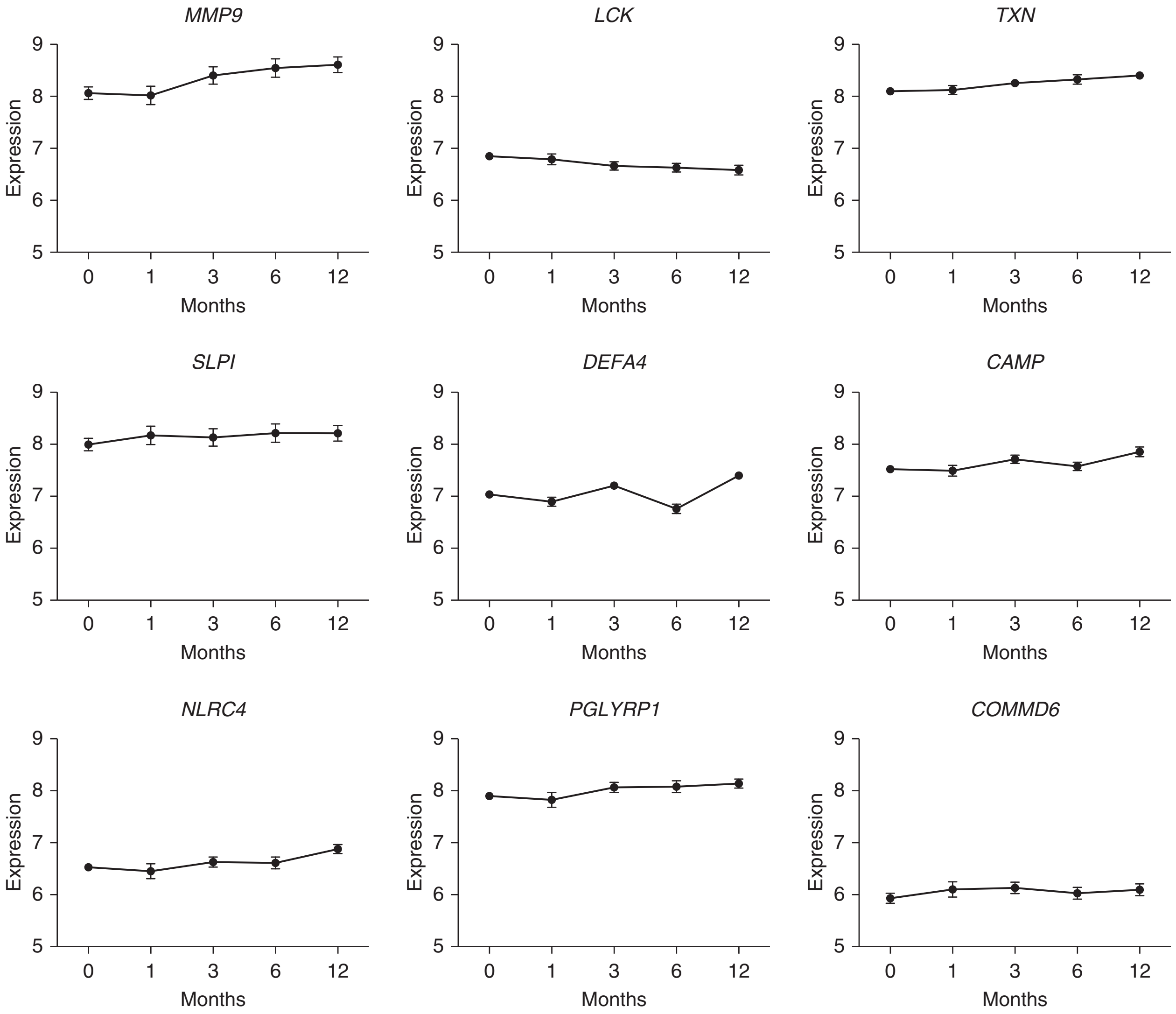


Figure 3. Expression values over time in genes of interest. The expression of *MMP9*, *CAMP*, *DEFA4*, *NLRC4*, and *TXN* increased over time. *LCK* expression decreased. There was no significant change in the expression levels of *SLPI*, *COMMD6*, and *POLYGPR* over the 12-month period.

increases were seen in *MMP9* expression levels ($\Delta = 0.65$), with substantial increases in expression also observed for *TXN* ($\Delta = 0.39$) and *DEFA4* ($\Delta = 0.36$). The expression levels of *LCK* dropped in the IPF cohort over 12 months ($\Delta = -0.31$), whereas those of *SLPI*, *COMMD6*, and *POLYGPR* remained stable.

Next, the subjects with IPF were dichotomized into progressive (defined as 6-month decline in FVC ⩾ 10% or decline in D_{LCO} ⩾ 15% or death) and stable disease cohorts, and longitudinal expression profiles were again constructed (Figure 4). The direction of gene expression change in all of the genes was the same for both stable and progressive IPF. The absolute expression levels, however, varied depending on the nature of the disease. Expression levels of *LCK* ($P = 0.016$) and *STAT4* ($P = 0.008$) were higher in stable IPF at all time points, whereas the expression levels of *MMP9* and *SLPI* were increased at all time points in progressive disease compared with stable disease ($P = 0.05$ and $P = 0.008$, respectively).

Discussion

The integration of molecular microbial, BAL, phenotypic, and transcriptomic data has highlighted interactions between the environment and host in IPF. We and others previously demonstrated that changes in the respiratory microbiome are associated with disease progression in IPF. In the present study, we have demonstrated associated changes in the peripheral blood expression profile, suggesting that there is a host response to the presence of an altered or more abundant microbiome. These responses remained elevated in longitudinal follow-up and differed between stable and progressive disease, suggesting that the bacterial communities of the lower airways may act as persistent stimuli for repetitive alveolar injury.

By network analysis, we discovered five gene modules of transcripts with coregulated expression levels. Genes within the blue and green modules were of particular interest, being involved in the host defense response. Many transcripts

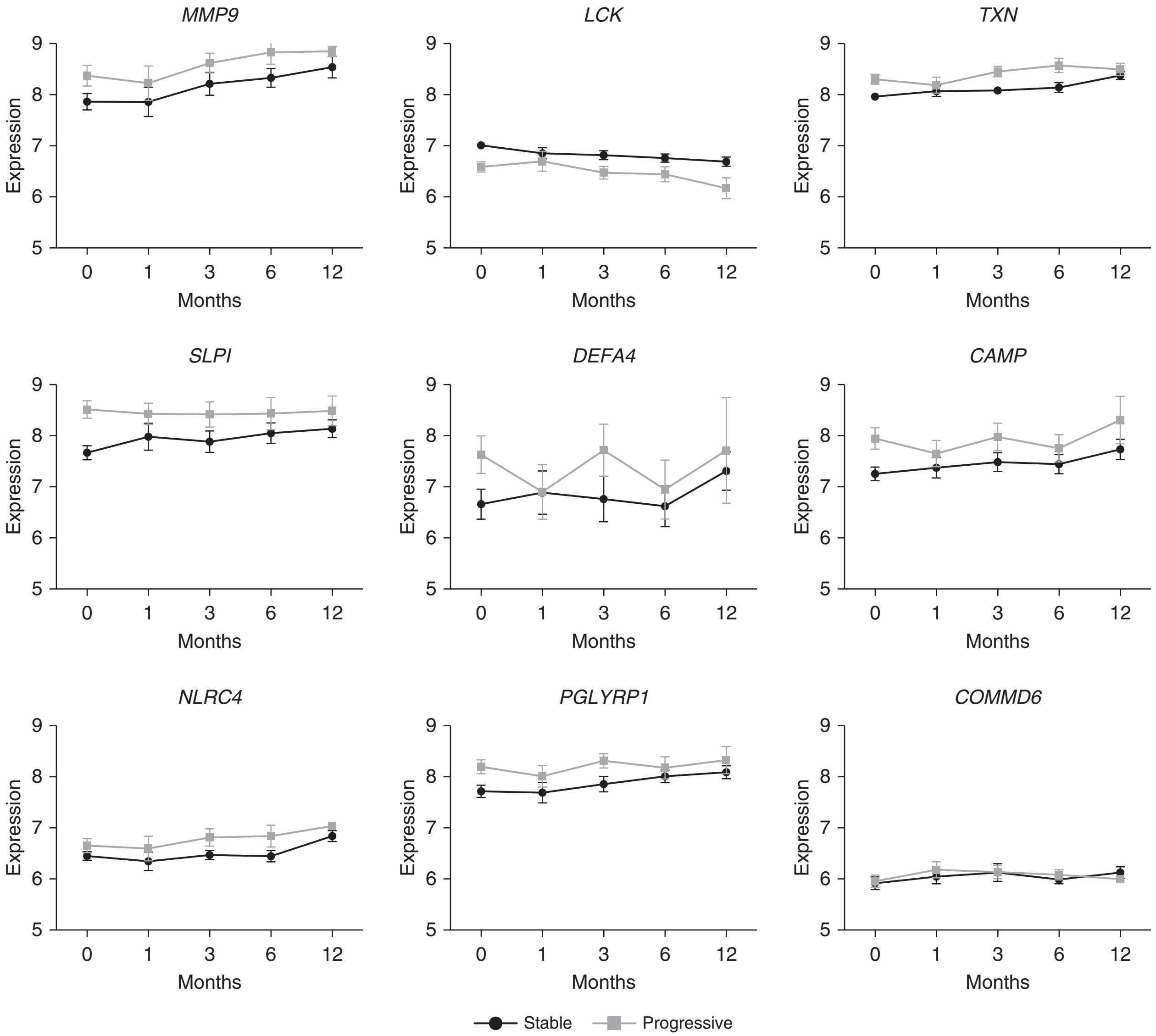


Figure 4. Expression values over time in genes of interest in stable and progressive idiopathic pulmonary fibrosis. The trend of change is the same in progressive and stable disease, but the absolute expression levels vary depending on disease state.

were associated with poor survival and remained overexpressed in subjects experiencing disease progression, strengthening their association with IPF.

The blue module showed the strongest negative correlation with survival in IPF, driven by a decline in lung function, and was positively correlated with increased BAL and blood neutrophilia as well as a higher BAL bacterial load. It was negatively correlated with the abundance of the *Neisseria* OTU, so overexpression of the genes within this module was associated with an underrepresentation of a *Neisseria* sp. within the microbiota of subjects with IPF.

One of the most highly connected hubs identified within the blue module was *NLRC4*. *NLRC4* encodes a key component of inflammasomes that plays a crucial role in the host response to proteins from pathogenic bacteria and fungi (24). *PGLYRP1*, another member of the blue module, encodes a novel antimicrobial protein with bactericidal activity against gram-positive bacteria (25). In addition, within the blue module, two genes encoding antimicrobial peptides (*SLPI* and *CAMP*) were significantly upregulated with large (2.29 and 2.11) FCs. SLPI is a serine protease inhibitor with antimicrobial properties found in mucosal fluids. SLPI has been shown to regulate intracellular enzyme synthesis, suppress matrix metalloproteinase production and activity, mediate normal wound healing, and prevent scar formation (26). Transforming growth factor-β activation by serine proteases is a possible pathogenic mechanism in IPF, and SLPI tightly regulates these protease inhibitors. In a bleomycin-induced pulmonary fibrosis model, SLPI-knockout mice demonstrate altered collagen deposition (27). In the present study, *SLPI* was overexpressed in subjects with IPF compared with control subjects, and the higher expression correlated with worse survival. At all times, the expression levels were higher in patients with progressive disease, although the expression levels of *SLPI* remained constant over the 12 months of follow-up.

CAMP protein has several functions in addition to antimicrobial activity, including cell chemotaxis, immune mediator induction, and inflammatory response regulation. *CAMP* gene expression has already been demonstrated to be up-regulated in peripheral blood in subjects with "severe" compared with "mild" IPF (28).

Within the blue module, overexpression of two genes (*MMP9* and *DEFA4*) previously widely associated with IPF was also observed (29, 30). There was a threefold increase in expression of *DEFA4* in subjects with IPF compared with healthy control subjects. Defensins are a family of microbiocidal and cytotoxic peptides involved in host defense. They are abundant in the granules of neutrophils and found in the bronchial epithelium. Subjects with IPF are known to have significantly increased concentrations of α-defensins in their plasma but not in BAL. Indeed, the plasma levels of α-defensins have been shown to correlate inversely with Pa_{O_2}, FVC, FEV_1, and $D_{L_{CO}}$ in patients with IPF (31). Plasma levels of α-defensins have also been demonstrated to be elevated in (32), and to correlate with, the clinical course of acute exacerbations in IPF (31). The role of α-defensins in fibrosis is further confirmed in patients with chronic hepatitis C, in whom the α-defensin levels and antibacterial activity correlate directly with the liver fibrosis that occurs (33).

The green module associated with the IPF phenotype but, in addition, had a strong association with higher proportions of the *Veillonella* sp. OTU, accompanied by elevated BAL and peripheral blood neutrophilia. One of the most highly connected genes (hubs) within the green module was *CD58*, which encodes a protein that plays a role in adhesion and activation of T lymphocytes during antigen presentation. The highly connected module gene *LY96* encodes a protein that associates with Toll-like receptor 4 on the cell surface and confers responsiveness to LPS, providing a potential mechanistic link between microbial exposure and the actions of this module.

The third module associated with a diagnosis of IPF, the brown module, in the present study appears to be related to cellular metabolism, in particular RNA metabolism and the citrate (trichloroacetic acid) cycle. This network shows weak associations with circulating neutrophils and possibly platelets and no association with the microbiome. It is not clear if its involvement reflects the metabolic response to a neutrophil activation and the presence serious illness or whether it captures some pathological process. Within the brown modules, there are two genes involved in modulating NF-κB activation, *COMMD6* and *TXNDC17*. NF-κB regulates a large number of genes involved in cell survival, differentiation, proliferation, apoptosis, and inflammation (34). Consequently, it is a key transcriptional regulator of the inflammatory response (35) and mediates the activity of tumor necrosis factor-α. On the basis of data generated in the bleomycin mouse model of fibrosis, it is believed that NF-κB plays a central role in the pathogenesis of lung injury and fibrosis (36).

The remaining modules (turquoise and yellow) were associated with the healthy phenotype, more stable lung function, and better survival. This improved survival was associated with significantly lower BAL bacterial burden. *LCK* is one of the most connected genes in the turquoise module and encodes a protein that is a key signaling molecule in the selection and maturation of developing T cells. Researchers in a prior study found that patients with IPF with higher levels of peripheral *LCK* had a longer survival time to transplant (37). The expression levels of *LCK* in the present study dropped in the IPF cohort as a whole over a 12-month period and were higher in patients with stable IPF at all time points, supporting the association of a lower expression level and worse survival. In this study, higher levels of *LCK* expression were associated with lower bacterial burden and improved survival. The yellow module was notably enriched for processes involved in chromosome and chromatin organization, and it was not associated with a specific cell count or enrichment for cell-specific markers. It may represent immune cellular inactivity in healthy subjects.

In this study, the 3-year follow-up and sampling allowed us to demonstrate, for the first time to our knowledge, the longitudinal expression profile of peripheral blood in subjects with IPF. We observed no significant changes in expression values from baseline to 1 month for any of the transcripts studied, though this study was not specifically powered to assess this. The rate of gene expression change over 12 months demonstrated that expression levels of the genes *MMP9*, *CAMP*, *DEFA4*, *NLRC4*, and *TXN* all increased in the IPF cohort as a whole. The expression levels of *LCK* dropped in the IPF cohort over 12 months, whereas those of *SLPI*, *COMMD6*, and *POLYGPR* remained stable. Dichotomizing the cohort into progressive and stable disease allowed comparison of longitudinal expression profiles. The direction of gene expression change in all of the genes was the same for both stable and progressive IPF. The absolute expression levels, however, varied depending on the nature of the disease. These longitudinal changes not only help support the differential baseline expression findings, providing validation of the signature, but also give confidence that these signatures could potentially be used as biomarkers at any stage of the disease process.

Despite the up-regulation of a number of potent antimicrobial factors, individuals with IPF still have an increased bacterial burden and altered microbiome compared with healthy individuals. Longitudinal microbial studies are needed to further elucidate the interaction between host and microbe. It is important to consider several other limitations of our study. Direct comparisons of longitudinal changes between subjects with IPF and healthy subjects were not feasible, owing to longitudinal healthy control samples not being available.

In the present study, we used WGCNA to identify coregulated processes that are qualitatively quite different from those derived from simple dimension reduction techniques such as principal component analysis (38). Although the results are consistent with a hypothesis that chronic infection drives IPF, it should be recognized that correlation between expression and microbial signatures does not establish causality. Importantly, gene expression changes in the peripheral blood may not directly reflect events in the lung parenchyma. Reassuringly, the blue and green neutrophil-associated modules in peripheral blood were also associated with BAL neutrophilia and with BAL bacterial burden and the *Veillonella* sp. OTU counts. Future experimental work is required to refine these findings with lung parenchymal expression profiles.

In the present study, we investigated host gene expression and known genotypes associated with IPF. We did not assess for the effect of epigenetic regulators such as DNA methylation, histone modifications, and noncoding RNAs (39), which in future studies may provide more insight into the regulation of host expression in response to environmental factors. Finally, larger studies with more ethnically diverse populations would be of great interest to establish the relevance and reproducibility of the findings in general IPF populations.

In summary, integrated analysis of the host transcriptome and microbial signatures has demonstrated, for the first time to our knowledge, an interaction between host and environment in IPF, supporting a potential pathogenic role for the lung microbiome in IPF. The bacterial communities of the lower airways may act as persistent stimuli for repetitive alveolar injury; however, the observed associations are insufficient to determine causality. Interventional studies with antibiotics or other measures able to alter the microbiome are required to further determine the clinical relevance of these findings. ■

Author disclosures are available with the text of this article at www.atsjournals.org.

References

1. Navaratnam V, Fleming KM, West J, Smith CJP, Jenkins RG, Fogarty A, Hubbard RB. The rising incidence of idiopathic pulmonary fibrosis in the U.K. *Thorax* 2011;66:462–467.
2. Maher TM, Wells AU, Laurent GJ. Idiopathic pulmonary fibrosis: multiple causes and multiple mechanisms? *Eur Respir J* 2007;30:835–839.
3. Molyneaux PL, Maher TM. The role of infection in the pathogenesis of idiopathic pulmonary fibrosis. *Eur Respir Rev* 2013;22:376–381.
4. Song JW, Hong SB, Lim CM, Koh Y, Kim DS. Acute exacerbation of idiopathic pulmonary fibrosis: incidence, risk factors and outcome. *Eur Respir J* 2011;37:356–363.
5. Idiopathic Pulmonary Fibrosis Clinical Research Network. Prednisone, azathioprine, and *N*-acetylcysteine for pulmonary fibrosis. *N Engl J Med* 2012;366:1968–1977.
6. Shulgina L, Cahn AP, Chilvers ER, Parfrey H, Clark AB, Wilson ECF, Twentyman OP, Davison AG, Curtin JJ, Crawford MB, *et al*. Treating idiopathic pulmonary fibrosis with the addition of co-trimoxazole: a randomised controlled trial. *Thorax* 2013;68:155–162.
7. Zhang Y, Noth I, Garcia JGN, Kaminski N. A variant in the promoter of *MUC5B* and idiopathic pulmonary fibrosis. *N Engl J Med* 2011;364: 1576–1577.
8. Peljto AL, Zhang Y, Fingerlin TE, Ma SF, Garcia JGN, Richards TJ, Silveira LJ, Lindell KO, Steele MP, Loyd JE, *et al*. Association between the *MUC5B* promoter polymorphism and survival in patients with idiopathic pulmonary fibrosis. *JAMA* 2013;309:2232–2239.
9. Seibold MA, Wise AL, Speer MC, Steele MP, Brown KK, Loyd JE, Fingerlin TE, Zhang W, Gudmundsson G, Groshong SD, *et al*. A common *MUC5B* promoter polymorphism and pulmonary fibrosis. *N Engl J Med* 2011;364:1503–1512.
10. Shah JA, Vary JC, Chau TTH, Bang ND, Yen NTB, Farrar JJ, Dunstan SJ, Hawn TR. Human TOLLIP regulates TLR2 and TLR4 signaling and its polymorphisms are associated with susceptibility to tuberculosis. *J Immunol* 2012;189:1737–1746.
11. Noth I, Zhang Y, Ma SF, Flores C, Barber M, Huang Y, Broderick SM, Wade MS, Hysi P, Scuirba J, *et al*. Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. *Lancet Respir Med* 2013;1:309–317.
12. Roy MG, Livraghi-Butrico A, Fletcher AA, McElwee MM, Evans SE, Boerner RM, Alexander SN, Bellinghausen LK, Song AS, Petrova YM, *et al*. Muc5b is required for airway defence. *Nature* 2014;505: 412–416.
13. Molyneaux PL, Cox MJ, Willis-Owen SAG, Mallia P, Russell KE, Russell AM, Murphy E, Johnston SL, Schwartz DA, Wells AU, *et al*. The role of bacteria in the pathogenesis and progression of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2014;190:906–913.
14. Han MK, Zhou Y, Murray S, Tayob N, Noth I, Lama VN, Moore BB, White ES, Flaherty KR, Huffnagle GB, *et al*.; COMET Investigators. Lung microbiome and disease progression in idiopathic pulmonary fibrosis: an analysis of the COMET study. *Lancet Respir Med* 2014;2: 548–556.
15. Molyneaux PL, Willis-Owen SA, Blanchard A, Lukey P, Simpson J, Marshall R, Cookson WO, Moffatt MF, Maher TM. The longitudinal peripheral whole blood transcriptome in idiopathic pulmonary fibrosis [abstract]. *Am J Respir Crit Care Med* 2015;191:A2163–.
16. Raghu G, Collard HR, Egan JJ, Martinez FJ, Behr J, Brown KK, Colby TV, Cordier JF, Flaherty KR, Lasky JA, *et al*.; ATS/ERS/JRS/ALAT Committee on Idiopathic Pulmonary Fibrosis. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *Am J Respir Crit Care Med* 2011;183:788–824.
17. Molyneaux PL, Mallia P, Cox MJ, Footitt J, Willis-Owen SAG, Homola D, Trujillo-Torralbo MB, Elkin S, Kon OM, Cookson WOC, *et al*. Outgrowth of the bacterial airway microbiome after rhinovirus exacerbation of chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2013;188:1224–1231.
18. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;9:559.
19. Langfelder P, Horvath S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol* 2007;1:54.
20. Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;4:44–57.
21. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;13:2498–2504.
22. Miller JA, Cai C, Langfelder P, Geschwind DH, Kurian SM, Salomon DR, Horvath S. Strategies for aggregating gene expression data: the collapseRows R function. *BMC Bioinformatics* 2011;12:322.
23. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 2001;98:5116–5121.
24. Zhao Y, Yang J, Shi J, Gong YN, Lu Q, Xu H, Liu L, Shao F. The NLRC4 inflammasome receptors for bacterial flagellin and type III secretion apparatus. *Nature* 2011;477:596–600.
25. Cho JH, Fraser IP, Fukase K, Kusumoto S, Fujimoto Y, Stahl GL, Ezekowitz RAB. Human peptidoglycan recognition protein S is an effector of neutrophil-mediated innate immunity. *Blood* 2005;106: 2551–2558.
26. Ashcroft GS, Lei K, Jin W, Longenecker G, Kulkarni AB, Greenwell-Wild T, Hale-Donze H, McGrady G, Song XY, Wahl SM. Secretory leukocyte protease inhibitor mediates non-redundant functions necessary for normal wound healing. *Nat Med* 2000; 6:1147–1153.
27. Habgood AN, Tatler AL, Porte J, Wahl SM, Laurent GJ, John AE, Johnson SR, Jenkins G. Secretory leukocyte protease inhibitor gene deletion alters bleomycin-induced lung injury, but not development of pulmonary fibrosis. *Lab Invest* 2016;96:623–631.
28. Yang IV, Burch LH, Steele MP, Savov JD, Hollingsworth JW, McElvania-Tekippe E, Berman KG, Speer MC, Sporn TA, Brown KK, *et al*. Gene expression profiling of familial and sporadic interstitial pneumonia. *Am J Respir Crit Care Med* 2007;175:45–54.
29. Craig VJ, Zhang L, Hagood JS, Owen CA. Matrix metalloproteinases as therapeutic targets for idiopathic pulmonary fibrosis. *Am J Respir Cell Mol Biol* 2015;53:585–600.
30. Konishi K, Gibson KF, Lindell KO, Richards TJ, Zhang Y, Dhir R, Bisceglia M, Gilbert S, Yousem SA, Song JW, *et al*. Gene expression profiles of acute exacerbations of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2009;180:167–175.

31. Mukae H, Iiboshi H, Nakazato M, Hiratsuka T, Tokojima M, Abe K, Ashitani J, Kadota J, Matsukura S, Kohno S. Raised plasma concentrations of α-defensins in patients with idiopathic pulmonary fibrosis. *Thorax* 2002;57:623–628.
32. Konishi K, Gibson KF, Lindell KO, Richards TJ, Zhang Y, Dhir R, Bisceglia M, Gilbert S, Yousem SA, Song JW, *et al*. Gene expression profiles of acute exacerbations of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2009;180:167–175.
33. Aceti A, Mangoni ML, Pasquazzi C, Fiocco D, Marangi M, Miele R, Zechini B, Borro M, Versace I, Simmaco M. α-Defensin increase in peripheral blood mononuclear cells from patients with hepatitis C virus chronic infection. *J Viral Hepat* 2006;13:821–827.
34. Ma Y, Wang M, Li N, Wu R, Wang X. Bleomycin-induced nuclear factor-κB activation in human bronchial epithelial cells involves the phosphorylation of glycogen synthase kinase 3β. *Toxicol Lett* 2009; 187:194–200.
35. Karin M, Ben-Neriah Y. Phosphorylation meets ubiquitination: the control of NF-κB activity. *Annu Rev Immunol* 2000;18:621–663.
36. Gurujeyalakshmi G, Wang Y, Giri SN. Taurine and niacin block lung injury and fibrosis by down-regulating bleomycin-induced activation of transcription nuclear factor-κB in mice. *J Pharmacol Exp Ther* 2000;293:82–90.
37. Herazo-Maya JD, Noth I, Duncan SR, Kim S, Ma SF, Tseng GC, Feingold E, Juan-Guardela BM, Richards TJ, Lussier Y, *et al*. Peripheral blood mononuclear cell gene expression profiles predict poor outcome in idiopathic pulmonary fibrosis. *Sci Transl Med* 2013; 5:205ra136.
38. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, *et al*. Variations in DNA elucidate molecular networks that cause disease. *Nature* 2008;452:429–435.
39. Yallapu MM, Jaggi M, Chauhan SC. Curcumin nanoformulations: a future nanomedicine for cancer. *Drug Discov Today* 2012;17:71–80.