A Computational Systems Approach to Elucidate New Mechanisms Involved in Progressive Lung Disease

by

Katy Norman

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biomedical Engineering)
in the University of Michigan
2020

Doctoral Committee:

Assistant Professor Kelly Arnold, Chair
Professor Jeffrey Curtis
Professor Jennifer Linderman
Professor Bethany Moore
Professor David Sept

Katy C. Norman

kcnorman@umich.edu

ORCID iD: 0000-0001-8841-0212

Acknowledgements

When I look back on myself in 2015 as I just entered the PhD program, I have to give a little laugh about all the surprises and changes that have happened along the way. I have grown more than I could ever have imagined – I have become so much more comfortable and confident in my communication skills, in my scientific skills, and in myself. These changes and this growth could not have happened without my huge support network, and I am honored to have the chance to thank everyone for what they have given me and for what they have helped me achieve.

I would first like to thank my adviser, Kelly Arnold, for giving me a chance to try out computational systems biology research when I had absolutely no background in the field coming in. I appreciate how you were easily able to create a space where I felt comfortable to learn, to ask questions, and to make mistakes as I worked towards gaining expertise in this field. Thank you for always being there for me with extra support when my nerves were getting the best of me before big presentations or before a big deadline, and for supporting my involvement in activities outside of the lab as well. Thank you so much for all of your extra support throughout the past few months as I have been writing my thesis and planning my next steps. I have been lucky to have been part of the lab at such a unique time; I cannot wait to see and hear about what directions you take the lab next.

I could not have done any of this work without help and support from my clinical collaborators: Beth Moore and Dave O'Dwyer in my IPF work, and Jeff Curtis and Christine Freeman in my COPD work. Beth and Dave, it has been so much fun to collaborate with you on the COMET analysis. I am still amazed that I was able to work with so much human data from a landmark study right when I started my PhD. Thank you for answering all my questions and always supporting me when we had ideas of new experiments or analysis that we wanted to try. I always left our meetings energized and excited to continue our projects; it has been a joy working with you. Jeff and Christine, thank you for giving me so much free reign in both the computational and experimental side of your lab. I could not have completed the PBMC stimulation experiments without guidance and help from both of you and from Valerie, and I learned so much from being able to drive the analysis of the AE-COPD data. It has been an honor to work with such a unique and complex dataset, and I enjoyed working out the details of the final story we got to tell. Beth and Jeff, thank you also for serving on my committee – your suggestions throughout this process have helped me understand the current disease research so I could effectively highlight the most interesting biological insights from our signatures.

I would also like to thank my committee members David Sept and Jennifer Linderman. I have learned so much about how to communicate scientific results effectively from our meetings throughout the years; thank you for creating a supportive environment for me to grow in this area. Dr. Sept, thank you for always having suggestions about other prior knowledge tools or data visualization techniques to use; I’ve enjoyed following up on your ideas, and they have helped strengthen my results. Dr. Linderman, thank you so much for both your scientific and career advice, which have helped me plan for the future and keep positive in the present.

Thank you also so the other members of the Arnold lab: Melissa Lemke, Kat Di Lillo, Christina Lee, Suzie Shoffner, and Emily Bozich! I have been so lucky to have worked with so many smart and dedicated women who are always working to improve our experimental and computational methods. I have learned so much from all of you, and you all push me to be my best. Melissa, I literally cannot imagine lab without you; thank you for being patient with me, for convincing me to accept helpful changes even when I complain, and for sharing chips from the cafeteria with me. Kat, I have loved working with you on the lung computational and experimental projects; thank you for all the long days (and nights) we pulled when doing Luminex on the SPIROMICS samples. I could not have done those experiments without you! I am so excited to see what you do next with SPIROMICS. Christina, you are a computational whiz and a beautiful digital artist. I'm looking forward to the time we have left together so I can learn all I can from you! Suzie, I'm so glad our time in the Arnold lab has overlapped; thank you for your expert clinical advice and for always bringing Freddie to our lab meetings. Although I'm sorry that we were just getting regular lab events going when the pandemic hit, I have loved our weekly lab meetings during quarantine. They've been a highlight of my week! I hope we all keep in touch; I can't wait to hear about all of your accomplishments as time goes on.

My PhD experience wouldn't have been the same without all of my friends that I've been able to make during my time here. I would love to thank everyone I've gotten to work with in the BME Grad Student Council – the club made me feel like I had a home here, and I have been so happy that I've been able to give back to the department through my involvement. Thank you Lauren and Melissa for being the best Co-Presidents I could have asked for, especially during big recruitment. Tiffany and Elissa, I am so happy that we've become so close; I could not imagine going through the last year without your love and support. Special thanks to Melissa for always being there for me in all aspects of my life; you always make everything better. Thanks for being the most fun office mate; I miss our chats so much!

A huge thank you to my boyfriend Rob. I still cannot believe how long we have known each other and everything that we have shared in this time. You have been a constant supply of smiles, laughter, and support throughout our PhDs, even from 2300 miles away. Thank you for always deeply listening to me, for working with me on ways to make the distance feel a little less, and for the best surprise visit earlier this year. I know I can always talk to you, and you are the first person I want to share all of my news with. Thank you for being there for me throughout this challenging and rewarding period in my life; I am beyond excited for whatever the next steps hold because I know we will be taking these steps together.

The support from my family throughout my PhD has been unbelievable. Mom and Dad, thank you for being my number one supporters. I know I can always give you a call if I need a smile, and you have done everything you can to help me stress less and smile more. Although unexpected, it was so much fun to spend a few months quarantining at home earlier this year – I loved that extra time with you guys. Thank you so much for everything you do for me; you guys are the best. Ellie, I have loved our phone calls and how I've gotten to watch you grow up so much throughout undergrad. I'm glad we've been able to be there for each other when things are rough, and I'm so excited to see all that you will accomplish in grad school in the fall! Pat-Pat and Cheryl, thank you for all of your expert help when I was applying to grad school in the first place! I wouldn't be here without your guidance, and I've appreciated all of your advice on how to navigate higher education as I've progressed in my degree. And lastly, to Nana, Kat-Kat, Hoser, and little Hazel – thank you for all that you have given and taught me; I will forever treasure the time we got to spend together.

Table of Contents

List of Tables

List of Figures

List of Appendices

Abstract

Mucosal surfaces in the lung interface with the outside environment for breathing purposes, but also provide the first line of defense against invading pathogens. The intricate balance of effective immune protection at the pulmonary epithelium without problematic inflammation is not well understood, but is an important consideration in complex lung diseases such as idiopathic pulmonary fibrosis (IPF) and chronic obstructive pulmonary disease (COPD). Although IPF is a fibrotic interstitial lung disease of unknown origin and COPD is an obstructive lung disease, they do share some similarities. Both are heterogeneous and progressive in nature, have no cure and few treatment options, advance through unknown mechanisms, and involve an aberrant immune response. As research has focused into the role the immune system plays in IPF and COPD, it has become clear that disease progression is caused by a complex dysregulation of immune factors and cells across the tissue compartments of the lungs and blood.

Data-driven modeling approaches offer the opportunity to infer protein interaction networks, which are able to identify diagnostic and prognostic biomarkers and also serve as the basis for new insight into systems-level mechanisms that define a disease state. Additionally, these approaches are able to integrate data from across multiple tissue compartments, allowing for a more holistic picture of a disease to be formed. Here, we have applied data-driven modeling approaches including partial least squares discriminant analysis, principal component analysis, decision tree analysis, and hierarchical clustering to high-throughput cell and cytokine measurements from human blood and lung samples to gain systems-level insight into IPF and COPD.

Overall we found that these approaches were useful for identifying signatures of proteins that differentiated disease state and progression better than current classifiers. We also found that integrating protein and cell measurements across tissue compartments generally improved classification and was useful for generating new mechanistic insight into progression and exacerbation events. In evaluating IPF progression, we showed that the blood proteome of progressors, but not of non-progressors, changes over time, and that our data-driven modeling techniques were able to capture these changes. Curiously, our models showed that complement system components may be associated with both COPD disease state and IPF disease progression. Lastly, though our analysis suggested that circulating blood cytokines were not useful for differentiating disease state or progression, preliminary work suggested that cell-cell communication networks arising from stimulated peripheral blood proteins may be more useful for classification and gaining mechanistic insight from minimally invasive blood samples. Overall, we believe that this approach will be useful for studying the mucosal immune response present in other diseases that are also progressive or heterogeneous in nature.

Chapter 1 Introduction

Besides being the site of respiration, the lung is also a key site of immunity because it interfaces with the outside environment. However, there is still much that is not understood about how the body maintains proper protection without experiencing an excessive immune response at this surface[1]. The importance of this problem is highlighted by the increasing number of global cases of chronic respiratory disease from 1990 to 2017 and the rise in the incidences of asthma, interstitial lung diseases, and pulmonary sarcoidosis over the past 15 years[2]. This intricate balance of effective immune protection without dysregulation is affected in lung diseases idiopathic pulmonary fibrosis (IPF) and chronic obstructive pulmonary disease (COPD). Although IPF and COPD differ in both their clinical presentation and their natural history, they are both progressive and heterogeneous diseases with few treatment options, and the mechanisms underlying development and progression are not well understood. Better understanding of the complex immunological mechanisms that are associated with disease state and disease progression will be a critical step on the path to development of better diagnostic and treatment options. This thesis aims to use systems-focused, data-driven modeling approaches to help identify signatures of key immune factors associated with IPF and COPD in order to gain increased insight into potential mechanisms associated with these two lung diseases.

1.1 Idiopathic pulmonary fibrosis disease pathogenesis and treatment

Idiopathic pulmonary fibrosis is a progressive and heterogeneous interstitial pneumonia of unknown origin with a median survival rate of 3-5 years[3,4]. The diagnosis process for IPF can be challenging, and there are few treatment options available to patients. It presents in patients as shortness of breath (dyspnea), dry cough, and fatigue[5]. IPF is more commonly diagnosed in older populations who have a history of smoking or occupation-related exposure to inhaled particles, especially in men[4], and there are also genetic variants that are associated with increased risk for the disease.

Although the exact mechanisms behind disease pathogenesis are unknown, current hypotheses involve some environmentally-caused repetitive injuries to lung alveolar epithelial cells (AECs). Additionally, there are certain genetic predisposition towards IPF as well, which includes mutations in genes coding for surfactant proteins A2 and C[6], single nucleotide polymorphisms (SNP) in the mucin 5B (MUC5B) gene (rs35705950)[7–10], and SNPs in the Toll-interacting protein (TOLLIP) gene[7]. Overall disease pathogenesis is attributed to a dysregulated healing response[11] that results in both the collapse of alveoli, which decreases the surface area available for gas exchange in the lung, and in the fibrosis of the interstitial surfaces, which can spread and cause symptoms associated with restrictive lung diseases[12]. This response is enacted in part by neutrophils, macrophages, and T cells. Neutrophils have been reported to be increased in the bronchoalveolar lavage (BAL) fluid of IPF patients compared to healthy controls[13] and classically secrete the protease neutrophil elastase (NE), pro-inflammatory cytokines, and reactive oxygen species (ROS)[14]. According to patterns seen in other chronic inflammatory diseases, macrophages may start out in the lung with a pro-inflammatory phenotype due to activation by lipopolysaccharide (LPS) or interferon γ (IFNγ), but as the disease progresses, macrophages activated by IL-13 may become more abundant and could be the cause of increases in CCL18 seen in IPF patients[15,16]. Although there is much literature that describes macrophage activation on the classically activated/M1 vs. alternatively activated/M2 axis, other studies have shown that macrophage activation is more complex than originally understood and is better described as a spectrum rather than an axis[17–19], and must be kept in mind when discussing cellular mechanisms. Th1 and Th2 CD4^{+} T cells have also been historically reported as being associated with IPF, with newer studies linking Th17, Th9, and Tregs to IPF pathogenesis as well, although the balance of these cells' action in humans is still not well understood[16]. In this environment, some AECs are reprogrammed to transition into mesenchymal cells, whereas some experience senescence or apoptosis. The transition of the epithelial cells to a more mesenchymal state and the increase in the number of IL-13-activated macrophages result in the secretion of pro-fibrotic cytokines and growth factors that attract fibroblasts to the interstitial space surrounding the alveoli. Once recruited, the fibroblasts can also differentiate into myofibroblasts if they experience an environment characterized by high mechanical stress or high concentrations of signaling molecules such as transforming growth factor β1 (TGF-β1) or specialized matrix proteins such as the fibronectin ED-A splice variant[20]. Myofibroblasts are a type of mesenchymal cell with the ability to secrete extracellular matrix (ECM) proteins like fibroblasts, and can also generate contractile forces through the production of α-smooth muscle actin (α-SMA), like smooth muscle cells[21]. Additionally, others have reported that myofibroblasts in IPF may also arise from other cell sources, such as epithelial cells that experience epithelial to mesenchymal transition (EMT)[22], lung-resident mesenchymal cells[23], or potentially from bone marrow progenitor recruitment in murine models of pulmonary fibrosis[24]. However, as El Agha et al. have summarized over multiple studies, the origin of the myofibroblast may determine if it has a pathogenic effect in the development of IPF or not[25]. Once present, both fibroblasts and myofibroblasts secrete high levels of ECM and other pro-fibrotic signaling cytokines to encourage more ECM production and fibroblast growth[26].

IPF progression, like the disease state itself, is both heterogeneous and not well understood. Patients experience progression as increased shortness of breath and cough, which is accompanied by decreased quality of life[27]. Progression is generally tracked in clinical trials through lung function measurements such as forced vital capacity (FVC) and the diffusion capacity of the lung for carbon monoxide (DLCO)[27,28]. Unfortunately, in IPF, once lung function is lost, it cannot be regained. Patients heterogeneously lose lung function over time, with some patients experiencing steep declines in lung function, others progressing slowly but steadily, and others experiencing times of relative stability interspersed with periods of steep decline[3,29]. The causes of those periods of extreme worsening may be directly due to an infection or a comorbidity, or could be caused by an acute exacerbation of IPF (AE-IPF), which could have been triggered by an external stimulus or could have an unknown cause[27,30]. AE-IPF events can present as increases in shortness of breath, cough, fever, and/or sputum production[3], and are associated with up to 46% of the deaths in IPF[30]. They are more common in patients with advanced disease, though this could be due to patients with advanced disease being more likely to seek treatment[30]. Much like the slower progressive periods in IPF, the mechanisms behind AE-IPF events remain unclear, though it is hypothesized that neutrophils or anti-inflammatory macrophages potentially activated by IL-4 or IL-13 could be involved due to their presence in the lungs of IPF patients experiencing an exacerbation[31,32].

To improve patients' quality of life, in the past decade clinicians and researchers have focused on discovering new diagnostic and prognostic markers as well as pharmacological treatment options to better patient outcomes. Since a proper diagnosis of IPF can be a challenge, streamlining this process has been the goal of many international pulmonary organizations. As reported by Raghu et al., the current guidelines for IPF diagnosis involve first ruling out any other environmental or genetic causes of the fibrosis[33]. If no other potential cause of the fibrosis can be identified, then chest high-resolution computed tomography (HRCT) scans are taken and analyzed for the presence of usual interstitial pneumonia (UIP) patterns, which includes fibrosis in a honeycomb pattern that is primarily present in the subpleural and basal regions of the lung. There are then multidisciplinary discussions with pulmonologists, radiologists, and pathologists over the patient's history and HRCT scans, especially in cases where the UIP pattern is not obvious, to gauge next steps. For patients with indeterminate UIP patterns on their HRCT scans and no history of a co-existing rheumatological disease, a surgical lung biopsy may be recommended to confirm the presence of the UIP pattern in the lung tissue itself[33] in order to completely validate an IPF diagnosis. While a biopsy is not always required for an IPF diagnosis, this procedure does present a challenge because not all patients are healthy enough to undergo a lung biopsy due to the risk of further injury that could result in a progressive event. Thus the current diagnostic guidelines could result in an unclear diagnosis in some patients[33].

Challenges associated with the diagnosis process for IPF involve the low prevalence of the disease due to high lethality rates (it is estimated to affect between 10-60 people out of 100,000[26]), the large number of other diseases that share the same presenting symptoms as IPF, and the lack of biomarkers specific for the disease. The presenting symptoms of IPF are very similar to more common lung afflictions, such as asthma or pneumonia, as well as other interstitial lung diseases (ILDs) or other immunological diseases affecting the pulmonary environment (such as hypersensitivity pneumonitis or sarcoidosis). The Interstitial Lung Disease Patient Diagnostic Journey (INTENSITY) survey reported that out of 600 ILD patients, over half (55%) received at least one misdiagnosis before receiving their current diagnosis[34]. Misdiagnosis is problematic because it prevents patients from receiving helpful treatment in a timely manner and could mean that they receive potentially harmful treatment based on their incorrect diagnosis. Thus, researchers have begun exploring potential molecular biomarkers from peripheral blood or the lungs that could help in the diagnosis of IPF and also predict the course of the disease. Some promising single biomarkers have been identified (matrix metalloproteinase 7 (MMP-7)[35,36], surfactant protein D (SP-D)[36,37], human mucin-1 (MUC1/KL-6)[36]) that can differentiate IPF from some ILDs or IPF from healthy controls, but biomarkers that are specific to only an IPF diagnosis have not yet been identified[38]. On the prognostic biomarker side, blood MMP-7[39,40], CCL18[41], KL-6[36,42], and SP-D[43,44] have shown promising results. However, it has been difficult to replicate these findings in other cohorts[38,45], especially when validating the exact concentration cut off of single biomarkers to use for diagnostic or prognostic purposes[46], and thus there are currently no biomarkers recommended for clinical use[33].

There has been great improvement in treatment options for IPF patients in the past decade, but currently there is still no cure other than lung transplantation. It was originally thought that IPF was mostly an inflammatory disease until it was reported in the PANTHER-IPF (Prednisone, Azathioprine, and N-Acetylcysteine: A Study That Evaluates Response in IPF) study that patients on an immunosuppressive, anti-inflammatory three-drug combination had increased risk of death and hospitalization as compared to the placebo group[47]. Since then, two anti-fibrotic drugs, pirfenidone[48] and nintedanib[49], were approved for the treatment of IPF. Believed to act through different mechanisms, both drugs have been shown to temporarily slow disease progression (as measured by decline in FVC), but current studies did not report significant improvements in quality of life or shortness of breath and were not powered to investigate the drugs’ effect on AE-IPF occurrence or mortality[50]. Altogether, the result is that the only current option for an IPF cure is a lung transplant. However, this procedure comes with a high risk of rejection (median survival of 4.5 years post-transplantation[51]) and is not recommended for patients over 70 years of age[52].

1.2 Chronic obstructive pulmonary disease pathogenesis

Chronic obstructive pulmonary disease (COPD) is a progressive lung disease associated with smoking that is currently the fourth leading cause of death in the United States[53]. It was estimated that at least 174 million people were living with COPD worldwide in 2015 (although underdiagnosis is common[54,55]), and that it resulted in the death of 3.2 million people that same year[56]. COPD is a costly disease, with an estimated healthcare-related spending of $36 billion in 2010 that is only projected to increase[57]. COPD is diagnosed via lung spirometry and patient history/experiences, with COPD patients meeting the following criteria: (1) a ratio of the recorded post-bronchodilator forced expiratory volume in one second (FEV_1) to FVC (FEV_1/FVC) that is less than 70%; (2) the presence of symptoms such as cough, sputum production, shortness of breath, and wheezing; and (3) significant exposure to harmful stimuli, such as cigarette or biomass smoke[58,59]. According to the Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines, disease severity can be characterized by comparing the measured FEV_1 to the FEV_1 that is predicted based on the patient's age, sex, and height. The lower the measured FEV_1 is compared to the predicted, the more severe the airflow obstruction and the higher the GOLD stage, with stages ranging from GOLD 1 (patients with $FEV_1 \geq 80\%$ predicted) to GOLD 4 (patients with $FEV_1 \leq 30\%$ predicted)[58]. Although all COPD patients share similar spirometry patterns and general symptoms, the underlying biological processes causing these test results and symptoms can vary across patients. Some patients may experience these symptoms due to emphysema-related processes, in which alveolar tissue is destroyed, leading to gas trapping and hyperinflation; whereas others may experience symptoms due to small airways disease-related processes, which involves airway remodeling and narrowing at the bronchiole level[60].

Although COPD is most commonly associated with smoking, the exact factors that lead to disease development and that are involved in disease pathogenesis are not well understood. Not all smokers develop COPD, and other risk factors for developing the disease include genetics, environmental exposure (workplace or biomass fuels exposure, for example), and poor lung growth and development[60–62]. In general, the disease is believed to develop due to repeated exposure to harmful stimuli, though most of the research into disease pathophysiology comes from studying the effects of cigarette smoke. It is hypothesized that reactive oxygen species present in cigarette smoke accumulate in the lung and lead to an increased expression of genes involved in mucus secretion, inflammation, and anti-protease inactivation[63]. COPD is characterized by increased pulmonary and systemic inflammation, especially with severe disease[60], and this inflammation involves mediators from both the adaptive and the innate immune system. On the innate side, macrophages have been found to be increased in the sputum and bronchoalveolar lavage (BAL) fluid of COPD patients[64]. It is hypothesized that due to the oxidant/anti-oxidant dysregulation, epithelial cell injury caused by cigarette smoke, and underlying genetic and epigenetic factors[65,66], macrophages increase secretion rates of pro-inflammatory cytokines and chemokines to recruit other immune cells into the lung and also display decreased phagocytic responses[66,67]. Pro-inflammatory macrophages are commonly found within the lung tissue itself, but macrophages that are anti-inflammatory in function have also been reported in the epithelial lining fluid of the alveoli[64]. More research into the spectrum of macrophage activation is needed to fully understand the significance of the location of these different macrophage populations in COPD. Neutrophils are also a major source of inflammation in COPD and are increased in the sputum and BAL of patients[66]. Neutrophil secretions add to the tissue destruction associated with emphysema by secreting neutrophil elastase and proteinases, and a higher number of the cells in the sputum is associated with increasingly severe disease[63,68]. From the adaptive immune response, $CD4^+$ and $CD8^+$ T cells are both increased in the lungs of COPD patients, with Th1 T cells especially contributing to the pro-inflammatory environment[66]. Unfortunately, cessation of smoking is only able to slow FEV_1 decline in COPD patients, but does not result in a decrease of inflammation in the lung[69,70].

COPD symptomology is also characterized by acute exacerbation (AE-COPD) events, which are deadly, with inpatient mortality rates reported to be between 3.9-7%[71–73]. They are the most costly event associated with the disease[55], with each hospital visit due to exacerbation averaging an estimated $40,000[73]. AE-COPD events are characterized by acute increases in inflammation and in symptoms severity such as cough, sputum production and shortness of breath that does not return to baseline levels without either a change in medication or a hospital stay[58]. The effects of exacerbations can be permanent, as reported in a longitudinal study where only 75% of patients returned to their baseline peak expiratory flow rates within 35 days of experiencing an AE-COPD event, and 7% of patients did not experience a full return to baseline values 90 days after exacerbation[74]. Overall, AE-COPD events can lead to a downward spiral of worsening symptoms: they have a negative impact on the patient’s quality of life, increase the rate of lung function decline, and are associated with hospital stays and death[75]. As another example of the heterogeneity of the disease, there may be a subset of “frequent exacerbator” patients who experience more exacerbations than the average COPD patient, which is defined as more than two per year[76,77]. For these frequent exacerbator patients, the negative impacts of exacerbations are even greater[78]; however, it has also been shown that frequent exacerbators may not consistently experience two exacerbations every year[79].

Current treatment options for COPD aim at managing symptoms and preventing exacerbations[80], but better definitions of COPD subgroups may hold the key to the development of more effective and personalized treatment options in the future. Common pharmacological treatments prescribed for the stable state of COPD include a combination of long-acting β_2 agonists (LABAs), long-acting muscarinic antagonists (LAMAs), and inhaled corticosteroids (ICS), depending on the patient's symptoms and exacerbation risk[80]. Pharmacologic treatments for severe exacerbations include antibiotics for bacterial infection-associated exacerbations, systemic corticosteroids, and bronchodilators[81]. Due to the heterogeneity of the disease, it can be difficult to find an effective treatment regimen for each patient. Looking forward, one of the goals for clinicians and researchers is to identify and define subpopulations of COPD patients in order to easily prescribe personalized treatment. Some researchers and clinicians focus on subpopulations that can be identified through biomarkers, as these patients may all share common mechanisms of action for disease pathogenesis and will be explored in depth below, whereas others focus on groups of patients who exhibit similar symptoms (e.g. frequent exacerbators and the GOLD ABCD classification based on symptom severity and exacerbation history[58]), as these patients may share a common phenotype[82]. Identifying these subgroups of patients is of importance so that more personalized treatments can be administered. Shifting gears to disease progression, although AE-COPD events are common and are associated with progression, the precise definition of these events is still debated by physicians, as COPD patients may experience changes in therapy that are not caused by the presence of an exacerbation[83]. Thus, a stronger definition or marker of exacerbation would aid clinicians in prescribing the correct treatments to patients.

To aid in identification of COPD subgroups and AE-COPD events, researchers have been focusing on genetic, cellular, and proteomic biomarkers that could help distinguish between patients with COPD and smokers without airway obstruction, help define patient subgroups within COPD, and help understand and predict exacerbation events. In terms of markers for COPD, alpha-1 antitrypsin (A1AT) deficiency, which is caused by a mutation in the SERPINA1 gene, is commonly reported in COPD patients (both ex-smokers and never smokers) and may also be responsible for a faster rate of emphysema development after exposure to cigarette smoke[84]. In terms of potential COPD patient subgroups, it has been reported that high levels of eosinophils are associated with a subset of COPD patients that tends to respond well to inhaled corticosteroid (ICS) in terms of FEV1 decline[85,86] and exacerbation frequency[86,87], but this has not been seen in all studies[85]. It has also been reported that some COPD patients may be characterized by an IL-17 airway epithelial response, and that these patients are less responsive to corticosteroids[88]. For AE-COPD events specifically, many studies have reported single biomarkers that are associated with the exacerbated state as opposed to the stable state (e.g. blood C-reactive protein (CRP)[89–91], sputum IL-1β[92,93], and blood growth differentiation factor 15 (GDF-15)[90,91]). However, these markers are not always unique to AE-COPD alone[89,94], and replication across multiple cohorts has been difficult[95]. Currently plasma fibrinogen is the only marker that is associated with AE-COPD, but it can only be used as an enrichment tool for clinical trials studying exacerbation[96]. A new focus in the field has involved taking a computational approach to the analysis of imaging scans to better view and predict disease progression. Computed tomography (CT) scans have recently been reported to be able to identify the presence of small airway damage in COPD with the use of parametric response mapping (PRM) analysis, which could identify unexpected damage in patients and could be used to track disease progression[97]. It has also been reported that many patients entering the hospital for AE-COPD events present with consolidation (the presence of liquid in the lung where air should be) on their chest X-rays, and that this is associated with higher mortality and may require different therapeutic steps[83]. The next steps in this area are to explore if these imaging patterns are associated with biological expression of genes, cells, or proteins that could lead to mechanistic insight into disease progression. Overall, positive steps have been made towards better identification of COPD patients, subgroups of COPD patients, and exacerbations, yet the challenge still remains to identify robust biomarkers and therapies.

1.3 Systems biology approaches to immunological disorders

Although it is clear that IPF and COPD do not act through the exact same mechanisms, they do share some similarities. As discussed above, pathogenesis of both diseases involves tissue reorganization, as seen in the aberrant collagen deposition in IPF and the airway remodeling or breakdown in COPD. Additionally, there is evidence that immune dysregulation and inflammation are involved to some extent in both, although inflammation may play a greater role over time in COPD than in IPF, especially in the disease natural history[26,47,66,68]. Lastly, our current understanding of each disease has led to similar focus areas in the related research: identification of diagnostic and prognostic biomarkers is of key importance in each disease, as is determination of the key mechanistic underpinnings of disease state and progression for the purposes of developing more targeted treatments.

Current difficulties in identifying biomarkers and treatment options suggests it may be possible that no single factor entirely accounts for disease development or progression. As IPF and COPD are complex, immunological diseases and the associated disease progression in both is multifaceted and heterogeneous, it is plausible that both the development and progression of both conditions result from disrupted systems of immune cells and cytokine communication networks rather than individual events. Approaches to infer these key players and the associated networks could provide valuable new insight into systems-level relationships driving each disease and the associated progressive events.

Systems biology-focused computational approaches, including data-driven modeling, may aid in identifying key networks of immune cells and factors involved in lung disease. These approaches add value in that they allow for evaluation of how components may interact together in a physiological system of interest, rather than as individual proteins, genes, or cells in isolated environments[98]. The increasing ease and decreasing cost of collecting quality “omics” data from biological systems has made the application of these analytical approaches more accessible over the past 20 years[99]. Data-driven modeling approaches can be applied to high-throughput data to identify small signatures of proteins, cells, or genes that covary with each other and are associated with clinically relevant groups of interest. Importantly, these approaches do not rely on prior knowledge of the system in order to identify these signatures. Additionally, unsupervised modeling approaches could be used to identify potentially novel subgroups within a patient population[100]. Through the use of knowledge-based bioinformatics databases and experimental follow-up and validation, the identified signatures can then be linked to mechanisms or cell types involved in disease phenotypes or pathogenic states. The identification of critical players in the network and the linkage to mechanisms provide starting points for potential diagnostic or prognostic criteria, insight into specific disease biology, and identification of potential targets for combinatorial therapeutic intervention[101]. In the future, these tools can also be used to help differentiate heterogeneous responses to drugs and can connect these responses to the potential underlying biology with the patient population[99].

Data-driven modeling approaches have already been applied with success to identify signatures associated with inflammation and infectious disease susceptibility in mucosal tissues of the female reproductive tract[102,103], in identifying potential sub-groups of systemic lupus erythematosus (SLE) patients[104,105], and in gaining deeper understanding into abnormal CD4$^+$ T cell and fibroblast response in rheumatoid arthritis (RA)[106,107]. These approaches have likewise been recently applied to better understand IPF and COPD disease state and disease progression. For example, in IPF, researchers have identified and validated a data-driven signature of 15 transcripts measured in lung samples that accurately distinguished healthy controls and IPF subjects[108]. In another study, a combinatorial classifier of 5 plasma proteins (MMP-7, MMP-8, MMP-1, TNRSF1A, and IGFBP1) was found via decision tree analysis to differentiate healthy controls and IPF patients with 98.6% sensitivity and 98.1% specificity[35]. This same study also reported MMP-7 and MMP-1 expression as increased in IPF compared to patients with hypersensitivity pneumonitis, but not in COPD or sarcoidosis[35]. In IPF progression, one study applied multivariate analysis to identify a signature of plasma proteins that differentiated IPF patients by progression-associated outcomes (e.g. decline in FVC and DLCO) and were associated with epithelial cell function[109]. In COPD, Christenson et al. identified a signature of transcripts associated with the response of airway epithelial cells to IL-17A exposure that was increased in a subset of COPD patients in two independent COPD studies, and corresponded with more severe airway obstruction in these patients[88]. For COPD disease progression, Bafadhel et al. used feature selection and unsupervised modeling techniques to identify signatures of proteins associated with four biologic clusters of COPD exacerbations: bacterial, eosinophil- or viral- predominant, or “pauciinflammatory”. These model-identified biological clusters were found to be very similar to previously defined clinical phenotypes of exacerbations[92].

While data-driven models have been promising in identifying potential diagnostic and prognostic markers and associated mechanisms, the limitation in current applications in IPF and COPD is that the approaches used have emphasized only the additive significance of each protein in differentiating clinical groups, rather than co-variance, which may improve classification ability and can better assist with network inference[110–112]. Additionally, there is currently a lack of studies that incorporate data from multiple tissue compartments into single models. Based on the number of previously identified markers of disease state and disease progression from the blood, it is likely that although IPF and COPD are localized in the lung, these diseases also exert measurable systemic changes. It is then plausible that systemic factors may also influence the pulmonary environment in return, and that characterizing these cross-tissue compartment proteomic and cellular networks will lead to a deeper understanding of the natural history of COPD and IPF.

1.4 Structure of thesis

With this background in mind, the goal of this study was to identify key relationships between cytokines, secreted factors, and immune cells in the blood and lungs of human patients that suggested new systems-level mechanisms of action that underpin the disease state and disease progression of IPF and COPD. We decided to focus our analysis on proteins and cells in IPF and COPD because these are biologically active factors that directly reflect the current state in patients. We also wanted to highlight relationships between proteins across tissue compartments as, to our knowledge, there is a lack of published work in this area. We accomplished our goal through the following three aims, with part A of each aim focused on IPF and part B focused on COPD:

Aim 1 will use quantitative models of high-throughput data to infer protein relationships in the blood that define patients' disease state and progression status.

Aim 2 will use computational systems analytical techniques to infer relationships in the lungs from omics samples and datasets that are associated with disease state and progression.

Aim 3 will use data-driven analytical techniques to integrate multiple types of data across various tissue compartments and assays to characterize proteomic, transcriptomic, and cellular relationships associated with disease state and progression.

Completion of these aims will be presented in the following format: **Chapter 2** presents published work that describes how these approaches can be used to identify a proteomic blood signature that differentiates healthy and IPF patients with high accuracy (Aim 1A). The related supplemental materials for this work are presented in **Appendix A**. **Chapter 3** describes published work in which these approaches were applied to identify temporal and cross-tissue compartment signatures of blood and bronchoalveolar lavage (BAL) proteins that were able to differentiate IPF progressors and non-progressors (Aim 1A and 3A). The related supplemental materials are presented in **Appendix B**. **Chapter 4** presents unpublished work of signatures of BAL proteins that differentiated healthy and IPF patients, as well as IPF progressors and non-progressors, which highlighted the importance of lung cytokines in IPF progression status (Aim 2A). **Chapter 5** includes published work to identify mechanistic hypotheses related to acute exacerbations of COPD (Aim 1B, 2B, and 3B), with the related supplemental materials presented in **Appendix C**. **Chapter 6** presents unpublished work, where plasma and BAL signatures successfully differentiated COPD disease state and severity (Aim 1B, 2B, and 3B). The supplementary materials for this work are presented in **Appendix D**. **Chapter 6** also presents preliminary results illustrating how evaluation of immune cell-cell communication networks in peripheral blood mononuclear cells (PBMCs) may be useful in evaluating lung disease (Aim 1B). **Chapter 7** contains a discussion of the key findings discovered throughout all the aims.

Chapters 2, **3**, and **5** are based off of previously published manuscripts and are presented in this thesis with minimal changes compared to their published counterparts. **Appendix E** contains details and figures on sets of models that were not included in this thesis and explanations of why we made these decisions.

Chapter 2 The Peripheral Blood Proteome Signature of Idiopathic Pulmonary Fibrosis Is Distinct From Normal and Is Associated With Novel Immunological Processes

David N. O'Dwyer[1, †], Katy C. Norman[2, †], Meng Xia[3], Yong Huang[4], Stephen J. Gurczynski[1], Shanna L. Ashley[5], Eric S. White[1], Kevin J. Flaherty[1], Fernando J. Martinez[6], Susan Murray[3], Imre Noth[4], Kelly B. Arnold[2, #], and Bethany B. Moore[1,7, #]

[†]Authors contributed equally; [#]Shared senior authorship

[1]Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA. [2]Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI, USA. [3]Biostatistics Department, University of Michigan School of Public Health, Ann Arbor, MI, USA. [4]Section of Pulmonary and Critical Care Medicine, University of Chicago, Chicago, IL, USA. [5]Immunology Graduate Program, University of Michigan, Ann Arbor, MI, USA. [6]Department of Internal Medicine, Weill Cornell Medical College, New York, NY, USA. [7]Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI, USA.

Scientific Reports. 2017; 7: 46560.

2.1 Abstract

Idiopathic pulmonary fibrosis (IPF) is a progressive and fatal interstitial pneumonia. The disease pathophysiology is poorly understood and the etiology remains unclear. Recent advances have generated new therapies and improved knowledge of the natural history of IPF. These gains have been brokered by advances in technology and improved insight into the role of various genes in mediating disease, but gene expression and protein levels do not always correlate. Thus, in this paper we apply a novel, large scale, high throughput aptamer approach to identify more than 1100 proteins in the peripheral blood of well-characterized IPF patients and normal volunteers. We use systems biology approaches to identify a unique IPF proteome signature and give insight into biological processes driving IPF. We found IPF plasma to be enriched for proteins involved in defense response, wound healing and protein phosphorylation when compared to normal human plasma. Analysis also revealed a minimal protein signature that differentiated IPF patients from normal controls, which may allow for accurate diagnosis of IPF based on easily-accessible peripheral blood. This report introduces large scale unbiased protein discovery analysis to IPF and describes distinct biological processes that further inform disease biology.

2.2 Introduction

Idiopathic Pulmonary Fibrosis (IPF) is the most common idiopathic interstitial pneumonia and is a fatal progressive disease with a median survival of 2 to 3 years[3]. The etiology of IPF remains unclear and, despite recent advances in therapy, IPF persists as an incurable disease[48,49]. IPF is characterized by certain clinical features with radiological and histopathological findings of usual interstitial pneumonia[3]. The disease results in progressive fibrotic remodeling of the pulmonary parenchyma with loss of structural integrity, impaired gas exchange, and respiratory failure. The pathophysiology of IPF features a paradigm that involves injury, loss of the epithelial cell barrier with aberrant re-epithelialization, fibroblast activation, and unregulated myofibroblast deposition of extracellular matrix components[6].

The natural history of IPF is variable and patients can experience different and dynamic clinical courses with phenotypes ranging from accelerated disease with early mortality to slowly progressive disease[113]. Considerable resources have been employed to facilitate prediction and early identification of these phenotypes to improve transplantation strategies and the selection of appropriate patients for therapeutic trials. Studies have identified proteins and chemokines that may discriminate between disease phenotypes and predict clinical outcomes[35,41,114]. Several genomic expression profiles have reported associations with disease progression in IPF[115,116] and the peripheral blood transcriptome may discriminate between mild and severe disease graded by diffusion capacity[117]. Genetic risk loci include single nucleotide polymorphisms in the Toll interacting (TOLLIP) gene, toll like receptor (TLR) 3 gene and MUC5B promoter[7,118,119]. These key advances have elucidated new potential mechanisms and therapeutic targets and have advanced the role of “omics” in IPF. However, a greater understanding of the relationship between genomic risks and the mechanistic impact on IPF pathophysiology is required. For instance, disease susceptibility is increased by the MUC5B polymorphism yet survival is improved[120]. The genome is subject to post transcriptional manipulation by micro-RNA (miRNA). Altered levels of miR-200 and miR-21 have reported associations with fibrogenesis in experimental models and human IPF patients[121,122]. Furthermore, circulating miRNA’s have been found in the blood of IPF patients and several miRNAs are differentially expressed in rapidly progressive disease[123]. Micro-RNA may act as regulators of disease progression and therefore the transcriptome and genome may be subject to significant modifications in IPF. An accurate "snapshot" of disease biology may require analysis of protein or the "proteome" in IPF patients. IPF is heterogeneous with distinct individual variation in the clinical courses that patients encounter. It is plausible that distinct and dynamic biological processes manifest as a common clinical phenotype, as evidenced by the UIP pattern on histopathology and imaging. The application of a new approach focused on identifying these processes or "molecular endotypes" may facilitate improved understanding of disease biology, molecular pathways, and the mechanisms behind the IPF clinical phenotypes[124,125].

Studies of the IPF proteome to date have focused on bronchoalveolar lavage fluid (BALF) and lung tissue analysis[126–129]. Novel targets have been reported including CCL24[126], and putative molecular pathways have been identified including the unfolded protein response through proteomic studies[127]. While BALF may be desirable for analysis given it is an accessible component of the lung environment, it is acquired through an invasive endoscopic procedure and subject to variability in representative sampling and processing. Furthermore, many patients may be unable to undergo the sampling procedure; thus, accurate analyses from peripheral blood would be optimal for patients. New proteomic assays have been developed that utilize modified aptamers termed SOMAmers© (slow off rate modified aptamers)[130]. This assay can readily analyze over 1,000 proteins at varying levels of abundance in the peripheral blood. The SOMAmer© platform has been employed in biomarker discovery in several diseases to date[131–135]. We have previously published a panel of 6 SOMAmer© measured proteins which accurately predicts disease progression in IPF[136]. In this paper, for the first time, we apply aptamer technology to identify on a large scale the differentially expressed proteins in the blood of IPF patients compared to normal controls. We then use this information to describe in detail the biological processes and molecular pathways that may discriminate the disease biology of IPF.

The ultimate goal of this work is not to identify or validate particular proteins as biomarkers, but rather to understand what biological pathways are aberrant in IPF vs. control patients based on the peripheral blood proteome.

2.3 Results

2.3.1 The peripheral proteome of IPF patients is distinct from controls

The demographics and clinical characteristics of study subjects are summarized in **Supplementary Table A.S1** in **Appendix A**. This population of IPF patients was a sub cohort of the COMET trial. The initial proteomic analysis included all 1129 available analytes which span a wide variety of biological processes and molecular pathways. Relevant comorbidities are reported in **Supplementary Table A.S2** in **Appendix A**. We applied analysis (see schematic in **Supplemental Figure A.S1**) to the blood proteins measured in the SOMAscan assay in order to find differences in the blood protein profiles of healthy and fibrotic patients. From a total of 1129 plasma proteins, 203 were found to have a mean value that was significantly different (both upregulated and downregulated) than the mean value of the same analyte in control patients, with a Bonferroni corrected α of 1% ($P < 0.0000089$) (**Fig 2.1A**). The top 10 significantly different values (all significant after Bonferroni correction with $P < 4E-19$) included glycogen synthase kinase-3 alpha/beta (GSK3A/GSK3B; 3.73 fold change), proto-oncogene tyrosine-protein kinase Src (SRC; 3.85 fold change), complement C1r subcomponent (C1R; 4.39 fold change), Proprotein convertase subtilisin/kexin type 7 (PCSK7; fold change 2.07), cGMP-specific 3',5'-cyclic phosphodiesterase (PDE5A; 4.44 fold change), sphingosine kinase 1 (SPHK1; 4.92 fold change), tyrosine-protein kinase BTK (BTK; 10.45 fold change), B-cell activating factor (BAFF; fold change 2.13), nascent polypeptide-associated complex subunit alpha (NACA; 2.28 fold change), and GTP-binding nuclear protein Ran (RAN; 10.78 fold change). Interestingly, these 10 proteins that were most significantly different between control and IPF patients were all increased in the IPF patients.

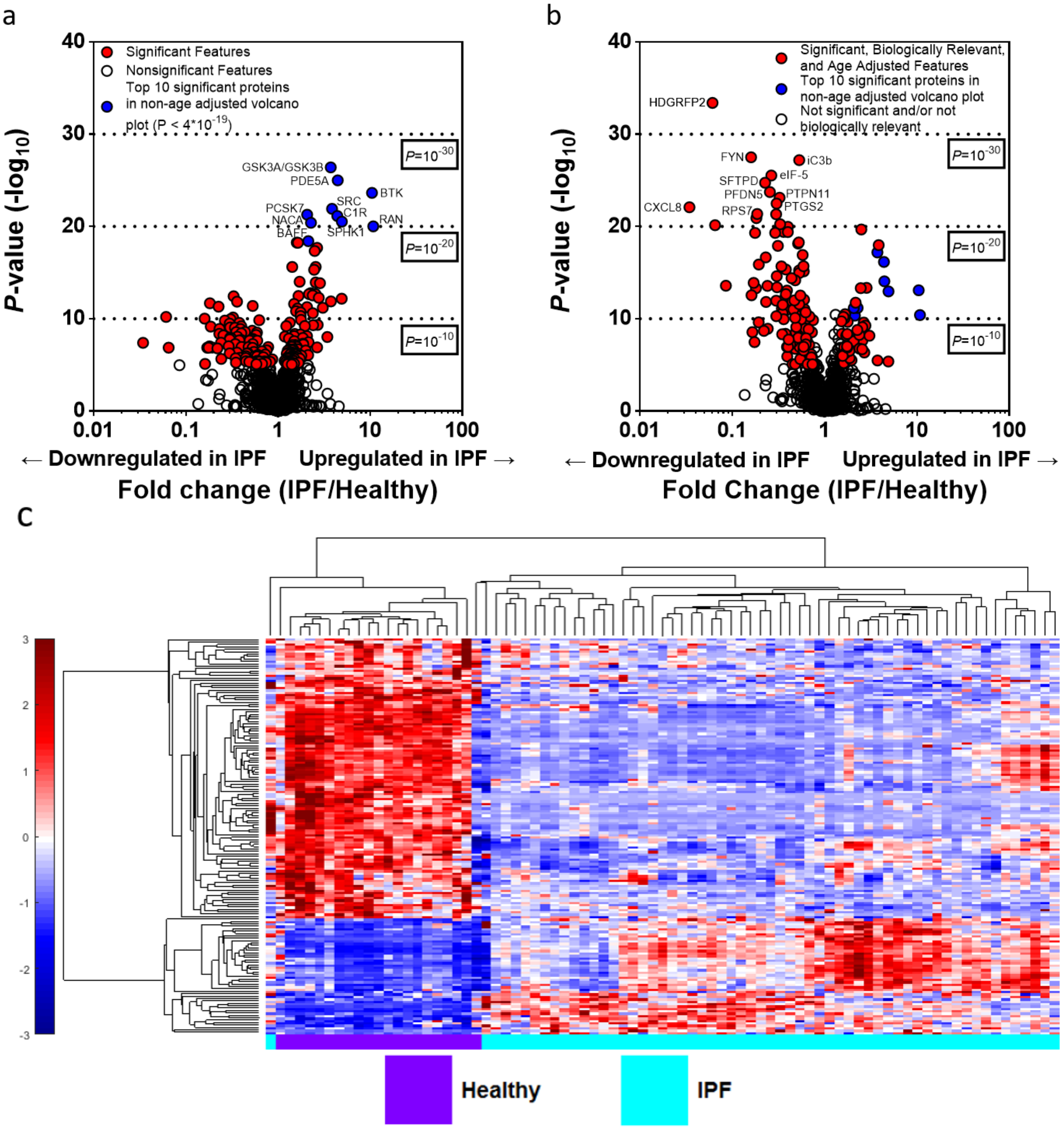


Figure 2.1. The peripheral plasma in IPF is distinct from normal controls.
(**a**) Volcano plots highlight fold change (x axis) and the significance level the y axis of the blood proteins measured by the SOMAmer Aptamer assay in the COMET study. Points in red indicate proteins that are significantly different in the healthy versus IPF patients when correcting for multiple comparisons using the Bonferroni method with a corrected P-value of 0.01. Points in blue are the top ten most significant proteins when age is not considered. (**b**) Volcano plot with age adjustment. Points in red indicate proteins that are significantly different between healthy and IPF patients when adjusted for the age difference between the two groups and when correcting for multiple comparisons using the Bonferroni method with a corrected P-value of 0.01. (**c**) Hierarchical clustering of age-adjusted blood proteins that were determined to be significantly different and biologically relevant between healthy and IFP patients show visually distinct blood proteomes between healthy and IPF patients. With the exception of two individuals, this subset of proteins in the blood was able to perfectly differentiate between healthy and IPF patients. This abundance of each protein is shown in color, with red meaning overabundant proteins, white unchanged, and blue being underabundant proteins, all compared to the mean (color bar scale is to the left of the figure). Hierarchical clustering of proteins was generated by unsupervised average linkage using Pearson’s correlation as the distance metric.

We next applied a secondary method to account for age differences between control and IPF cohorts. This screen identified 48 proteins which were expressed at significantly elevated or upregulated levels (≥ 1.5 fold) in the blood of IPF patients at screening when compared to controls (**Supplementary Table A.S3**). This represents 4.3% of total screened analytes. The screening process further identified 116 proteins which were expressed at significantly reduced or downregulated levels (≤0.75 fold) in the blood of IPF patients when compared to controls (**Supplementary Table A.S4**). This represents 10.3% of the screened analytes. A list of all significant proteins with their fold expression is reported in **Supplementary Table A.S5**. These biologically relevant, age-adjusted, significantly different proteins were then highlighted in a volcano plot (**Fig 2.1B**). The top ten significantly different, age-adjusted proteins were hepatoma-derived growth factor-related protein 2 (HDGFRP2; fold change 0.06), inactivated complement 3b (iC3b; fold change 0.53), tyrosine-protein kinase FYN (FYN; fold change 0.16), pulmonary surfactant-associated protein D (SFTPD; fold change 0.23), eukaryotic translation initiation factor 5 (EIF5; fold change 0.26), prefoldin subunit 5 (PFDN5; fold change 0.25), tyrosine-protein phosphatase non-receptor type 11 (PTPN11; fold change 0.33), prostaglandin G/H synthase 2 (PTGS2; fold change 0.30) 40S ribosomal protein S7 (RPS7; fold change 0.19), interleukin-8 (IL8; fold change 0.034). Interestingly, when the effects of age were addressed when performing the t-tests, the top ten significantly different proteins were all increased in healthy patients.

To better visualize how this age-adjusted, biologically relevant protein signature differentiated the two groups, we performed hierarchical clustering on the 48 upregulated and the 116 downregulated, age-adjusted, significantly different proteins (identified in **Fig 2.1B**) between healthy and IPF patients. The result was almost ideal differentiation of the healthy and

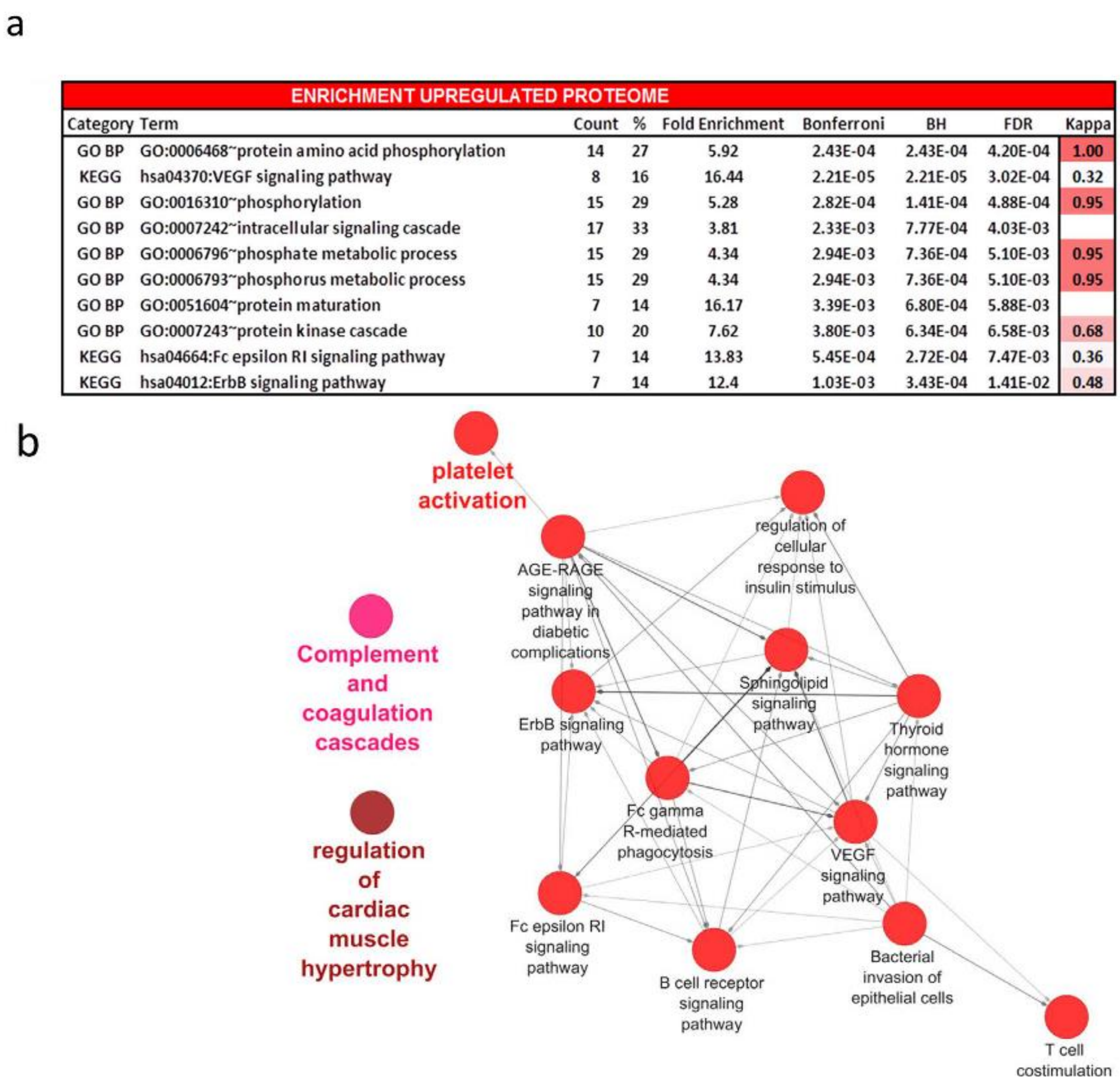
IPF groups (**Fig 2.1C**). Overall this analysis indicated visually distinct proteomes could be measured in healthy and IPF patients using a subset of 164 analytes within the SOMAscan Assay®.

The two most common co-morbidities in this patient cohort were gastroesophageal reflux disease (GERD) and obstructive sleep apnea (OSA) (**Supplementary Table A.S2**). Principal component analysis demonstrates that the greatest differences in the proteomic data arise from variation between the healthy and IPF groups, with no apparent clustering due to the co-morbidities (**Supplementary Figure A.S2**). Comorbidity information was not available for the healthy controls.

2.3.2 Enrichment and network analysis of the upregulated IPF plasma proteome

The next step was to utilize our differentially expressed proteins to gain systems level insight into the disease biology of IPF. This was achieved through enrichment analysis using the online DAVID software tool. DAVID associates proteins to hierarchically clustered functional terms (Gene Ontology, Kegg Pathway), and an enrichment score is calculated. The most significantly enriched processes included protein amino acid phosphorylation, VEGF signaling, and intracellular signaling cascade (see **Fig 2.2A**).

We next looked at possible networks and relationships between these proteins using the ClueGo application in Cytoscape. Proteins are clustered within enriched terms (Gene Ontology, Kegg Pathway) and the degree of similarity between clusters is calculated using Kappa statistics. The significantly enriched clusters included platelet activation (P = 17.0E-12), the regulation of cardiac muscle hypertrophy (P = 2.9E-6) and complement and coagulation cascades (P = 53.0E-6) (**Fig 2.2B**). The level of agreement between each cluster and term is reported by Kappa statistics (supplemental **Fig A.S3**). Statistical values for each reported term are listed in



a

ENRICHMENT UPREGULATED PROTEOME								
Category	Term	Count	%	Fold Enrichment	Bonferroni	BH	FDR	Kappa
GO BP	GO:0006468~protein amino acid phosphorylation	14	27	5.92	2.43E-04	2.43E-04	4.20E-04	1.00
KEGG	hsa04370:VEGF signaling pathway	8	16	16.44	2.21E-05	2.21E-05	3.02E-04	0.32
GO BP	GO:0016310~phosphorylation	15	29	5.28	2.82E-04	1.41E-04	4.88E-04	0.95
GO BP	GO:0007242~intracellular signaling cascade	17	33	3.81	2.33E-03	7.77E-04	4.03E-03	
GO BP	GO:0006796~phosphate metabolic process	15	29	4.34	2.94E-03	7.36E-04	5.10E-03	0.95
GO BP	GO:0006793~phosphorus metabolic process	15	29	4.34	2.94E-03	7.36E-04	5.10E-03	0.95
GO BP	GO:0051604~protein maturation	7	14	16.17	3.39E-03	6.80E-04	5.88E-03	
GO BP	GO:0007243~protein kinase cascade	10	20	7.62	3.80E-03	6.34E-04	6.58E-03	0.68
KEGG	hsa04664:Fc epsilon RI signaling pathway	7	14	13.83	5.45E-04	2.72E-04	7.47E-03	0.36
KEGG	hsa04012:ErbB signaling pathway	7	14	12.4	1.03E-03	3.43E-04	1.41E-02	0.48

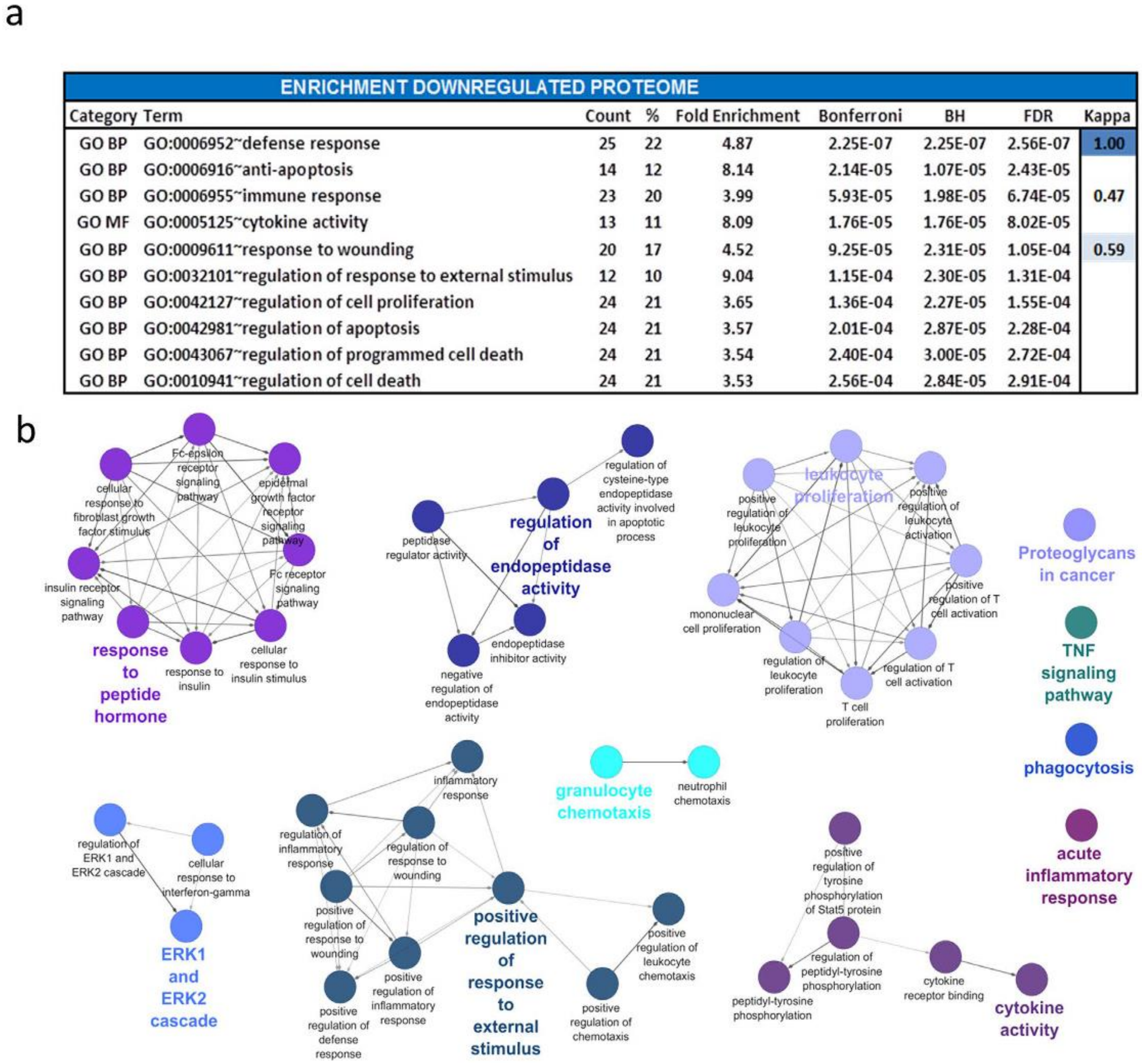
b
platelet activation
Complement and coagulation cascades
regulation of cardiac muscle hypertrophy
regulation of cellular response to insulin stimulus
AGE-RAGE signaling pathway in diabetic complications
Sphingolipid signaling pathway
ErbB signaling pathway
Thyroid hormone signaling pathway
Fc gamma R-mediated phagocytosis
VEGF signaling pathway
Fc epsilon RI signaling pathway
B cell receptor signaling pathway
Bacterial invasion of epithelial cells
T cell costimulation

Figure 2.2. Enrichment and network analysis of the upregulated IPF plasma proteome.
(**a**) DAVID enrichment analysis was employed to select the most significantly enriched terms within the sample of upregulated proteins (n = 48). Bonferroni corrected P value, Benjamini-Hochberg (BH) P value and False Discovery Rates (FDR) are reported. Kappa statistics reporting similarity to most significant term (low > 0.25, moderate 0.25-0.5, high 0.5-0.75, very high 0.75-1). (**b**) ClueGO visualization and analysis of biological role (GO, Kegg pathways) was undertaken. GO terms are mapped in clusters by Kappa statistics [Hexagon=Kegg pathway, Ellipse=Gene ontology term, arrow depicts direction of association]. The major overview term (smallest P value within the cluster) is depicted in color. Node size depicts Bonferroni corrected P value < 0.0005 for all terms reported. Further details can be found in online supplement/**Appendix A**.

Supplemental Table S6. In order to biologically validate our proteomic pathway discovery findings, we analyzed threshold values of transcriptomic data from peripheral blood cells in the same patients and report that VEGF-related genes correlate with VEGF-related proteins as measured by aptamers (data not shown). These differentially expressed VEGF-related genes when analyzed by Kegg pathway are enriched in biological pathways that are plausibly related to VEGF signaling, providing biological validation for our findings.

2.3.3 Enrichment and network analysis of the downregulated IPF plasma proteome

The downregulated proteins were analyzed for enrichment using the DAVID online software tool. The most significantly enriched terms (GO ontology, Kegg pathway) included



ENRICHMENT DOWNREGULATED PROTEOME

Category	Term	Count	%	Fold Enrichment	Bonferroni	BH	FDR	Kappa
GO BP	GO:0006952~defense response	25	22	4.87	2.25E-07	2.25E-07	2.56E-07	1.00
GO BP	GO:0006916~anti-apoptosis	14	12	8.14	2.14E-05	1.07E-05	2.43E-05	
GO BP	GO:0006955~immune response	23	20	3.99	5.93E-05	1.98E-05	6.74E-05	0.47
GO MF	GO:0005125~cytokine activity	13	11	8.09	1.76E-05	1.76E-05	8.02E-05	
GO BP	GO:0009611~response to wounding	20	17	4.52	9.25E-05	2.31E-05	1.05E-04	0.59
GO BP	GO:0032101~regulation of response to external stimulus	12	10	9.04	1.15E-04	2.30E-05	1.31E-04	
GO BP	GO:0042127~regulation of cell proliferation	24	21	3.65	1.36E-04	2.27E-05	1.55E-04	
GO BP	GO:0042981~regulation of apoptosis	24	21	3.57	2.01E-04	2.87E-05	2.28E-04	
GO BP	GO:0043067~regulation of programmed cell death	24	21	3.54	2.40E-04	3.00E-05	2.72E-04	
GO BP	GO:0010941~regulation of cell death	24	21	3.53	2.56E-04	2.84E-05	2.91E-04	

b
Fc-epsilon receptor signaling pathway
cellular response to fibroblast growth factor stimulus
epidermal growth factor receptor signaling pathway
insulin receptor signaling pathway
Fc receptor signaling pathway
response to peptide hormone
response to insulin
cellular response to insulin stimulus
regulation of cysteine-type endopeptidase activity involved in apoptotic process
peptidase regulator activity
regulation of endopeptidase activity
endopeptidase inhibitor activity
negative regulation of endopeptidase activity
leukocyte proliferation
positive regulation of leukocyte proliferation
positive regulation of leukocyte activation
mononuclear cell proliferation
positive regulation of T cell activation
regulation of leukocyte proliferation
regulation of T cell activation
T cell proliferation
Proteoglycans in cancer
TNF signaling pathway
phagocytosis
inflammatory response
granulocyte chemotaxis
neutrophil chemotaxis
regulation of ERK1 and ERK2 cascade
cellular response to interferon-gamma
ERK1 and ERK2 cascade
regulation of inflammatory response
regulation of response to wounding
positive regulation of response to wounding
positive regulation of inflammatory response
positive regulation of defense response
positive regulation of response to external stimulus
positive regulation of leukocyte chemotaxis
positive regulation of chemotaxis
positive regulation of tyrosine phosphorylation of Stat5 protein
regulation of peptidyl-tyrosine phosphorylation
peptidyl-tyrosine phosphorylation
cytokine receptor binding
cytokine activity
acute inflammatory response

Figure 2.3. Enrichment and network analysis for the downregulated IPF plasma proteome.
(**a**) DAVID enrichment analysis was employed to select the most significantly enriched terms within the sample of downregulated proteins (n=116). Bonferroni corrected P value, BH P value and FDRs are reported. Kappa statistics reported similarity to most significant term (low > 0.25, moderate 0.25-0.5, high 0.5-0.75, very high 0.75-1). (**b**) ClueGO visualization and analysis of biological role (GO, Kegg pathways) was undertaken. GO terms are mapped in clusters by Kappa statistics [Hexagon = Kegg pathway, Ellipse = Gene ontology term, arrow depicts direction of association]. The major overview term (smallest P value within cluster) is depicted in color. Node size depicts Bonferroni corrected P value < 0.0005 for all terms reported. Further details can be found in **Appendix A**.

defense response, anti-apoptosis and immune response (see **Fig 2.3A**). Cytoscape and ClueGo were then utilized to examine possible networks and relationships between enriched terms and their associated proteins. These significant clusters included acute inflammatory response (P = 740.0E-9), response to peptide hormone (P = 3.4E-15), phagocytosis (P = 1.8E-6), regulation of endopeptidase activity (P = 14.0E-12), leukocyte proliferation (P = 25.0E-9), ERK1/2 cascades (P = 150.0E-12), granulocyte chemotaxis (P = 22.0E-9), positive regulation of a response to an external stimulus (P = 74.0E-24), TNF signaling pathway (P = 4.2E-6), proteoglycans in cancer (P = 530.0E-9), and cytokine activity (P = 140.0E-15) (**Fig 2.3B**). Kappa statistics for similarity between gene, terms and clusters can be found in **Supplement Figure A.S4**. Statistical values for each reported term are listed in **Supplemental Table A.S7**.

2.3.4 A unique protein signature involved in immune processes differentiates IPF patients from controls

We next wanted to find a minimum set of proteins that best differentiated the healthy and IPF patients based on covariance, or relationships between proteins. This signature could potentially be used as a diagnostic tool based on non-invasive measurements made from peripheral blood. To identify the minimum multivariate protein signature that differentiated healthy and IPF patients, we used the Least Absolute Shrinkage and Selection Operator (LASSO) method as a feature selection tool, followed by Partial Least Squares Determinant Analysis (PLSDA) to assess the usefulness of the identified signature. LASSO identified an age-adjusted signature of eight proteins that best differentiated the healthy patients from the patients with IPF. A PLSDA model of these eight selected proteins classified the two groups perfectly, with 100% calibration accuracy and 100% cross-validation accuracy, as well as 100% sensitivity and specificity for both the healthy and the IPF groups. Latent variable 1 (LV1) was able to completely differentiate between healthy patients (negative scores on LV1) and patients with IPF (positive scores on LV1; **Fig 2.4A**). Two of the eight proteins were loaded positively on LV1 (**Fig 2.4B**), indicating that they were positively associated with the IPF patients, whereas six proteins were loaded negatively on LV1, indicating that they were negatively associated with

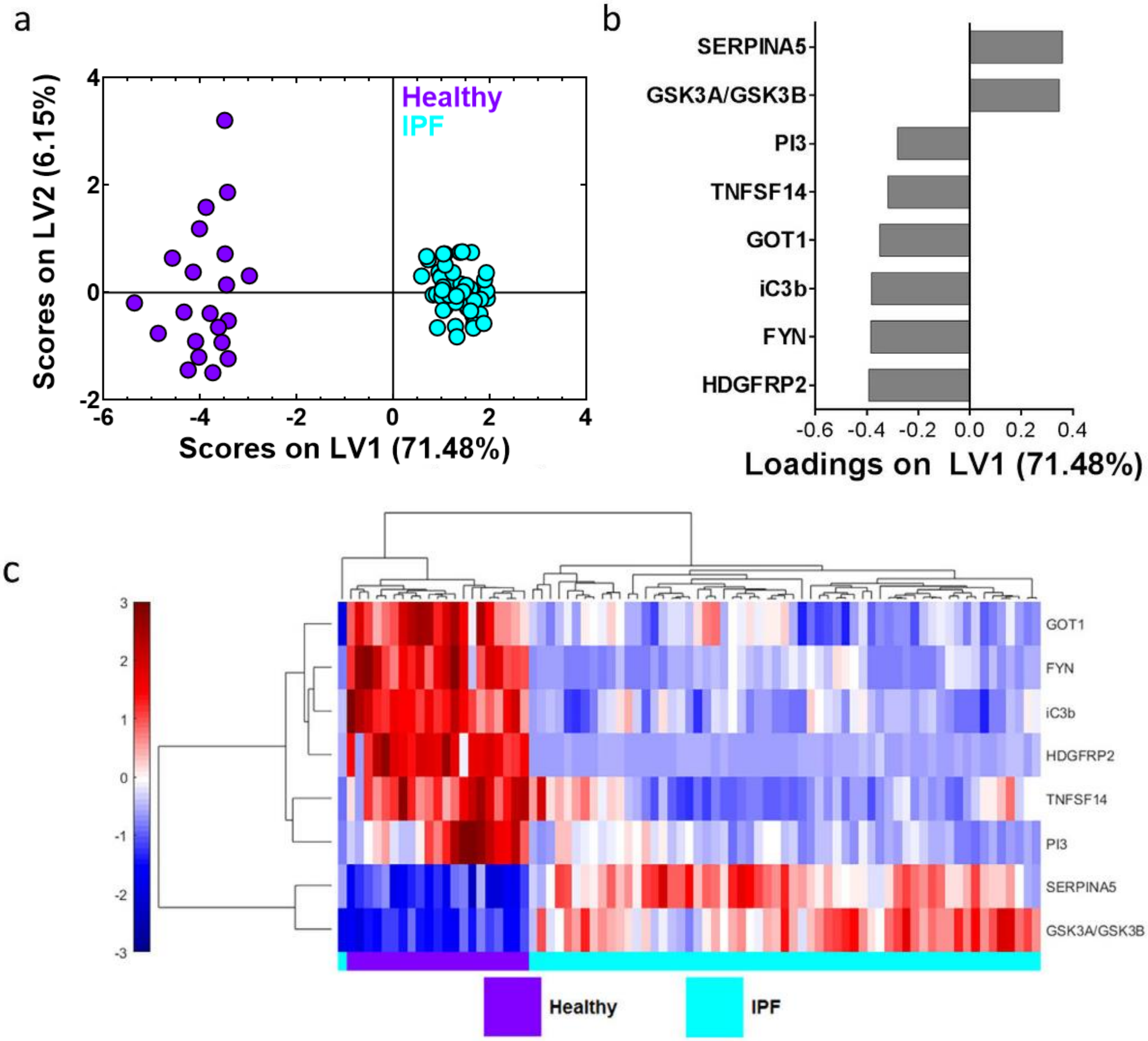


Figure 2.4. LASSO/PLSDA identified a minimum protein signature of 8 age-adjusted proteins that best differentiated healthy and IPF patients.
(**a**) LASSO identified an 8-protein signature that differentiated healthy (purple) and IPF (cyan) patients, with 100% calibration accuracy and 100% cross-validation accuracy, with 100% sensitivity and specificity for both healthy and IPF patients. Latent variable 1 (LV1) accounted for 71.48% of the variance in the data, and latent variable 2 (LV2) accounted for 6.15% of the variance in the data. (**b**) The loadings plot indicates protein contributions to the LASSO-identified signature, with positive loadings positively associated with IPF, and negative loadings comparatively reduced in IPF. (**c**) Hierarchical clustering further emphasizes the visual difference between healthy and IPF patients based on the LASSO-identified signature. Abundance of each protein is shown in color, with red indicating overabundance, white unchanged, and blue indicating underabundant proteins compared to the mean. Color bar scale is to the left of figure.

the IPF patients (**Fig 2.4B**). Not surprisingly, all of the proteins identified by LASSO were also found to be significantly different between healthy and IPF patients in the volcano plot (**Fig 2.1B**). LASSO and PLSDA were able to successfully separate individuals that were healthy from individuals with IPF; this suggests that the eight proteins in the signature may have relationships that are of biological interest. The LASSO-signature does include proteins that have clear immunological functions: inactivated (iC3b) and tumor necrosis factor ligand superfamily member 14 (TNFSF14 or LIGHT). This further suggests the potential importance of immune processes in the pathogenesis of IPF and warrants further investigation**.**

In order to better visualize patient clustering using our LASSO-identified signature, we performed hierarchical clustering and created a heat map of the LASSO-identified protein signature (**Fig 2.4C**). The result was readily-identifiable, near-perfect clustering of the healthy and IPF patients, with only one patient being misclassified. Interestingly, the two proteins in the hierarchical cluster that were overabundant in the IPF patients are the same two proteins that PLSDA identified as being positively associated with the IPF patients. Recalling that all eight of the proteins were also included in the biologically relevant, age-adjusted significantly different protein panel, these findings validate the LASSO-identified blood protein signature as being the preferred signature to differentiate the two groups of patients, and also support the idea that there are large differences in the blood proteome seen in healthy and IPF patients. We also analyzed the LASSO-identified protein signature using GO terms for biological process and molecular function. The most significantly upregulated functional annotation cluster involved peptidase inhibitors, endopeptidase regulators and catalytic activity (FE = 3.46, Bonferroni corrected P value = 0.0135) (**Supplementary Figure A.S5**). Overall these results provide proof-of-concept and suggest value for these approaches in the future development of a non-invasive diagnostic or prognostic assay for IPF. This could be especially useful for a diagnosis of IPF with relatively normal pulmonary function levels and/or atypical radiological findings.

2.4 Discussion

IPF remains a disease of unknown etiology with poorly understood pathophysiological mechanisms. Major advances have occurred in recent years through hypothesis-driven studies of potential biomarkers of the genome, transcriptome, chemokines and cytokines. In this paper we apply novel modified aptamer technology to produce large scale studies of proteins of variable abundance in the blood of IPF patients and normal controls for the first time. This novel approach to IPF has generated new hypothesis-provoking insight regarding the possible key functional biological abnormalities in IPF. The design and main focus of this study was to identify differentially expressed proteins in the blood of IPF patients compared to normal healthy controls and, through the employment of systems biology and bioinformatics tools, generate knowledge about the enriched biological processes that these proteins may represent.

Analysis of the downregulated protein profile identified a role for defense response encompassing a reaction to the presence of a foreign body or injury with an associated attempt to restrict damage and initiate repair. This is the most significantly enriched process within the downregulated protein panel. These data suggest that compared to a normal host, IPF patients have reduced levels of circulating proteins that support host defense. Indeed, the cohort of patients studied in this work (COMET study cohort) was previously employed in a project that supported a role for dysbiosis in the lung and disease progression. Alterations in the microbiome, namely an increase in *Streptococcal* and *Staphylococcal* operational taxonomic units were associated with disease progression in IPF[137]. Molyneaux *et al.* have reported an association between disease progression and increased bacterial burden in the lung[138]. An increased quantity of *Streptococcus* species was noted. Knippenberg *et al.* using murine models have demonstrated a mechanism by which a pneumococcal toxin, pneumolysin, exacerbates pulmonary fibrosis[139]. Our study of the proteome at trial screening suggesting a reduction in processes supporting host defense, supports a potential role for pathogens, particularly given further findings in the downregulated proteome involving the regulation of responses to external stimuli. These data enrich the evidence for a potential role for dysbiosis in IPF progression.

Features of acute inflammation including leucocyte chemotaxis, proliferation and phagocytosis are subject to downregulation in the blood compared to normal controls in our study. Several proteins involved in regulating the response to wounding appear inhibited in the plasma of patients with IPF compared to controls. We hypothesize that this finding is indicative of the recurrent injury and loss of the alveolar epithelial barrier. The proteome findings in this study support the paradigm of recurrent injury or wounding with aberrant repair. Indeed, our findings support an intrinsic impairment of the immune response to stimuli which may, in turn, promote insufficient or even exuberant responses to improve pathogen clearance but worsen bystander damage. The response of Toll like receptors (TLRs) and other pathogen recognition receptors to pathogen associated molecular patterns (PAMPs) and danger associated molecular patterns (DAMPs) is crucial to mounting a response to infection and injury[140]. IPF patients may have impaired responses to DAMPs and PAMPs. Studies of pathogen recognition receptors involved in responses to PAMPs/DAMPs including TLR 3 and TOLLIP have reported associations with IPF pathophysiology[7,119]. Furthermore, the role of immunosuppression is associated with poorer survival and higher levels of hospitalization in IPF patients[47]. The addition of agents responsible for attenuated immune responses may contribute negatively to a disease biology that features impaired responses to PAMPs and DAMPs.

The upregulated protein profile identified T cell co-stimulation as a process discriminating between normal and IPF patients. The role of T cell co-stimulation in regulation of lung fibrosis is controversial and complicated by the fact that measurements have been based on samples taken from different human compartments versus murine models. Studies to date have supported a role for decreased expression of inducible T cell co-stimulator (ICOS) in peripheral blood mononuclear cells (PBMCs) as a marker of disease progression and a predictor of poor survival outcomes[115,116]. However, animal models of bleomycin-induced pulmonary fibrosis reported higher levels of ICOS ligand (ICOSL) expression on macrophages and B cells in ICOS deficient mice compared to wild type which correlated with higher levels of fibrosis, thus highlighting a role for ICOSL expression in positively regulating pulmonary fibrosis. ICOS deficient mice had attenuated pulmonary fibrosis upon bleomycin challenge[141]. The role of ICOS and T cell co-stimulation warrants further study given our findings of enrichment of this process in the upregulated proteins when comparing IPF patients to normal controls. We have shown that ICOS may be secreted by activated T lymphocytes[137] and hypothesize that the loss of ICOS expression on cells may correlate with elevated plasma levels and that this may be accompanied by reduced transcription. Taken together, these changes suggest a crucial regulatory step in the pathobiology of IPF. Interestingly, the positive regulation of T cell activation is notably enriched within the downregulated plasma proteome in IPF patients suggesting that overall, IPF patients may have impaired T cell activity and this may be linked to disease biology, potentially via impaired defense against pathogens such as herpesviruses[142].

Protein phosphorylation is a fundamental mechanism of signal transduction and is achieved by kinase activity. The high signal for phosphorylation in our upregulated proteome may represent heightened kinase activity and both these processes are enriched within the upregulated proteome. In vitro studies and animal models have produced robust evidence to support a central role for protein kinase activity in pulmonary fibrosis, particularly tyrosine kinase activity including platelet derived growth factor (PDGF), epidermal growth factor (EGF), fibroblast growth factor (FGF), and vascular endothelial growth factor (VEGF)[143]. Nintedanib, a novel and approved tyrosine kinase inhibitor for IPF, robustly inhibits VEGF receptor, PDGF receptor, and FGF receptor with resultant modification of IPF fibroblast biology and improved patients outcomes[49,144,145]. VEGF signaling was additionally enriched within the upregulated plasma proteome of IPF patients in our work, consolidating its role in IPF pathogenesis. A key downstream event of ligation between these tyrosine kinases and their receptors is autophosphorylation and phosphatidylinositide 3-kinase activity[146,147]. ErbB signaling enrichment is also notable. These are a family of tyrosine kinase receptors, which include Her1 (epidermal growth factor receptor (EGFR)), Her2, Her3, and Her4. Several of these receptors have reported roles in epithelial remodeling and proliferation, and are found to play significant roles in models of fibrosis[148–150]. Further dysfunction within this pathway is supported by the finding of enrichment within the downregulated proteome for EGFR (Her1) signaling. EGFR is vital for normal epithelial repair so downregulation of this pathway could indicate impaired wound healing. Alternatively, we cannot rule out the possibility that EGFR signaling within the lung promotes fibrosis, but that the signature is lost in peripheral blood. Further investigation of the role of ErbB signaling in the pathogenesis of IPF is likely needed.

Platelet activation leads to the release of several profibrotic mediators and IPF patients have reported evidence of increased platelet reactivity and activation in a previous study[151]. It is possible that this is reflective of the IPF plasma environment. Complement and coagulation cascades have reported associations with IPF. Complement receptor polymorphisms may be associated with the development of IPF[152]. Furthermore, complement can augment epithelial injury in pulmonary fibrosis through crosstalk with Transforming Growth Factor-β (TGF-β)[153]. Gu et al. demonstrated that the inhibition of both complement component C3a and C5a receptors can lead to the arrest of fibrosis and may have therapeutic potential in IPF[154]. The enrichment within the plasma proteome of platelet activation and complement cascades is suggestive of ongoing injury that is detectable in the blood and will require further study.

The LASSO/PSLDA proteome signature we have identified includes novel proteins that have no previous reported associations with IPF. Armed with these target proteins however, it is interesting to speculate on their putative roles in pulmonary fibrosis. TNFSF14 (Tumor necrosis factor ligand superfamily member 14 or LIGHT) is an inflammatory molecule and a member of the TNF superfamily that our analysis also shows to be downregulated in IPF plasma compared to normal. Seemingly contradictory, the genetic deletion of LIGHT attenuates bleomycin-induced pulmonary fibrosis in animal models through the abolition of Thymic stromal lymphopoietin (TSLP) expression[155]. In addition, Herro et al. demonstrated that the administration of recombinant LIGHT to murine models produced features of fibrotic lung disease similar to the bleomycin fibrotic phenotype, via a TSLP-dependent mechanism. Human bronchial epithelial cells challenged with LIGHT in vitro generate TSLP production[155]. LIGHT appears to have potential as a regulator of fibrosis and its role in IPF requires further exploration. LIGHT can function as a mediator of herpes viral cell entry, hence its acronym Herpes Virus Entry Mediator (HVEM), and one may speculate a further mechanistic role for LIGHT in this context given the evolving roles of herpes virus in fibrotic lung disease exacerbations[142], but it may be informative to compare circulating vs. tissue measurements. Glycogen synthase kinase-3 alpha/Glycogen synthase kinase-3 beta(beta (GSK3A/GSK3B) are negative regulators of glucose homeostasis, Wnt signaling, and transcription factors, and this protein is positively associated with IPF. GSK3A/GSK3B inhibition in bleomycin-exposed mice has been shown to reduce alveolitis, lung fibrosis, and alveolar cell apoptosis[156]. GSK3A/GSK3B inhibition also decreased the production of monocyte chemoattractant protein-1 (MCP-1/CCL2) and tumor necrosis factor-α (TNF-α) by lung macrophages after bleomycin exposure in this study. Plasma serine protease inhibitor (SERPINA5), a molecule we find at elevated levels in IPF relative to control patients, has been shown to be upregulated in the intra-alveolar space of patients with interstitial lung diseases (IPF included), and is involved in the inhibition of fibrinolysis, especially in IPF[157]. A reduction in fibrinolysis causes more collagen, fibrin, and other extracellular matrix fibers to accumulate in the intra-alveolar space of these patients, leading to a stiffer lung and to formation of a matrix where fibroblasts can proliferate and release more collagen[158].

The acquisition of a distinct signature in the blood proteome of IPF patients that allows for discrimination between IPF and healthy controls is a significant proof of concept discovery. While we recognize that a blood test is not necessary to diagnose IPF patients from healthy volunteers, our work suggests that this methodology could be employed to help diagnose IPF from other forms of chronic lung disease. This will require further validation with larger numbers of patients, and exploration in other chronic lung diseases to determine whether differential signatures are producible in similar diseases. If true, the potential for change in clinical practice is considerable. The use of peripheral blood to identify disease-specific signatures may result in obviating the need for biopsy in patients who present with imaging features that are not consistent with IPF or possibly improve diagnostic confidence in patients who are not suitable for a surgical biopsy. Previous studies of plasma proteins in IPF patients identified both MMP-7 and MMP-1 as predictors of disease progression that were differentially expressed compared to normal plasma[35]. While there remain significant methodological differences between studies, we have found that MMP-7 is also upregulated in IPF plasma compared to normal.

There are several limitations to our study. The study numbers are limited and the IPF cohort, while extensively characterized, was not subject to death over the course of 80 week follow up. This population may not be fully representative of the IPF disease spectrum and we are not able to adjust for all potential confounding variables including co-morbidities within the IPF population. The absence of a validation cohort is a weakness; however, the main goal of this work was to generate hypotheses based on the proteomic data accrued. The use of slow off rate modified aptamers is novel and the aptamer results may not correlate with other protein measurement platforms. The aptamers bind to non-linear sequences with very high specificity for the selected target; this may explain some of the variance when measuring identical targets with other platforms such as ELISA[130]. However, several studies have demonstrated very high levels of agreement between the modified aptamer platform and ELISA[136,155].

Although we did not have a validation cohort to test the accuracy of our PLSDA model, we did investigate model accuracy through cross-validation. This involved excluding a small portion of the data (called the test set), building a model based on the rest of the data, and testing the accuracy of the model using the test set. By repeating this process many times and using different test sets, we were able to obtain the cross-validation accuracy by averaging the accuracy of each individual model. Thus despite the fact that there was not a validation cohort, we were still able to report a metric of model accuracy, which was calculated based on testing the model with unseen data. The final model we have reported on performed perfectly during cross-validation testing with 100% cross-validation accuracy.

Our work identified biological processes that discriminate IPF from healthy controls and generates hypotheses and new targets for investigation into disease mechanisms. Our study patients were recruited to a clinical trial with the highest standards of diagnostic approach and management. The prime purpose of this work was to introduce the approach of large scale unbiased biomarker screening and the generation of subsequent mechanistic hypotheses. However, given the proposed single organ nature of IPF, the biological signal detectable in blood is dilute and may not accurately reflect ongoing change within the lung. However, the peripheral blood has been employed in several biomarker studies in IPF to date[35,114,116] and represents an easily-accessible compartment for analysis. The fact that the identified proteome clustered differently between IPF and controls gives some confidence that analyses of peripheral blood may be useful.

In conclusion, this work furthers the evolving evidence supporting impaired host defense as a key marker of IPF disease biology and validates some of our current understanding. We generate further hypotheses about novel potential therapeutic targets and introduce a new approach to biomarker studies in IPF. The ability to identify a minimal signature that allows clinicians and researchers alike to discriminate IPF cases from normal serves as a proof of principle that this approach may have potential in defining other forms of chronic interstitial lung disease and the further evaluation of molecular endotyping in pulmonary fibrosis.

2.5 Methods

2.5.1 Study population

Subjects included in this analysis were a subset of patients who participated in a prospective observational study correlating biomarkers with disease progression (clinicaltrials.gov, clinical trials ID no. NCT01071707) (Correlating Outcomes with biochemical

Markers to Estimate Time-progression in Idiopathic Pulmonary Fibrosis - COMET). This cohort consisted of 60 patients who had samples available for analysis for at least 3 follow up time points, but this report focuses only on the baseline samples. Inclusion criteria required patients to be aged 35-80 years with a diagnosis of IPF. Exclusion criteria included a diagnosis of IPF that was >4 years prior to screening, a diagnosis of collagen-vascular disorder, FEV1/FVC<0.6, evidence of active infection at screening, or comorbid conditions other than IPF likely to result in death within one year. Subject follow up was for 80 weeks. Informed consent was obtained from all participating patients. The study protocol was reviewed and approved by the institutional review board of each participating center and methods were carried out in accordance with the relevant guidelines and regulations. Participating centers included: University of California Los Angeles. Los Angeles, CA, United States—University of California, San Francisco. San Francisco, CA, United States—National Jewish medical and Research Center, Denver, CO, United States—University of Chicago, Chicago, IL, United States—University of Michigan Ann Arbor, MI, United States—Cleveland Clinical Foundation, Cleveland, OH, United States—Temple University, Philadelphia, PA, United States—Brown University, Providence, RI, United States—Vanderbilt University, Nashville, TN, United States. Patients were enrolled from March 2010 to March 2011. Blood samples and demographic data were also acquired from healthy human controls (n = 21). Demographics are displayed separately for IPF patients and healthy normal participants, with mean and standard deviation for the continuous predictor age and the number and percentage enrolled for the categorical variable gender. Statistical significance of differences between the two groups of people for age and gender were assessed via Student's t test and Pearson's Chi-squared test, respectively (**Supplementary Table A.S1**). Patients were diagnosed as having IPF using a multidisciplinary approach as per published international guidelines[3]. In brief, the diagnosis of IPF was on the basis of features on computed tomography (CT) scans of the chest or usual interstitial pneumonia (UIP) pathology confirmed by lung biopsy. Cases were reviewed with expertise from radiologists, pathologists and clinicians at the local enrolling center. The number of biopsy proven cases was 35 of 60 patients, representing 57% of the study cohort. All cases and controls were of Caucasian ethnicity.

2.5.2 Sample acquisition and preparation

Peripheral blood was collected in EDTA-containing vacutainers at study centers and samples were shipped by overnight mail using cold packs to the University of Michigan. Samples were collected at 3 time points, namely screening, week 48 and week 80. Samples from healthy human controls were obtained from MedImmune and analyzed simultaneously with the COMET specimens. Whole blood was centrifuged at 2500 rpm for 10 minutes and plasma was collected and frozen at -80°C in small aliquots. Samples were shipped to SomaLogics for analysis on the SOMAscan® panel (1129 analytes). Plasma samples were diluted at 3 different concentrations for analysis on the aptamer array at the optimal concentrations for each SOMAmer©.

2.5.3 SOMAscan assay

The SOMAscan® proteomic assay has been described extensively in previous publications[130]. In brief, each of the listed proteins is measured using a modified aptamer reagent and measured quantitatively in relative fluorescence units (RFU’s) using a custom Agilent hybridization chip. Normalization and inter-run calibration were performed according to SOMAscan v3 assay data quality-control procedures as defined in the SomaLogic good laboratory practice quality system. A complete list of SOMAscan© analytes may be found online (http://www.somalogic.com/somalogic/media/Assets/PDFs/SSM-045-REV-1-SOMAscan-Assay-1-3k-Content.pdf).

2.5.4 Statistical analysis of SOMAscan assay results

Proteomic data is reported quantitatively as RFU's for 1129 analytes in 60 IPF patients and 21 healthy controls. For a graphic summary of our investigative approach see **Supplemental Figure A.S1**.

The initial approach first identified 203 proteins that differentiated IPF from controls. Relative fold change in blood protein levels were calculated by dividing the average intensity in IPF samples by the average intensity in the healthy samples. Statistical analysis between the healthy and IPF patients was performed by a standard two-tailed and two-sample t-test. Graphical representation of the proteomic data was created using GraphPad Prism software (v6.01 for Windows, GraphPad Software, La Jolla, CA). Significantly different proteins were those that passed a set false discovery rate threshold of 1%. Hierarchical clustering of significantly different proteins was generated by unsupervised average linkage hierarchical clustering using Pearson's correlation coefficient as the distance metric[159].

Upon comparison of epidemiological factors between the two groups, we found age to be slightly increased in the normal group. To account for this and identify age-adjusted proteomic differences, we performed linear regression with all biomarkers and age as predictors based on comparison between the IPF and normal cohort, and assessed mean analyte differences between IPF patients and controls adjusted for age. To account for multiple comparisons, we considered Benjamini-Hochberg false discovery rate methods[160], but eventually decided upon a more conservative Bonferroni correction to maintain an overall type I error of 0.01 and more aggressively screen analytes from the pool of candidates[161,162]. Altogether, this resulted in a refined volcano plot showing the age-adjusted proteome. Hierarchical clustering was then used to visualize how these proteins differentiated the healthy and IPF patients.

2.5.5 Analysis of the differentially expressed IPF proteome with DAVID and Cytoscape

To identify significantly enriched biological process that differentiated IPF from control, those proteins that passed initial screening steps (a Bonferroni correction and linear regression modelling for age) were catalogued into “upregulated” and “downregulated” profiles. In brief, proteins that were meaningfully “upregulated” or “downregulated” were deemed to have potentially significant biological roles in IPF patients compared to the control cohort. A fold increase over control mean of 1.5 and a fold decrease below control mean of 0.75 were used as thresholds for “upregulated” and “downregulated” proteins, respectively. These criteria selected out 48 upregulated proteins and 116 downregulated proteins when comparing IPF patients to controls (**Supplementary Table A.S2** and **A.S3**). Certain proteins were measured in combination (see **Supplementary Table A.S2** and **A.S3**). Certain proteins, i.e. inactivated or splice variants, measured by the SOMAscan array do not have unique UniProt identifiers available, and therefore the parent protein UniProt Identifier is reported. Functional annotation and visualization was employed using the Cytoscape (v3.3.0) software environment and the ClueGO (v2.2.5) plugin application[163,164]. In brief, for ClueGo analysis, Gene ontology levels and Kegg Pathways were explored with medium specificity and a Kappa score of >0.4. The Bonferroni correction was employed for each P value calculation. GO fusion was used to reduce redundancy with child-parent term fusion. P value of 0.05 was regarded as significant. Visualization was applied with Overview term labelling and term P value for nodal size. Functional annotation clustering and enrichment analysis was performed using Gene Ontology (GO) biological processes (BP FAT), molecular function (MF FAT), Kyoto Encyclopedia of

Genes and Genomes (KEGG). Enrichment analysis was undertaken by submitting these proteins to the Database for Annotation, Visualization and Integrated Discovery (DAVID) (http://david.abcc.ncifcrf.gov/)[165,166]. Enrichment analysis was performed on the basis of *uniprot_accession* as identifier and *gene list* as list type, medium stringency and Bonferroni correction was applied. Enrichment chart analysis was performed using Gene Ontology (GO) biological processes (BP FAT), GO molecular function (MF FAT) and Kyoto Encyclopedia of Genes and Genomes (KEGG). The top functional annotation clusters with significant enrichment scores were identified.

2.5.6 Identification of a minimal IPF proteomic signature with hierarchical clustering and PLSDA

The Least Absolute Shrinkage and Selection Operator (LASSO) method[159] was used to identify a minimum, age-adjusted protein signature that best differentiated IPF and normal proteomes and was implemented using Matlab software[167] (Mathworks, Natick, MA). *K*-fold cross-validation was used to generate the model that had the lowest possible mean squared error for prediction. Associated features for this model were chosen as the minimum set of biomarkers. In order to allow for age-adjustment in the LASSO model, age was forced into the model as a parameter and assigned zero penalty. PLSDA assessed the usefulness of the LASSO-identified protein signature for differentiating healthy and IPF patients. Data were normalized with mean centering and variance scaling, and cross-validation was performed by iteratively excluding random subsets in groups of 9-10 data points during model calibration. Excluded data samples would then be used to test model predictions. Hierarchical clustering of LASSO-identified proteins was generated by unsupervised average linkage hierarchical clustering using Pearson’s correlation coefficient as the distance metric.

2.5.7 Investigating the effect of comorbidities in IPF had on the LASSO and PLSDA analysis

To investigate whether or not the comorbidities present in some IPF patients affected the feature selection by LASSO or the clustering in PLSDA, we performed a principal component analysis (PCA) on all of the measured blood proteins in the healthy and IPF patients. PCA was chosen as the method of analysis due to the lack of knowledge of the comorbidities seen within the healthy cohort. Gastroesophageal reflux disease (GERD) and obstructive sleep apnea (OSA) were examined based on their prevalence in the IPF patients (34 patients with GERD and 12 patients with OSA).

Chapter 3 Identification of a Unique Temporal Signature in Blood and BAL Associated with IPF Progression

Katy C. Norman[1], David N. O'Dwyer[2], Margaret L. Salisbury[3], Katarina M. DiLillo[1], Vibha N. Lama[2], Meng Xia[4], Stephen J. Gurczynski[2], Eric S. White[2], Kevin R. Flaherty[2], Fernando J. Martinez[5], Susan Murray[4], Bethany B. Moore[2,6], and Kelly B. Arnold[1]

[1] Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI 48109

[2] Department of Internal Medicine, Division of Pulmonary and Critical Care Medicine, University of Michigan Medical School, Ann Arbor, MI, USA

[3] Department of Medicine, Division of Allergy, Pulmonary and Critical Care Medicine Vanderbilt University Medical Center, Nashville, TN, USA

[4] Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, USA

[5] Department of Internal Medicine, Weill Cornell School of Medicine, New York, NY, USA

[6] Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, MI, USA

Scientific Reports. 2020; 10(1): 12049.

3.1 Abstract

Idiopathic pulmonary fibrosis (IPF) is a progressive and heterogeneous interstitial lung disease of unknown origin with a low survival rate. There are few treatment options available due to the fact that mechanisms underlying disease progression are not well understood, likely because they arise from dysregulation of complex signaling networks spanning multiple tissue compartments. To better characterize these networks, we used systems-focused data-driven modeling approaches to identify cross-tissue compartment (blood and bronchoalveolar lavage) and temporal proteomic signatures that differentiated IPF progressors and non-progressors. Partial least squares discriminant analysis identified a signature of 54 baseline (week 0) blood and lung proteins that differentiated IPF progression status by the end of 80 weeks of follow-up with 100% cross-validation accuracy. Overall we observed heterogeneous protein expression patterns in progressors compared to more homogenous signatures in non-progressors, and found that non-progressors were enriched for proteomic processes involving regulation of the immune/defense response. We also identified a temporal signature of blood proteins that was significantly different at early and late progressor time points (p<0.0001), but not present in non-progressors. Overall, this approach can be used to generate new hypotheses for mechanisms associated with IPF progression and could readily be translated to other complex and heterogeneous diseases.

3.2 Introduction

Idiopathic pulmonary fibrosis (IPF) is a heterogeneous and irreversible interstitial pneumonia, with symptoms including progressive cough, shortness of breath, and ultimately respiratory failure, with a median survival of only 3-5 years post diagnosis[5]. The disease is believed to be caused by a dysregulated wound healing response to various epithelial injuries leading to fibrosis of the lung interstitium[5]. Two medications (nintedanib[49] and pirfenidone[48]) are effective treatments for IPF; both are able to temporarily slow disease progression without reversing established fibrosis[168]. Thus, lung transplantation is currently the only option to cure IPF[3], even though this procedure has the highest failure rate of all organ transplantation options (54% at 5 years[169]). Better understanding of mechanisms underpinning progression of pulmonary fibrosis could lead to improved outcomes via identification of new therapeutic targets.

To add to the complexity surrounding IPF, disease progression is also heterogeneous, with some individual patients experiencing long-term stability and others rapid loss of lung function. A number of longitudinal cohort studies have been created with the goal of better characterizing IPF pathobiology using proteomic measurements[137,170–172]. These efforts have identified individual proteins, including blood MMP-7[39,40], CCL18[41], and blood surfactant protein D[43,44], as potential prognostic biomarkers. However, it has been difficult to replicate these findings across multiple cohorts[36,45], especially when attempting to validate specific, prognostically-relevant cut-off concentrations[45,46].

One potential explanation for failure to validate a specific prognostic biomarker is that disease progression is driven by dysregulated proteomic signaling networks rather than individual proteins. This hypothesis is supported by the multiple known actions of the two FDA-approved drugs that slow IPF progression: nintedanib[173] and pirfenidone[173]. The use of quantitative approaches to capture individual proteins within large clinical "omics" data sets has become a useful way to find new proteins associated with disease progression. Groups of proteins associated with progression that were identified by these approaches were characterized by biologically relevant functions, such as involvement in the immune system[111,114,136], tissue reorganization[109,114,136], and epithelial cell function[109]. While these results have highlighted potential prognostic biomarkers and biological functions associated with IPF progression, many of the techniques used in these discoveries emphasize the additive significance of each protein's individual ability to differentiate progression status but do not capture protein "signatures", or take into account potential protein networks associated with progression. In addition, none of these large scale blood proteomics studies investigated quantitative proteomic relationships across other tissue compartments such as the lung.

Data-driven ("machine learning") modeling approaches are able to integrate data across multiple tissue compartments and assays to identify signatures of factors that are associated with the disease state[92,102]. They serve as valuable tools for network inference by identifying co-varying factors that aid in generating new hypotheses for mechanisms of action based on protein interaction pathways rather than individual proteins. Once identified and validated, these signatures may be used for diagnostic or prognostic purposes, or for generating new hypotheses for future experimental work. We have previously used these approaches to successfully identify a blood protein signature that differentiated healthy and IPF patients with high accuracy[174], as well as signatures based on blood and sputum proteins and blood cell markers that differentiated stable and exacerbated chronic obstructive pulmonary disease (COPD) patients[175].

In this work, we applied data-driven modeling approaches to blood and bronchoalveolar lavage (BAL) samples from patients enrolled in the COMET-IPF (Correlating Outcomes with Biochemical Markers to Estimate Time-progression in Idiopathic Pulmonary Fibrosis) study to gain insight into cross-tissue compartment and temporal mechanisms of action associated with IPF progression. We identified a signature of blood and BAL proteins that differentiated IPF progressors and non-progressors with high accuracy. This signature indicated more heterogeneous progressor subgroups compared to non-progressors, and that proteins elevated in non-progressors were enriched for regulation of immune, defense, and inflammatory responses. Lastly, using measurements across multiple time points, we were able to identify a signature indicative of temporal changes in the blood of progressors that was not present in non-progressors. Overall these results provide insight into mechanisms of IPF progression that could be investigated further in follow-up murine studies.

3.3 Results

3.3.1 Only a small number of individual blood proteins are differentially expressed across IPF progressors and non-progressors

We evaluated a subset of participants (n=59) with an IPF diagnosis enrolled in the COMET IPF study. Participants were defined as progressors (n=34) if at the end of the 80 week study they had experienced death, lung transplantation, an acute exacerbation of IPF (AE-IPF), or a drop in forced vital capacity (FVC) of >10% or in diffusing capacity of the lung for carbon monoxide (DLCO) of >15%[137]. Otherwise participants were defined as non-progressors (n=25; demographics in **Supplemental Table B.S1**). Three blood draws from these 59 participants at week 0/baseline, 48, and 80 were used to measure the concentration of 1129 proteins (enriched for inflammation and cancer involvement) with SOMAmer$^{©}$ (slow off rate modified aptamer) technology (SomaLogic). One baseline (week 0) BAL sample was also collected from 51 individuals (31 progressors and 20 non-progressors, 50 of whom also had a baseline blood draw included in this analysis; demographics in **Supplemental Table B.S2**), and the concentration of 29 cytokines were measured with Luminex technology. There were no significant differences in demographic variables between the progressors and non-progressor groups, and all patients survived until the end of the 80-week study. Correlations in periostin SOMAmer aptamer and ELISA measurements within these samples have previously been published[136]. To build on this, in **Supplemental Table B.S3** we report significant Pearson's correlations (all p < 0.03) between

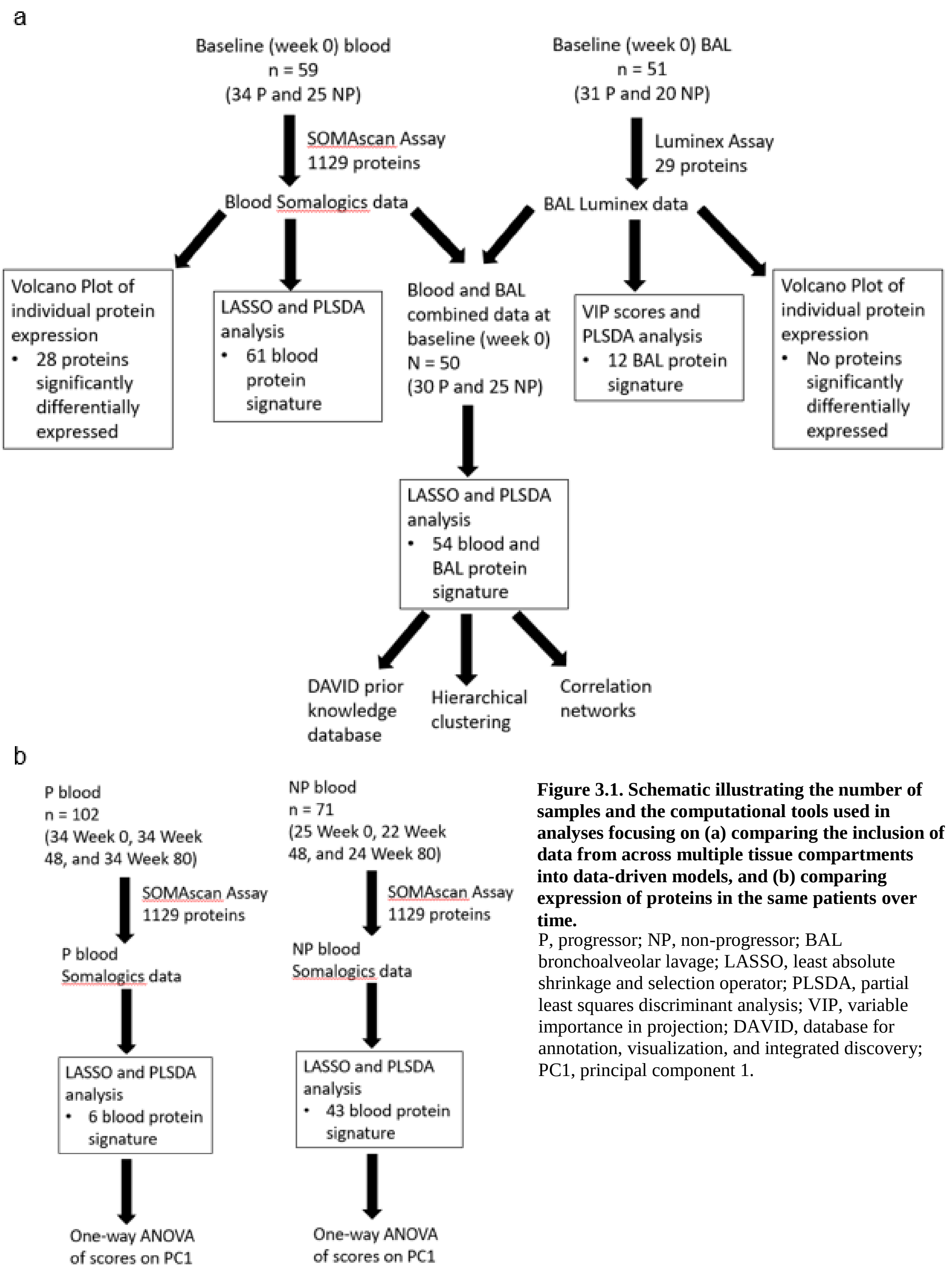
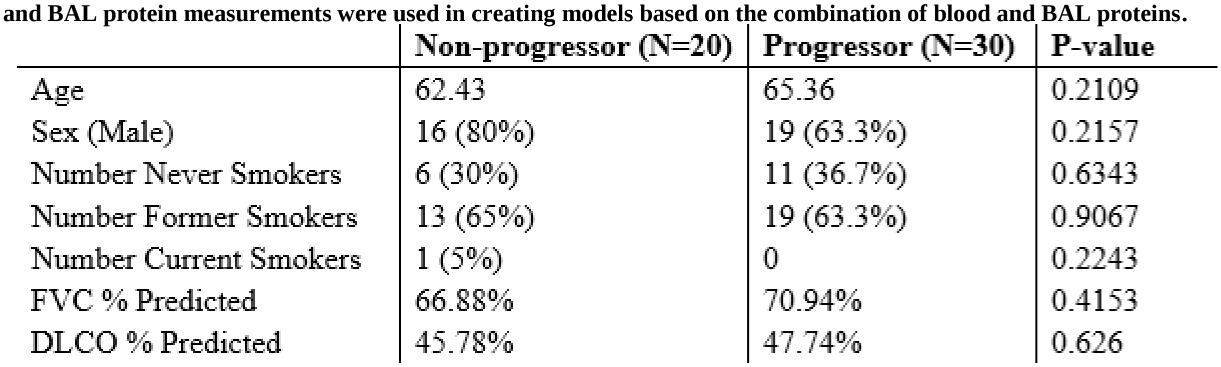

a
Baseline (week 0) blood
n = 59
(34 P and 25 NP)
SOMAscan Assay
1129 proteins
Blood Somalogics data
Baseline (week 0) BAL
n = 51
(31 P and 20 NP)
Luminex Assay
29 proteins
BAL Luminex data
Volcano Plot of individual protein expression
• 28 proteins significantly differentially expressed
LASSO and PLSDA analysis
• 61 blood protein signature
Blood and BAL combined data at baseline (week 0)
N = 50
(30 P and 25 NP)
VIP scores and PLSDA analysis
• 12 BAL protein signature
Volcano Plot of individual protein expression
• No proteins significantly differentially expressed
LASSO and PLSDA analysis
• 54 blood and BAL protein signature
DAVID prior knowledge database
Hierarchical clustering
Correlation networks
b
P blood
n = 102
(34 Week 0, 34 Week 48, and 34 Week 80)
SOMAscan Assay
1129 proteins
P blood
Somalogics data
LASSO and PLSDA analysis
• 6 blood protein signature
One-way ANOVA of scores on PC1
NP blood
n = 71
(25 Week 0, 22 Week 48, and 24 Week 80)
SOMAscan Assay
1129 proteins
NP blood
Somalogics data
LASSO and PLSDA analysis
• 43 blood protein signature
One-way ANOVA of scores on PC1

Figure 3.1. Schematic illustrating the number of samples and the computational tools used in analyses focusing on (a) comparing the inclusion of data from across multiple tissue compartments into data-driven models, and (b) comparing expression of proteins in the same patients over time.

P, progressor; NP, non-progressor; BAL bronchoalveolar lavage; LASSO, least absolute shrinkage and selection operator; PLSDA, partial least squares discriminant analysis; VIP, variable importance in projection; DAVID, database for annotation, visualization, and integrated discovery; PC1, principal component 1.

Table 3.1. Demographic and lung function test descriptions from progressors and non-progressors whose baseline blood and BAL protein measurements were used in creating models based on the combination of blood and BAL proteins.



	Non-progressor (N=20)	Progressor (N=30)	P-value
Age	62.43	65.36	0.2109
Sex (Male)	16 (80%)	19 (63.3%)	0.2157
Number Never Smokers	6 (30%)	11 (36.7%)	0.6343
Number Former Smokers	13 (65%)	19 (63.3%)	0.9067
Number Current Smokers	1 (5%)	0	0.2243
FVC % Predicted	66.88%	70.94%	0.4153
DLCO % Predicted	45.78%	47.74%	0.626

SOMAmer and ELISA concentrations for CCL22, CCL18, and CCL2, but not for IL-10 or CXCL12 (both $p > 0.45$). Our analysis pipeline is illustrated in **Figure 3.1**: **Figure 3.1A** focuses on analyses of baseline (week 0) expression of proteins in the blood and/or BAL samples of COMET patients, and **Figure 3.1B** focuses on analyses of the temporal change in blood protein expression (week 0, week 48, and week 80).

We first determined which of the measured baseline (week 0) 1129 blood and 29 BAL proteins were individually differentially expressed between IPF progressors (n=30) and non-progressors (n=20; demographics of these 50 patients are found in **Table 3.1**). A two-sample t-test was applied to each protein expression in progressors and non-progressors and revealed that 28 blood proteins were significantly different across the two groups; 17 proteins were increased in the progressors (fold change greater than 1) (**Figure 3.2A**; blue markers indicate a $p < 0.05$ and red indicate $p < 0.01$). The ten most significantly different blood proteins included E-Cadherin (cadherin E; fold change 1.19); DC-SIGN (CD209 antigen; fold change 1.30); a2-macroglobulin (fold change 1.24); ficolin-2 (FCN2; fold change 0.86); interleukin 17D (IL-17D; fold change 0.91); legumain (LGMN; fold change 0.87); C5b,6 complex (fold change 0.93); apolipoprotein B (ApoB; fold change 1.38); and neuroligin-4, X-linked (NLGNX; fold change 1.24). Except for TGM3 (protein-glutamine gamma-glutamyltransferase E; fold change of 2.47), all significant proteins had fold change values that ranged from 0.80 to 1.48. No BAL proteins

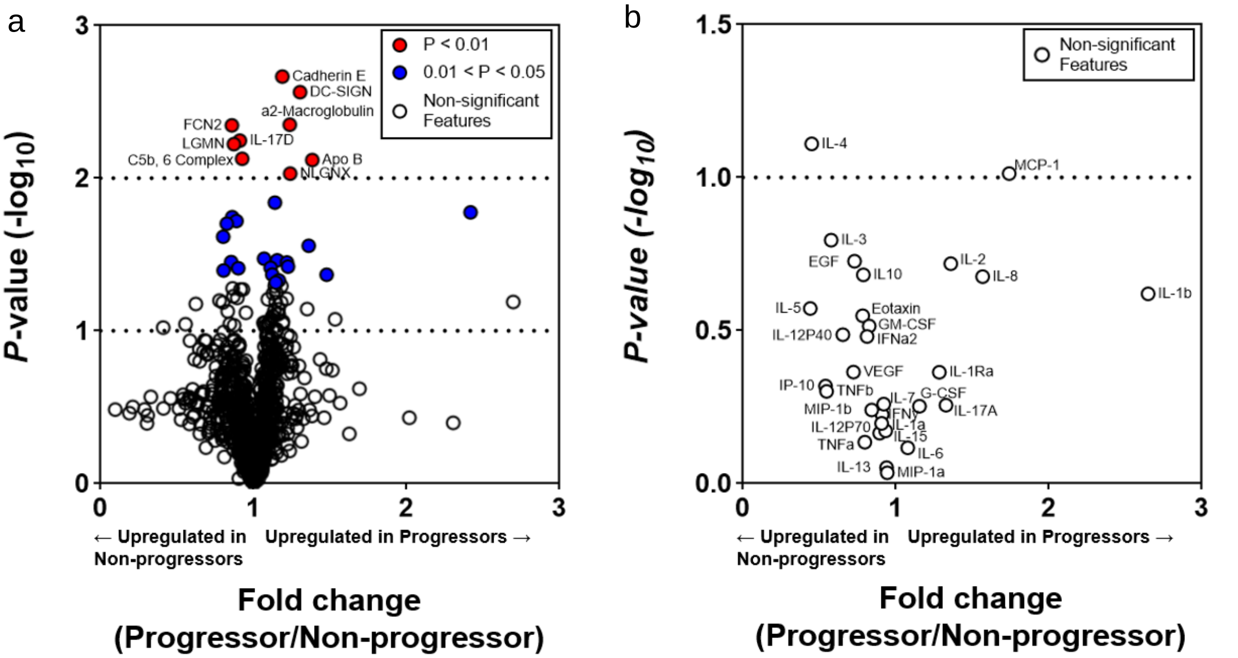


Figure 3.2. Volcano plot of blood (a) and BAL (b) proteins measured in COMET progressors and non-progressors. Proteins with a fold change greater than one are increased in progressors; fold changes less than one indicates elevation in non-progressors. Blue protein markers have a p-value < 0.05 after a two-tailed, two-sample t-test; red markers indicate p-value < 0.01 after the same test. No blood or BAL proteins were significantly different between progressors and non-progressors after adjusting for multiple comparisons using the Bonferroni correction.

were found significantly differentially expressed (**Figure 3.2B**). No proteins in blood or BAL were significant after application of the Bonferroni correction for multiple comparisons.

3.3.2 Data-driven analyses identify best signatures in single tissue compartments that differentiate IPF progression status

Due to the low number of significantly differentially expressed proteins in the univariate analysis, we next explored whether data-driven modeling techniques could identify signatures of proteins from single tissue compartments that differentiated IPF progressors and non-progressors. Our analysis pipeline that focused on baseline (week 0) expression of proteins in the blood and/or BAL samples of COMET patients is visualized in **Figure 3.1A**. We used the least absolute shrinkage and selection operator (LASSO[159]) as a feature selection tool to identify a signature of baseline (week 0) blood proteins that would best differentiate COMET participants based on progression status at 80 weeks. For every LASSO model in this analysis, k-fold cross-validation (k=10; see **Methods**) was performed to prevent over-fitting. Feature selection was accomplished in the BAL proteins through the use of variable importance in projection (VIP) scores. We then employed partial least squares discriminant analysis (PLSDA[176]) in order to visualize the separation power of the identified signatures. By highlighting co-varying relationships within protein signatures, PLSDA aids in generating new hypotheses about proteomic pathways associated with each group. For every PLSDA model in this analysis, we calculated calibration and k-fold cross-validation accuracy (k=10) to use as metrics of model performance for comparing PLSDA models generated from data in different tissue compartments (see Methods). LASSO identified a signature of 61 blood proteins that differentiated 25 non-progressors and 34 progressors (demographics in **Supplemental Table B.S1**); a PLSDA model based on this signature had 100% calibration and 96.53% cross-validation accuracy, and 97.06% sensitivity and 99.56% specificity for progressor identification (**Supplemental Figure B.S1A** and **B.S1B**; ROC curves in **Supplemental Figure B.S2**). The PLSDA model based on 12 VIP-selected baseline (week 0) BAL proteins differentiated 20 non-progressors and 31 progressors (demographics in **Supplemental Table B.S2**) with 78.55% calibration and 67.82% cross-validation accuracy (**Supplemental Figure B.S3A** and **B.S3B**; ROC curves in **Supplemental Figure B.S4**). Although these models performed with moderate to excellent accuracy, we wanted to explore the unique biological insight that might be gained from a model based on the combination of the data from the two tissue compartments.

3.3.3 Cross-tissue compartment signature differentiates COMET participants based on progression status

We combined measurements of the 1129 blood proteins and 29 BAL proteins from baseline samples to identify a cross-tissue compartment signature of co-varying proteins associated with progression. LASSO identified a signature of 54 baseline (week 0) proteins (51

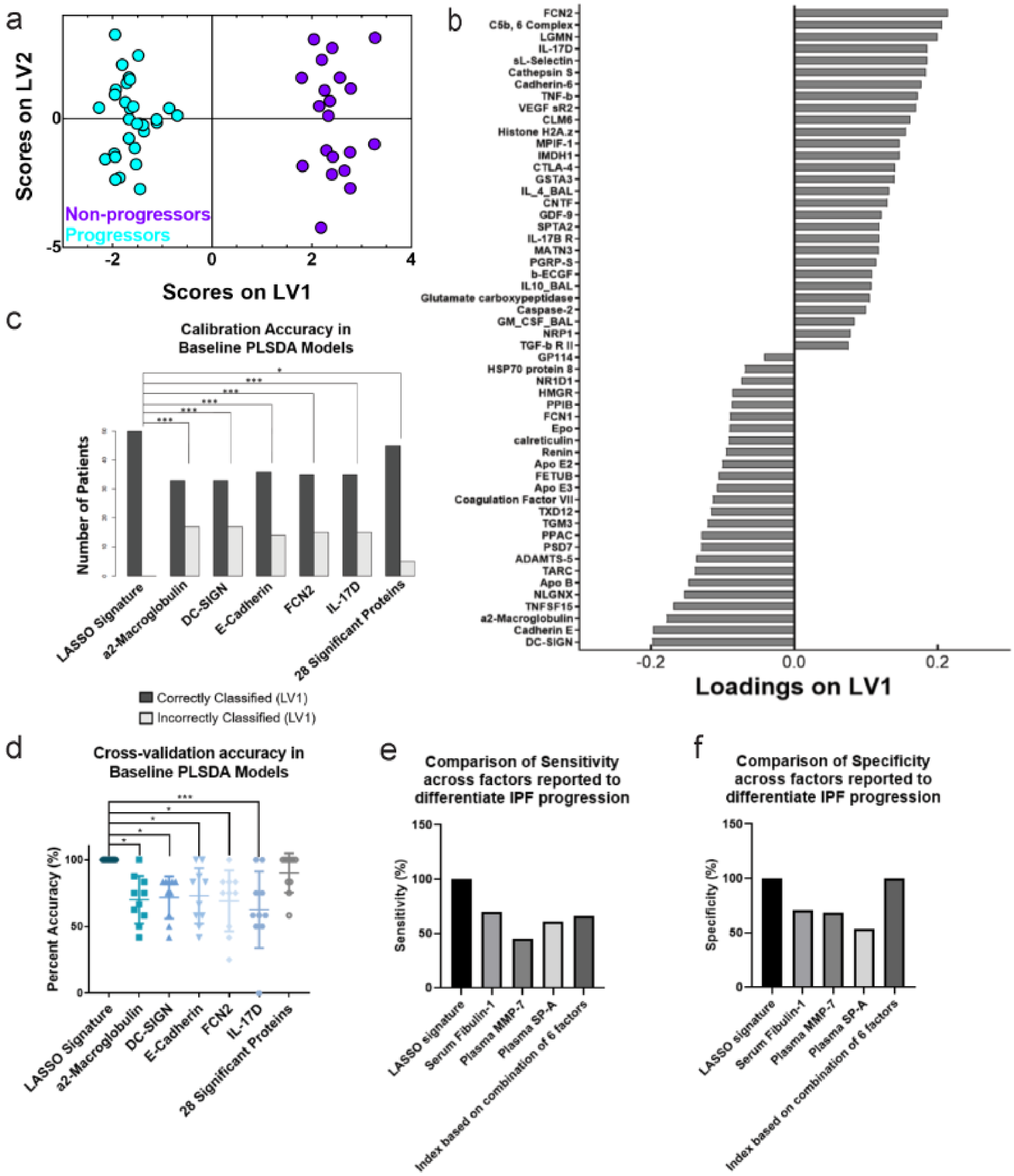


Figure 3.3. The LASSO-identified signature based on blood and BAL proteins separated progressors and non-progressors with high accuracy and significantly outperformed analyses based on individual factors.

(**a**) PLSDA scores plot based on blood and BAL proteins highlights strong differentiation between progressors (cyan) and non-progressors (purple); the model separated the two groups with 100% cross-validation and calibration accuracy. (**b**) The loadings on latent variable 1 (LV1) captured 8.75% of the total variance in the data, with negatively loaded proteins being comparatively increased in progressors and positively loaded proteins being comparatively reduced. (continued on next page)

Figure 3.3 caption continued (**c**) Comparison of the calibration accuracies between analyses based on data-driven signatures and univariate factors. The LASSO-selected PLSDA model based on blood and BAL proteins had significantly higher calibration accuracy than all analyses based on single proteins and a model based on the collection of all 28 significantly different proteins identified in Figure 3.2 (Cochran's Q test with McNemar's post hoc test; * indicates $p < 0.05$ and *** indicates $p < 0.001$). (**d**) Comparison of cross-validation accuracies between analyses based on data-driven signatures and univariate factors. The LASSO-selected PLSDA model based on blood and BAL proteins had significantly higher cross-validation accuracy than all analyses based on single proteins and trended towards better cross-validation accuracy than a model based on the 28 proteins identified in Figure 3.2 (one-way ANOVA with Tukey's post hoc test; * indicates $p < 0.05$ and *** indicates $p < 0.001$). (**e**) Comparison of sensitivity between the LASSO-selected PLSDA model based on blood and BAL proteins and previously published models of IPF progression (serum fibulin-1[177], plasma MMP-7[178], plasma SP-A[178], and an additive combination of blood factors[136]. (**f**) Comparison of specificity between the LASSO-selected PLSDA model based on blood and BAL proteins and previously published models of IPF progression.

in blood and 3 in BAL) that best separated progressors and non-progressors (comparison of protein signature expression in progressors and non-progressors can be found in **Supplemental Figure B.S5**). A PLSDA model based on this signature classified the two groups with 100% cross-validation and calibration accuracy (**Figure 3.3A**), with 100% sensitivity and specificity for each group (ROC curves in **Supplemental Figure B.S6**) and with positive and negative predictive values of 100%. Latent variable 1 (LV1) differentiated progressors (negative scores on LV1) from non-progressors (positive scores on LV1) (**Figure 3.3B**). Interestingly, we did not find significant Pearson's correlations between the scores on LV1 in this signature and the concentration of KL-6 ($r= 0.15$, $p=.31$), MMP7 ($r = -0.08$, $p=0.60$), or CCL18 ($r = 0.04$, $p=0.77$), which were other previously identified individual biomarkers of progression. However, we did see a significant correlation between the LV1 scores and the change in FVC percent predicted over the 80 weeks of the study ($r = 0.534$, $p = 0.00011$, Pearson's correlation coefficient).

We compared this model to cross-validated PLSDA analyses based on single significant proteins identified in the volcano plot, as well as a cross-validated PLSDA model based on the collection of the 28 differentially expressed blood proteins in the volcano plot (ROC curves for last model shown in **Supplemental Figure B.S7**). The model based on the LASSO-identified signature had significantly higher calibration accuracy than all of the analyses based on the individual proteins and the collection of the differentially expressed proteins (**Figure 3.3C**; Cochran's Q test with McNemar's post hoc test). In terms of cross-validation accuracy, the

LASSO-identified model also significantly outperformed analyses based on all of the individual Proteins, and trended towards outperforming the model based on the collection of the 28 differentially expressed proteins (**Figure 3.3D**; one-way ANOVA).

We also compared this model to other previously published single markers and combinations of markers that were shown to differentiate IPF progression status. The model based on our signature had 100% sensitivity and specificity, which outperformed previously published models that predicted IPF progression based on single factors (serum fibulin-1, 70% sensitivity and 71% specificity[177]; plasma MMP-7, 45.3% sensitivity and 68.5% specificity[178]; and plasma SP-A, 60.9% sensitivity and 53.9% specificity[178]), as well as a previously published model based on an additive combination of blood factors, where a score of ≥7 on the created index had a 66% sensitivity and 100% specificity for progression[136] (**Figure 3.3E**, **3.3F**).

We next sought to determine if the PLSDA model based on the combination of blood and BAL proteins was a better classifier than models based on signatures of blood or BAL proteins alone. The model based on blood proteins alone and the model based on blood and BAL proteins combined had significantly higher calibration accuracy than the model based on BAL proteins alone (**Supplemental Figure B.S8A**, $p = 0.0016$ for marked comparisons; Cochran’s Q test with McNemar’s post hoc test applied to calibration accuracy of patients that were included in all three models). McNemar’s post hoc test could not be applied when comparing the calibration accuracies of the blood protein model and the combination model because all patients were classified correctly in both models. When comparing cross-validation accuracies across the three models, again the model based only on BAL proteins performed significantly worse than the blood protein model and the combination model (**Supplemental Figure B.S8B**, $p = 0.0001$ for the blood protein vs. BAL protein model comparison and $p < 0.0001$ for the BAL protein vs.

combination model comparison, one-way ANOVA with Tukey's post hoc test applied to cross-validation accuracy based on all patients in all three model).

One reason the model based on BAL proteins had lower calibration and cross-validation accuracies might involve the high number of measured blood versus BAL proteins (1129 blood proteins vs. 29 BAL proteins). To investigate the potential effect of signature size on model accuracy, we created two new PLSDA models: one based on the top 12 loaded features of the blood signature; and the other based on the top 11 loaded proteins (all of which were blood proteins) and the top loaded BAL protein in the combination signature, for a total of 12 proteins in this shortened combination signature. When comparing the calibration accuracies of these models with the same signature size, there was no significant difference between the performance of the BAL protein model and the shortened blood protein model (p = 0.78, Cochran's Q test with McNemar's post hoc test). However, the calibration accuracy of the shortened combination model trended towards being significantly better than both of the BAL protein and the shortened blood protein models (p = 0.052 for both comparisons, Cochran's Q test with McNemar's post hoc test, **Supplemental Figure B.S9A**). There were no significant differences in cross-validation accuracy across any of the models, but again the shortened combination model trended towards significantly outperforming the BAL protein model (p = 0.12, one-way ANOVA with Tukey's post hoc test; **Supplemental Figure B.S9B**). Overall this suggests that the model based on blood proteins alone may have performed well due to the large panel of proteins measured, though the combination model still trends towards being significantly better than the BAL model even when the signature is shortened. We next explored the biological significance of the combination signature.

3.3.4 Non-progressors have enriched regulation of immune and defense response, and protein expression patterns suggest more heterogeneity in progressors

The database for annotation, visualization and integrated discovery (DAVID[165]) determined the proteins that were comparatively increased in the non-progressors in the LASSO-identified signature based on blood and BAL proteins were significantly enriched for processes involving immune and defense response regulation (**Figure 3.4**, enrichment score (ES) 4.83). Other functions enriched in non-progressors included cell signaling and regulation of basic cell processes (**Supplemental Figure B.S10A**, ES 2.57), and regulation of inflammatory, defense, and immune responses (**Supplemental Figure B.S10B**, ES 2.50). DAVID identified that proteins that were comparatively increased in progressors were only enriched for stress response regulation (**Supplemental Figure B.S11**, ES 2.05).

ciliary neurotrophic factor(CNTF)	IL17D	selectin L	TGFBR2	CD300c molecule	IL17RB	CCL23	CTLA4	ficolin 2	legumain	IL10	cathepsin S	lymphotoxin alpha(LTA)	peptidoglycan recognition protein 1(PGLYRP1)	IL4	complement C6	complement C5	Pathway	Bonferroni Corrected P-value
□	□	□	□	□	■	■	□	□	□	■	□	■	■	■	■	■	regulation of inflammatory response	0.001706961
□	□	□	□	□	□	□	■	■	■	□	■	■	■	■	■	■	positive regulation of immune response	0.030104629
□	□	□	□	□	■	■	□	□	■	■	■	■	■	■	■	■	regulation of defense response	0.002697466
■	□	□	■	□	■	■	□	□	■	■	■	■	■	■	■	■	regulation of response to stress	0.010235167
□	□	□	■	□	□	□	■	■	■	■	■	■	■	■	■	■	positive regulation of immune system process	0.00557428
□	□	■	□	■	□	□	■	■	■	■	■	■	■	■	■	■	regulation of immune response	0.000343634
□	■	□	□	■	■	■	□	■	■	■	■	■	■	■	■	■	defense response	0.007373938
□	□	■	■	■	□	□	■	■	■	■	■	■	■	■	■	■	regulation of immune system process	0.00265621

■ Indicates involvement in process □ Indicates no involvement in process

Figure 3.4. DAVID enrichment analysis of the blood and BAL LASSO-identified proteins that were comparatively elevated in the non-progressor group in the PLSDA loadings plot showed enrichment for pathways involved in the regulation of the inflammatory, defense, and immune responses after application of the Bonferroni correction (enrichment score 4.83).
Black squares indicate protein involvement in a particular pathway, while white squares indicate non-involvement.

We next used hierarchical clustering to visualize the individual expression of the proteins in the blood and BAL protein signature across all the patients. We saw four clusters that corresponded to the two groups, with one cluster composed only of non-progressors and three clusters that were mostly progressors (**Figure 3.5**). Only 5 non-progressors were misclassified out of 50 patients total (90% classification accuracy; 100% sensitivity and 75% specificity for identification of progressors). There were minor differences in classification accuracy of the PLSDA model and hierarchical cluster, likely due to underlying algorithmic differences associated with unsupervised identification of groups via the Pearson distance metric (hierarchical clustering) vs. supervised identification of groups based on maximized covariance in protein expression (PLSDA). Interestingly, there was heterogeneity within the progressor cluster, which was characterized by expression of different proteins. One of the progressor clusters had many apolipoproteins overexpressed compared to the mean (apolipoproteins E2, E3, and B), as well as cadherin E and DC-SIGN. Other progressors had high expression levels of

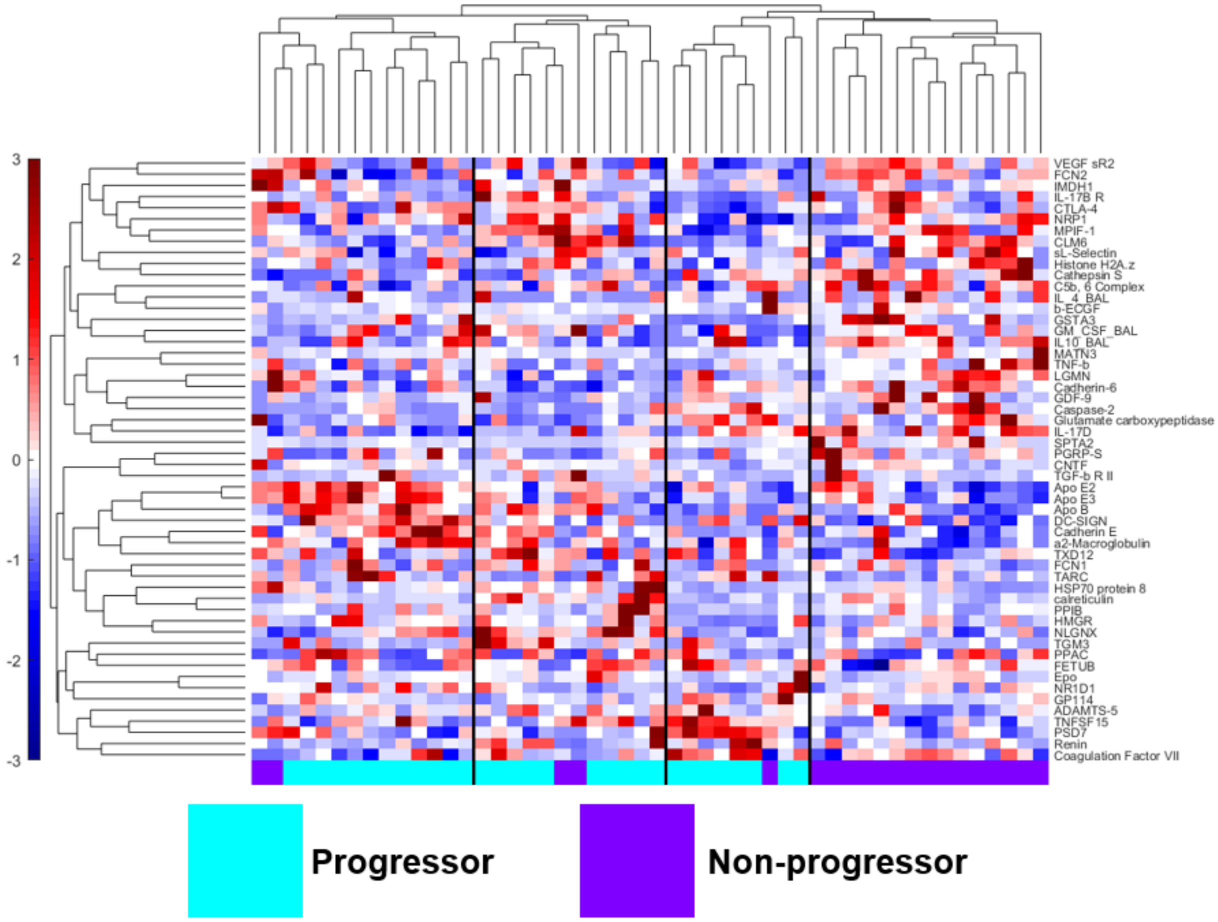


Figure 3.5. Hierarchical clustering of the COMET IPF patients by the LASSO-identified blood and BAL protein signature highlights a single group of non-progressors (purple) and three groups of progressors (cyan) with distinct expression levels of various proteins in the signature.

Only 5 out of the 50 patients were misclassified. Protein expression level is shown in the color scale on the left of the figure, with red indicating higher concentration compared to the mean, and blue lower concentration compared to the mean.

proteins that were also highly expressed in the first group of progressors (apolipoproteins E3 and B, and cadherin E), as well as proteins that were expressed highly in the non-progressor cluster (CTLA-4, MPIF-1/CCL23, and IL-17B receptor). The third group of progressors was characterized by high expression of TNFSF15 (also known as vascular endothelial growth inhibitor) and PSD7 (26S proteasome non-ATPase regulatory subunit 7). The presence of the three progressor groups in the hierarchical cluster may suggest heterogeneity among progressors compared to relative homogeneity among non-progressors, however based on the small sample size in this data it is not possible to determine whether these groups arise from other co-variates and/or random effects. We did evaluate whether any of the progressor clusters could be explained by other clinical and radiological variables collected during the COMET study, including progression metric (e.g. through AE-IPF or a >10% drop in FVC, etc.), smoking status, each participant's genotyping at the MUC5B rs35705950 and the TOLLIP rs5743890 SNPs, and the presence of ground glass and honeycombing in their baseline CT scan. We did not find any apparent clustering by any of these other variables (**Supplemental Figures B.S12A-H**).

3.3.5 Non-progressors exhibit fewer and stronger protein correlations at baseline (week 0) than progressors

Interestingly, when we used correlation networks to explore relationships between proteins in the LASSO-identified signature based on blood and BAL proteins, we found the network based on signature expression levels in progressors had a larger number of overall weaker correlations than the network based on non-progressors. The protein correlation network based on progressors' protein expression (**Figure 3.6A**) contained seven proteins with at least four significant correlations to other proteins. We speculate that the presence of numerous proteins with high numbers of significant correlations (i.e. hub proteins) may suggest a network

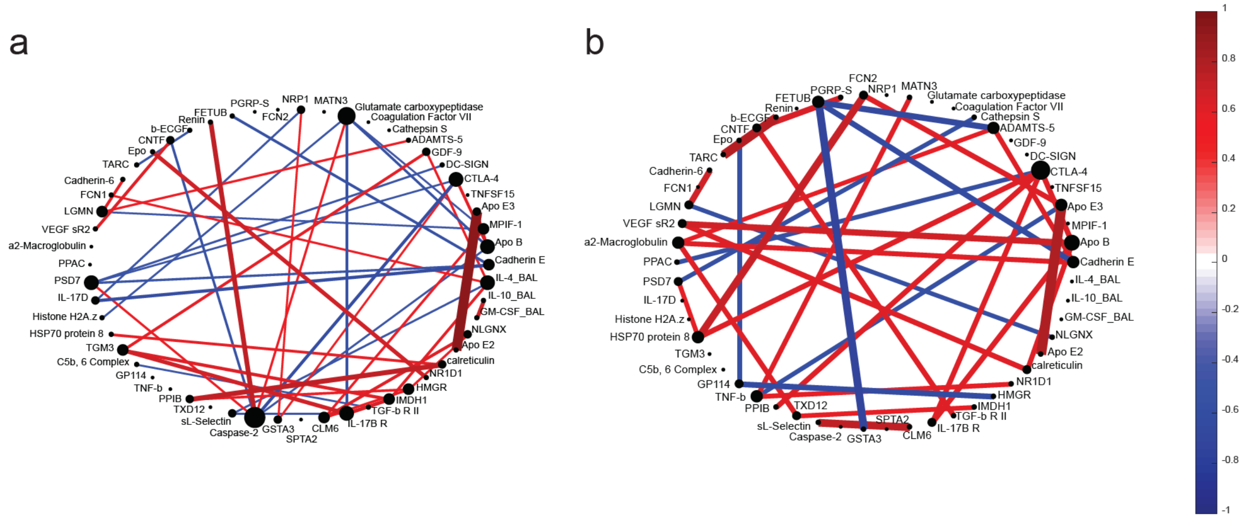


Figure 3.6. Protein correlation networks of the LASSO-identified blood and BAL protein signature present in progressors (a) and non-progressors (b) suggest that non-progressors have a higher degree of control over their proteomic networks than progressors.
A line connecting two proteins indicates the presence of a significant (p<0.05) correlation, as calculated by Pearson's correlation coefficient. Brighter and thicker lines indicate stronger, more significant correlations, respectively. The value of the correlation coefficient for both networks is displayed in the color bar scale on the right, with red indicating a positive relationship and blue a negative relationship. Node size is proportional to degree of connectivity.

with multiple potential drivers, especially when compared to the correlation network based on non-progressors' protein expression (**Figure 3.6B**), which only contained two proteins with four or more significant correlations. Blood caspase-2, CTLA-4, and ApoB, and BAL IL-4 were hub proteins in the progressor network, while blood CTLA-4 and ApoB were the hub proteins in non-progressors. When comparing the two networks, it was clear that there were fewer (45 correlations vs. 33 in the non-progressor network), but significantly stronger (higher absolute value; p = 0.0002, two-sample t-test) correlations present in the non-progressor network.

3.3.6 Trajectory principal component analysis (PCA) identified significant differences in the temporal signature of progressors that were not present in non-progressors

Finally, we found a time-dependent shift in protein expression in progressors that was not present in non-progressors. Our temporal analysis pipeline is illustrated in **Figure 3.1B**. We used LASSO and associated cross-validation to identify signatures that differentiated three time points of blood protein expression (week 0/baseline, week 48, and week 80) within progressors and non-progressors. We then created trajectory principal component analysis (PCA) models[179] based on these signatures to judge temporal separation. The trajectory PCA based on progressor measurements found significant differences in the temporal signature for week 0 and week 80 measurements, with week 48 time points falling in between the other two (**Figure 3.7A**). A one-way ANOVA with Tukey's post hoc test found that week 0 progressor scores on principal component 1 (PC1) were significantly different than scores from week 48 and week 80 ($p < 0.0001$ for both comparisons). We also created a kernel density plot based on the progressor scores on PC1 to further illustrate the differences in the spread of scores between week 0 and week 80 (**Figure 3.7B**). The accompanying loadings plot (**Figure 3.7C**) indicated a relative

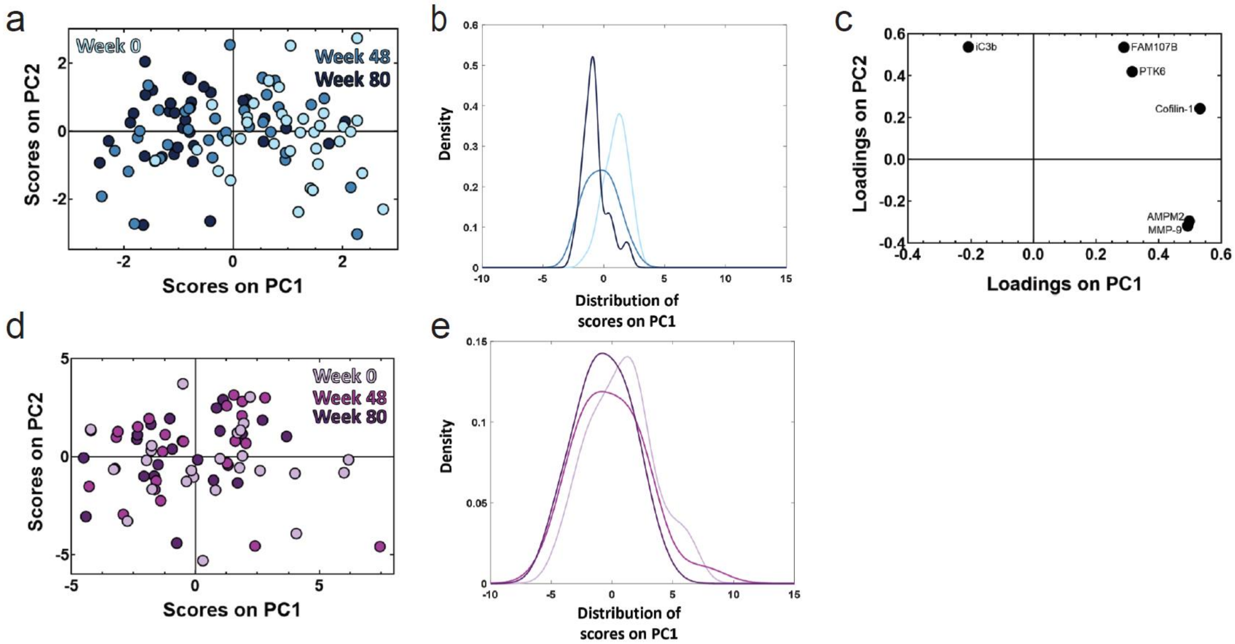


Figure 3.7. Trajectory PCA highlights changes in blood protein expression over time in progressors that is not seen in non-progressors.

(**a**) A trajectory PCA model based on three time points of progressor blood protein measurements highlights the change in protein expression patterns over time in IPF progressors. The week 0 scores on principal component 1 (PC1) were found to be significantly different from both the week 48 scores ($p < 0.001$) and the week 80 scores ($p < 0.001$) by one-way ANOVA with Tukey's post hoc test. The week 48 and week 80 scores were not found to be significantly different from one another by the same test ($p = 0.16$). (**b**) The kernel density plot of the scores on PC1 provides another way of viewing the differences in the scores distribution on PC1 of across all three time points of progressors. (**c**) The LASSO-identified signature separates the three time points of progressor measurements while capturing 49.95% of the natural variance in the data across the first two principal components. (**d**) A trajectory PCA model based on three time points of non-progressor protein measurements does not show clear separation across the three time points. None of the scores on PC1 of the three time points were significantly different from each other after one-way ANOVA with Tukey's post hoc test (all $p > 0.05$). (**e**) The kernel density plot of the scores on PC1 highlights the overlapping of the scores on PC1 from the three time points of non-progressors.

increased expression of inactivated complement C3b (iC3b) compared to matrix metalloproteinase 9 (MMP-9), methionine aminopeptidase 2 (AMPM2), cofilin-1, protein tyrosine kinase 6 (PTK6), and protein FAM107B at week 80, but relative increase of MMP-9, AMPM2, cofilin-1, PTK6, and protein FAM107B compared to iC3b at week 0. In contrast, a trajectory PCA model for non-progressors (**Figure 3.7D**) and a one-way ANOVA with Tukey's post hoc test indicated there were no significant differences in PC1 scores across the three time points ($p > 0.05$ for all comparisons; loadings plot shown in **Supplemental Figure B.S13**). The kernel distribution plot of the non-progressors' scores on PC1 highlights how all three time points are spread out among the same range of scores (**Figure 3.7E**).

3.4 Discussion

In this work we have identified cross-tissue compartment and temporal proteomic signatures that highlight differences between IPF progressors and non-progressors and generated new hypotheses for potential mechanisms of IPF progression. We discovered a multivariate signature based on proteins from the blood and lung tissue compartments that differentiated IPF progressors and non-progressors with 100% cross-validation and calibration accuracy and 100% sensitivity and specificity in a PLSDA model. This signature performed significantly better than analyses based on single proteins and a signature of BAL proteins. Through the use of other computational tools, we found that non-progressors were enriched for regulation of immune regulatory processes, and that the proteome of progressors had significantly weaker and a larger number of correlations than that of non-progressors. Using data from across multiple time points, we were able to identify significant proteomic differences in IPF progressors between week 0 and week 80 measurements that were not present in non-progressors. These results illustrate the value of data-driven modeling approaches for integrating measurements over different tissue compartments and experimental assays, and suggested potential prognostic signatures for progressive IPF for future validation.

The combined use of LASSO with PLSDA allowed us to find small signatures out of hundreds of proteins that were able to accurately differentiate clinical groups of interest. PLSDA and LASSO were able to incorporate data from multiple tissue compartments and assays in the same model to enable a more holistic understanding of IPF progression. The signature of co-varying blood and BAL proteins that we reported has the highest cross-validation and calibration accuracy compared to models based on single proteins, and either outperformed or matched the sensitivity and specificity of previously reported markers of IPF progression. Evaluating signature components allowed for further investigation of potential proteomic relationships and pathways associated with progression. Our identified signature was enriched for processes involving immune system regulation in non-progressors, which echoes results from other studies[109,111,136], and also included 4 of the 6 proteins previously identified in the COMET cohort as an index of IPF progression[136]. The complement cascade has also previously been associated with IPF disease severity[111]. Interestingly, our identified signature did not include MMP-7, which has been linked to IPF progression in several other studies[35,109,114], though some proteins in our signature did have proteolytic function (legumain, PSD7).

There were several limitations associated with this study. While we were able to integrate SOMAmer- and Luminex-based measurements in our models, the SomaLogic platform measured many more proteins than the Luminex platform, potentially biasing results toward blood measurements and toward the functions of the 29 BAL cytokines measured with Luminex. Larger (in the case of BAL proteins) and less directed screens of blood and BAL proteins in future experiments may uncover more unbiased signatures. Another consideration is that aptamer measurements do not always significantly correlate with ELISA concentrations, which could be due to different actions and binding sites of aptamers vs. antibodies. All subjects in the COMET study lived through the study end date, which means that our presented hypotheses might not representative of end-stage IPF patients. Although the model based on both blood and BAL proteins was found to be the most accurate at differentiating IPF progression status, this model would not currently be useful as a prognostic test due to 1) challenges associated with obtaining BAL measurements; and 2) the large number of proteins currently in the signatures. However, because our model is able to investigate covariation in protein expression across tissue compartments, we do believe that the analysis is useful for generating new insight into potential systemic and proteomic relationships associated with IPF progression. The blood protein signature identified here holds more promise as a prognostic signature (cross-validation accuracy of 96% was only moderately lower than the combined model); however, it would still require reduction in the number of proteins before it would be useful. Furthermore, development of a true prognostic signature for clinical use would require validation in new, larger cohorts. To our knowledge there is currently no appropriate validation cohort available, and the SOMAmer platform is no longer accessible for academic use. Therefore, we are unable to confirm the diagnostic or prognostic merit in any of the identified signatures. We did employ cross-validation which suggests that future validation of prognostic biomarkers could be valuable.

We identified signatures in our study to investigate potential mechanistic differences between IPF progressors and non-progressors, and found several emerging trends. A prior knowledge database (DAVID) indicated that significantly enriched processes in non-progressors involved regulation of immune or defense system responses, suggesting that this regulation is potentially lacking or deficient in progressors. We speculate that this idea that non-progressors have better control of proteomic processes was also reflected in the protein correlation networks, where non-progressors had fewer hub proteins and fewer significant correlations present, but these correlations were significantly stronger than those in the progressor network. We hypothesize that this finding indicates a more stable protein network in non-progressors that would be difficult to perturb. Stronger correlations could also indicate that non-progressors have finer control over the expression of these proteins, suggesting that the biological pathways these proteins are involved in are less dysregulated than they are in progressors. Additional experimental analysis would be needed to confirm these ideas.

IPF progressors were characterized by more heterogeneous proteomic expression across tissue compartments. Heterogeneity was suggested by both the correlation network (the large number of significant but weak correlations present in progressors), and also in the hierarchical cluster, which exhibited three progressor clusters that were characterized by unique expression patterns of proteins. We speculate this may suggest potential subgroups (endotypes) are present within the progressors; however, this study did not have the power to eliminate the effects of other co-variates or random influence. One progressor cluster showed increased expression of many apolipoproteins, in addition to DC-SIGN, E-cadherin, ficolin-1, and other proteins. Intriguingly, another cluster of progressors exhibited increased expression of both proteins that were also highly expressed in the non-progressor cluster and proteins that were highly expressed in another progressor cluster. We investigated this group of progressors but did not find a significant difference in the time from COMET enrollment to date of progressive event between this group and the other two groups of progressors identified in the hierarchical cluster. Unsupervised analytical and clustering techniques could be used in other larger studies to better characterize and confirm potential endotypes of IPF progressors.

Intriguingly, proteins from the complement system were signature components in both the temporal-focused and in the tissue compartment analyses. We observed that progressors at later time points (48 or 80 weeks post-baseline) were characterized by comparatively increased expression of iC3b compared to other proteins in the signature. iC3b plays a critical role in pathogen binding and clearance, and also regulates other functions including phagocytosis and IL-12 secretion[180,181]. To our knowledge there have been no studies directly focused on IPF and iC3b, but complement 3 (C3)'s involvement in IPF has been previously studied, with C3 gene expression reported to be higher in the lungs of IPF patients vs. those of healthy controls[182]. Likewise, C3 deficient mice exhibited reduced lung injury after exposure to bleomycin than their wild type counterparts[182], and depletion of the serum complement system inhibited bleomycin-induced lung collagen deposition in rats[183]. Although these studies investigated C3 expression and fibrosis, in our data progressor iC3b expression was positively and significantly correlated with progressor C3 expression over all time points (Pearson's correlation coefficient, $\rho = 0.52$, p-value = $2.1*10^{-8}$), suggesting that changes in iC3b expression levels may reflect similar changes in C3 concentration. Although appearances of iC3b in identified signatures suggest an association with IPF progression, future experimental and clinical studies would be needed to confirm any mechanistic role.

In conclusion, we were able to use systems-focused, data-driven modeling approaches to identify temporal and cross-tissue compartment proteomic signatures that led to increased insight into mechanisms associated with IPF progression. Overall, this work highlighted the ability of quantitative, systems-focused analytical techniques to aid in generating novel hypotheses for proteomic mechanisms associated with IPF progression. We envision these approaches could be easily applied to integrate spatiotemporal data in clinical samples from other diseases that have a progressive and/or heterogeneous patient population.

3.5 Methods

3.5.1 Ethical approval statement

All clinical investigations were conducted according to the Declaration of Helsinki. The human study protocol was approved by the institutional review board of all participating centers and methods were carried out in accordance with the relevant guidelines and regulations (University of California Los Angeles, Los Angeles, CA, United States; University of California, San Francisco, San Francisco, CA, United States; National Jewish Medical and Research Center, Denver, CO, United States; University of Chicago, Chicago, IL, United States; University of Michigan Ann Arbor, MI, United States; Cleveland Clinic Foundation, Cleveland, OH, United States; Temple University, Philadelphia, PA, United States; Brown University, Providence, RI, United States; Vanderbilt University, Nashville, TN, United States).

3.5.2 Subject population

The Correlating Outcomes with Biochemical Measurements to Estimate Time Progression in IPF study (COMET-IPF) (clinical trials ID no. NCT01071707) was a multi-center, prospective observational cohort aimed at identifying markers of IPF progression. All data and samples used in this study were de-identified. The study design has been described previously[137,174], but in brief, eligible patients were aged 35-80 with a multidisciplinary IPF diagnosis (confirmed by clinical history, chest computed tomography (CT) scan, and a lung biopsy when necessary). Subjects with an IPF diagnosis >4 years prior to screening, diagnosed collagen-vascular disorder, FEV1/FVC < 0.60, evidence of active infection at screening, or comorbid conditions likely to result in death within one year were excluded. Informed consent was obtained from all participating patients. Progression during an 80-week follow-up period was dichotomized by the composite occurrence of a relative decline in FVC of ≥10% or in the diffusion capacity of the lungs for carbon monoxide (DLCO) of >15%, acute exacerbation, lung transplant, or death. Seventy-one patients were originally screened for inclusion in the COMET cohort, of which 60 were included in the analysis described here. Patients were excluded from analysis based on a lack of blood samples at all three time points or missing data such as DLCO or 6 minute walk test as described in the original study[137].

3.5.3 Sample acquisition and measurements

Peripheral blood samples were collected from 60 COMET patients at three time points (week 0/baseline, week 48 and week 80). Slow off-rate modified aptamers (SOMAmer©) technology was used to measure 1129 proteins present in blood samples at each collection time point. A small number of blood proteins in fifteen of these samples were later also measured by ELISA; the concentrations of the two platforms were correlated using Pearson's correlation coefficient.

Bronchoscopy was performed at enrollment in patients who were clinically stable and without evidence of active infection. Luminex FlexMAP 3D (Luminex Corporation, Austin, TX) technology was used to measure 29 cytokines/chemokines in the BAL samples. Samples below the lower limit of detection were set to be ½ the lowest minimum detectable concentration across the standard curves of all analytes. Before inclusion in any analyses, all BAL protein concentrations were normalized to total protein concentration as quantified by a Pierce bicinchoninic acid (BCA) Protein Assay Kit (Pierce Protein Biology, Rockford, IL).

For more details on peripheral blood and BAL sample collection, please see **Appendix B**.

3.5.4 Data processing

Before beginning any analysis, a PCA model was created to identify potential negative drivers in the multivariate model. Negative drivers were defined as samples which disproportionally drove the final model such that model parameters solely explained the driver's variance, and were characterized as samples with a Hotelling's Reduced T^2 statistic value > 5. The sample with the highest Hotelling's Reduced T^2 statistic greater than 5 was subsequently removed and another PCA model was generated based on the remaining data. This process was iteratively implemented until all samples produced Hotelling's Reduced T^2 statistics <5, resulting in 4 unique datasets with the following features: (1) baseline blood proteins (59 samples; 34 progressors and 25 non-progressors; demographics detailed in **Supplemental Table B.S1**), (2) BAL proteins (51 samples; 31 progressors and 20 non-progressors; demographics in **Supplemental Table B.S2**), (3) baseline blood and BAL proteins (50 samples; 30 progressors and 20 non-progressors; demographics in **Table 3.1**), (4) temporal-dependent blood proteins for trajectory PCA (102 progressor and 71 non-progressor time point measurements in total). The associated univariate analyses contained the same spread of samples. All proteins, both those measured by SOMAmer aptamers and by Luminex, were measured in both progressors and non-progressors and included in the initial LASSO analysis.

3.5.5 Statistical analysis of differential protein expression in clinical cohorts

Two volcano plots illustrated individual blood and BAL proteins that were significantly and differentially expressed across IPF progressors and non-progressors. Relative fold-changes in blood and BAL protein levels were calculated by dividing the average expression of each protein in progressors by that in non-progressors. Statistical analysis between protein expression in the cohorts was performed by standard two-sample t-tests. P-values < 0.05 were regarded as significant.

3.5.6 Identification of proteomic signatures with feature selection tools and PLSDA

PLSDA was used in conjunction with feature selection tools to determine the protein signature which best differentiated clinical cohorts in various datasets. Prior to any analysis, data were normalized with mean centering and variance scaling. The LASSO was used when finding the minimum signature based on SOMAmer blood protein data. For all LASSO models, k-fold cross-validation (k=10) was used to generate the model with the lowest possible mean squared error for prediction, such that random subsets were iteratively excluded from the data set during model calibration and were later used to evaluate model predictions. VIP scores identified the differentiating signature of BAL proteins, with a VIP cutoff score for inclusion in the model of ≥1. All PLSDA models were built using k-fold cross-validation (k=10) and were orthogonalized to improve interpretability. ROC curves were generated based on the classification ability of a PLSDA model.

3.5.7 Analysis of differentially expressed proteome with DAVID

The Database for Annotation, Visualization, and Integrated Discovery (DAVID) was used to identify significantly enriched biological processes based on the protein signatures identified by multivariate methods. Protein signatures which resulted from these approaches were sorted into profiles based on their relative expression levels in progressor or non-progressor cohorts. The sign of the PLSDA loadings on LV1 determined if the protein was comparatively increased in progressors (negative loadings) or non-progressors (positive loadings). The resulting clustering and enrichment diagrams from DAVID were created by searching through Gene Ontology (GO) biological processes (BP FAT), GO molecular function (MF FAT), and Kyoto Encyclopedia of Genes and Genomes (KEGG). Only the clusters and pathways which were significant after applying the Bonferroni correction within DAVID were reported.

3.5.8 Comparison of PLSDA model performance parameters

In order to quantitatively compare calibration accuracy across multiple PLSDA models, each model of interest was probed to determine whether it correctly or incorrectly classified each individual patient. Patients who were not included in all of the models to be compared were unable to be included in this comparative analysis of calibration accuracy, which only affected the comparison of models based on multiple tissue compartments. A matrix of matched sets of proportions was generated where each patient's classification state (e.g. correctly or incorrectly classified by the model) was represented as dichotomous values for each of the models of interest. These proportions were then compared using Cochran's Q test in conjunction with McNemar's post hoc test; significance was defined as the adjusted p<0.05.

To compare cross-validation accuracy between models, we split the total data into ten groups (5-6 samples in each group) and then iteratively generated PLSDA models based on nine groups of data (training set), and tested the model with the unused group of data (test set). We recorded if these test samples were accurately classified by the model, and compared the percent accuracy from all ten groups associated with one overall PLSDA model to percent accuracy of other PLSDA models. Statistical significance between models was evaluated by a standard one-way ANOVA with Tukey's post hoc test. P-values <0.05 were deemed significant.

3.5.9 Visualization of classification ability of LASSO-identified signature using clustering

Hierarchical clustering of the LASSO-identified signature based on blood and BAL proteins was generated with supervised average linkage clustering. Pearson's correlation coefficient was used as the distance metric. Samples were colored by progression status as well as other clinical, radiologic, and genetic variables.

3.5.10 Exploration of network interactions between progressor and non-progressor cohorts

Protein correlation networks were constructed separately for progressors and non-progressors using pairwise Pearson's correlation coefficients between protein expression in the LASSO-identified signature within the two groups. Edge color and thickness correspond to coefficient value and statistical significance, respectively, with only significant correlations ($p < 0.05$) being shown. Node size is proportional to its degree of connectedness.

3.5.11 Investigating temporal dependences in progressor/non-progressor protein signatures

LASSO identified the minimum blood signature that differentiated the three collection time points (week 0, 48 and 80) in progressors and non-progressors separately. Trajectory PCA models[179] were then created based on each of these signatures. A one-way ANOVA with Tukey's post hoc test was used to evaluate the significance of temporal differences in protein expression by comparing the scores on PC1 at each collection time point. P-values < 0.05 were considered significant.

3.5.12 Visualization of time-dependent scores with density plots

PC1 scores from each of the three time points in the trajectory PCA were fit to a kernel distribution. The kernel distribution was reconstructed into a probability density function using the fitdist function with the normal smoothing function and the default bandwidth value.

3.5.13 Software summary

All volcano plots, hierarchical clustering, heat maps, correlation networks, and density plots were completed using Matlab (v2016b, Matlab, Natick, MA). LASSO was implemented using Matlab software[167]. PCA and PLSDA models, ROC curves, and VIP score calculations were generated using the PLS toolbox available in Matlab® (v8.2.1, Eigenvector, Mason, WA). All statistics, with the exception of Cochran's Q test, were performed using Prism version 7.00

and version 8.00 (GraphPad software, San Diego, CA). Cochran’s Q test with McNemar’s post hoc test was done in R software version 3.5.1 (R Core Team, Vienna, Austria).

Chapter 4 Unpublished IPF Results

Contributions. *The COMET investigators, and Jeff Curtis and Christine Freeman were involved in collecting the BAL samples from the IPF patients and the healthy patients. Vibha Lama stored the COMET IPF BAL samples, and Drs. Curtis and Freeman stored many of the healthy BAL samples.*

4.1 Introduction

In addition to the published work presented in **Chapters 2** and **3**, we have also generated additional models of IPF disease state and progression that have not been published. Key results in this chapter focus on analysis of BAL proteins, as well as the application of another data analysis tool to the IPF blood and lung protein data. The goal of this work was to identify proteomic signatures that could differentiate clinical groups and help us gain insight into processes involved in IPF disease state and progression.

Although IPF is a disease that is localized to the lung, collecting lung tissue biopsies that could allow for deeper insight into pathways associated with disease state or progression can be dangerous due to the potential injury or exacerbation events that could result from sample collection. Another technique for the collection of samples that describe the pulmonary environment in a less injury-inducing fashion is the bronchoalveolar lavage (BAL) procedure. This procedure involves the injection of sterile saline into the lung followed by immediate collection, which provides a sample of the epithelial lining fluid (ELF), secreted cytokines, and cell types present inside the lung. Although this procedure still requires entry into the lung environment through the trachea, there is lower risk for potential tissue damage when performed correctly for the BAL procedure than the surgical lung biopsy.

However, the BAL sample collection process is variable, which has resulted in very few reports of BAL protein biomarkers for IPF. The quantitation issues associated with the BAL procedure are caused by variations in lung structure across patients that affect the amount of saline that is recollected after the flush; the unknown quantity of ELF that is collected with each saline flush; and potential contamination from bronchi-level lung cells[184]. Due to these factors, it has been difficult to deal with the unknown ELF dilution factor and identify proteins that can separate groups of interest, even though these samples come from the tissue compartment of injury. There has been some success using BAL samples in IPF: previously, IL-33 and thymic stromal lymphopoietin (TSP) concentrations in BAL were reported to be able to differentiate IPF from other interstitial lung diseases[185]. Additionally, monocyte chemoattractant protein 1 (MCP-1), thymus- and activation-regulated chemokine (TARC/CCL17), and macrophage-derived chemokine (MDC) have been reported to be associated with poor outcomes in IPF[186]. However, there have been others who have reviewed results of studies focused on BAL samples that have concluded that measurements from BAL samples alone are not enough to diagnose patients with ILDs[187–189].

A potential reason why there are so few BAL proteins associated with disease state or progression could involve how both the BAL samples and the resulting proteomic data have been approached. There have been few studies where greater than ten BAL proteins were measured in each sample for IPF disease state or progression investigations[126,190–193], although this seems to be changing as multiplex protein assays become more common and attainable. Additionally, the approaches taken when analyzing BAL data often involved looking at proteins one at a time[185,186,190,191,193–196], with only few studies considering how covariation or networks might lead to increased differentiation ability or increased biological insight[126,192]. The design of past experiments in this way could have potentially contributed to the lack of proteomic network-level inferences made about lung processes involved in IPF disease state and progression. Based on this underutilization of BAL samples in the literature, we have applied our data-driven modeling techniques to identify signatures of BAL proteins that were able to differentiate IPF disease state and disease progression, and additionally created more cross-tissue compartment models using different classification algorithms, all of which led to increased insight into the potential role cytokines may play in IPF progression.

4.2 Results

4.2.1 BAL signature identified that differentiates healthy and IPF patients

Two-sample, two-tailed t-tests were used to identify 3 proteins out of the 29 measured proteins which were significantly differentially expressed across the healthy and IPF populations: interferon α2 (IFNα2), IL-7, and IL-15 ($p = 0.00012$, $p = 2.66*10^{-11}$, and $p = 0.034$, respectively). Variable Importance in Projection (VIP) scores selected and partial least squares discriminant analysis (PLSDA) visualized a signature of 4 out of 29 measured BAL proteins that differentiated healthy ($n = 5$) and IPF ($n = 51$) patients with 97.06% cross-validation and calibration accuracy (**Figure 4.1A**). Latent variable 1 (LV1) differentiated healthy (purple; more negative scores on LV1) from IPF patients (cyan; more positive scores on LV1) (**Figure 4.1B**). We then compared the calibration and cross-validation accuracy of the VIP-selected model to that of PLSDA analyses based on single differentially expressed proteins, and a PLSDA model based on all three of the significant proteins discovered. We saw that the VIP-selected model had

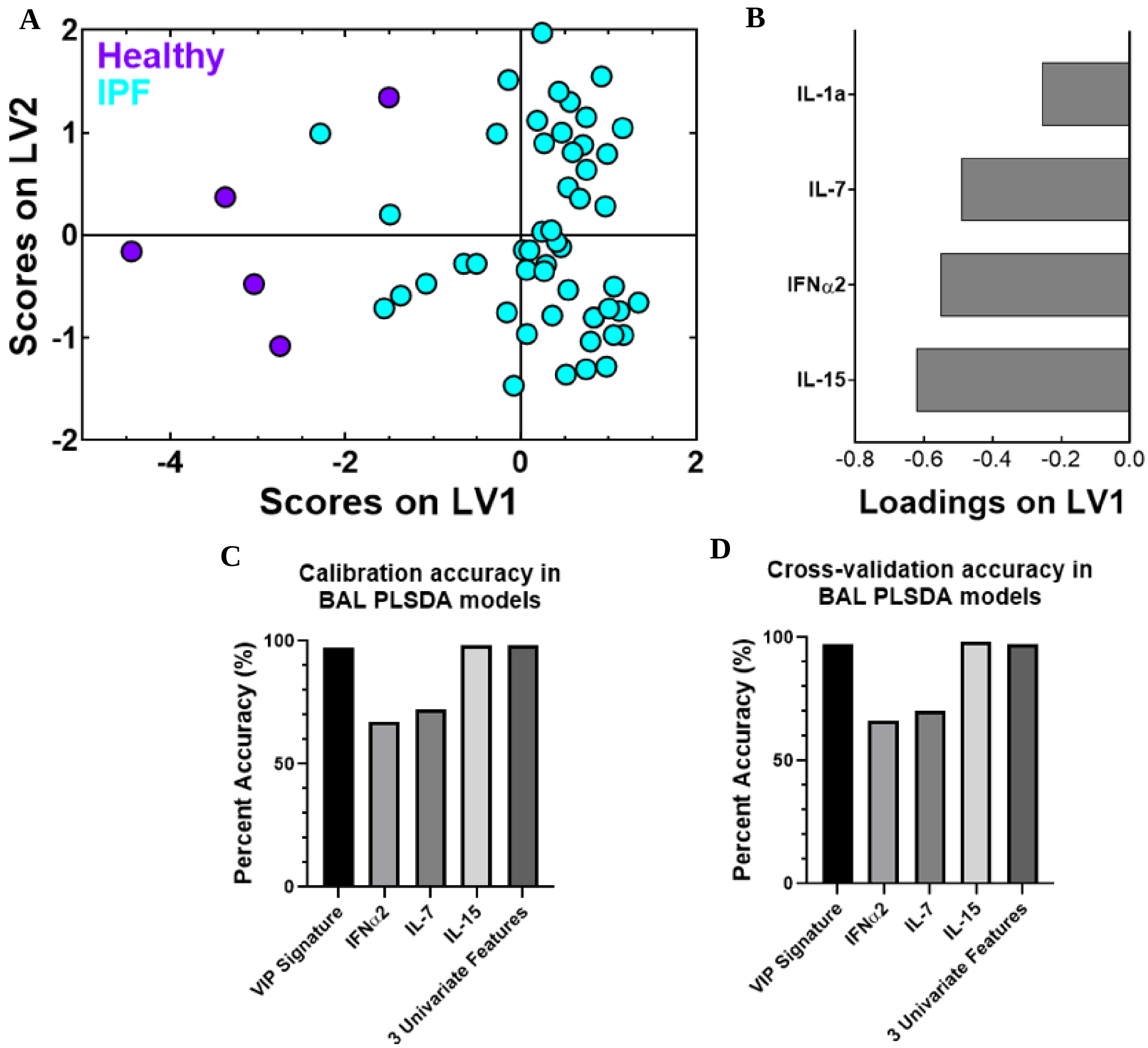


Figure 4.1. VIP-selected signature of BAL proteins classifies IPF disease state better than or just as good as single proteins.
(**A**) The PLSDA scores plot separated healthy (purple) and IPF (cyan) subjects with 97.06% calibration and cross-validation accuracy. (**B**) The loadings on latent variable 1 (LV1) captured 50.56% of the variance in the data. Proteins loaded negatively on LV1 are comparatively decreased in IPF. Comparisons of calibration (**C**) and cross-validation (**D**) accuracies associated with models and analyses based on univariate-identified proteins shows that the VIP signature is better than or nearly as good as these models.

much higher calibration (**Figure 4.1C**) and cross-validation (**Figure 4.1D**) accuracy than analyses based on IFNα2 or IL-7, but tended to be slightly (1%) worse than analyses based on IL-15 alone and based on all three significant proteins.

4.2.2 Lung signature that differentiates IPF progression status highlights importance of chemokines

Due to the low number of proteins that were significantly differentially expressed across the progressors and non-progressors (**Figure 3.2b**), we then turned to data-driven modeling techniques to identify signatures of covarying proteins that could differentiate the two groups. As discussed in **Chapter 3**, we used VIP scores to select a signature of 12 BAL cytokines that differentiated IPF progressors (n = 31) and non-progressors (n = 20) with 78.55% calibration and 67.82% cross-validation accuracy in a PLSDA model (**Supplemental Figure B.S3a** and **B.S3b**)[197]. The cytokine data included in this model were first normalized to protein albumin levels in the BAL samples using a bicinchoninic acid (BCA) assay (BCA-normalized BAL cytokines). This BCA-normalized model performed better than a model based on a signature of non-normalized BAL proteins, but performed significantly worse than models based on blood proteins alone and a model based on blood and BAL proteins combined (**Chapter 3**). We hypothesized that this might be due in part to the small number of BAL proteins measured compared to blood proteins, and that a targeted panel of BAL proteins (cytokines and chemokines) were measured compared to the less directed panel of blood proteins measured.

While better characterization of potential cross-tissue compartment proteomic relationships can help create a holistic understanding of IPF progression, we also wanted to focus on gaining insight into the potential lung mechanisms associated with progression because the lungs are the main tissue compartment of injury in IPF. We created correlation networks based on the expression of the VIP-selected BAL protein signature in progressors (**Figure 4.2A**) and non-progressors (**Figure 4.2B**). In the progressor network, the proteins with the highest number of significant correlations to other proteins included MCP-1, IL-8, granulocyte colony stimulating factor (G-CSF), granulocyte-macrophage colony stimulating factor (GM-CSF), and epidermal growth factor (EGF).

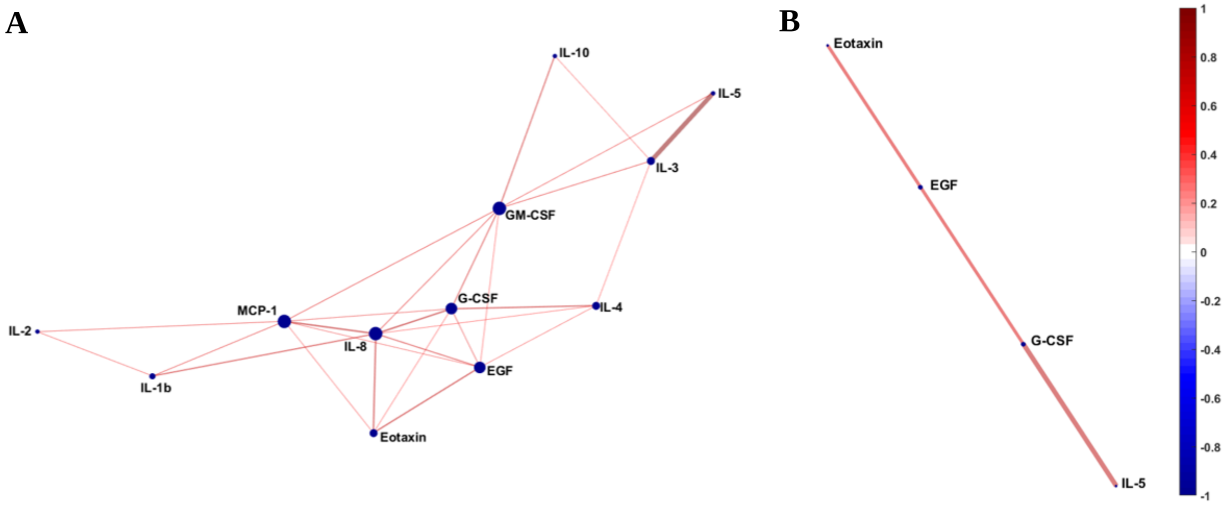


Figure 4.2. Correlation network of the VIP-selected BAL protein signature present in (A) progressors and (B) non-progressors.

Proteins connected by two lines are significant ($p < 0.05$) correlated by Pearson's correlation coefficient. Node size reflects the number of significant correlations to other proteins. Brighter and thicker lines indicate a stronger, more significant correlation, respectively. The value of the correlation coefficient for both networks is displayed in the color bar scale on the right, with red indicating a positive relationship and blue a negative relationship.

4.2.3 Gradients of proteins were not able to differentiate progression status better than single cytokines

We next investigated if the differential expression of proteins across the blood and the lungs could lead us to a deeper understanding of IPF progression. Standardized gradients have been successfully identified as being positively associated with higher disease risk in HIV, another immunological disease affecting a mucosal surface[103]. We applied decision tree analysis (DTA) to IPF to identify the gradient relationships that were best in differentiating progressors and non-progressors and the hierarchy of importance of these relationships in classifying the two groups. To calculate the gradients, we logarithmically transformed the raw blood and BAL protein data separately before standardizing the data by setting the mean of each protein to be zero and the standard deviation to be one. After that, we took these values and subtracted BAL protein – blood protein to find the gradient. This means that when interpreting the gradient

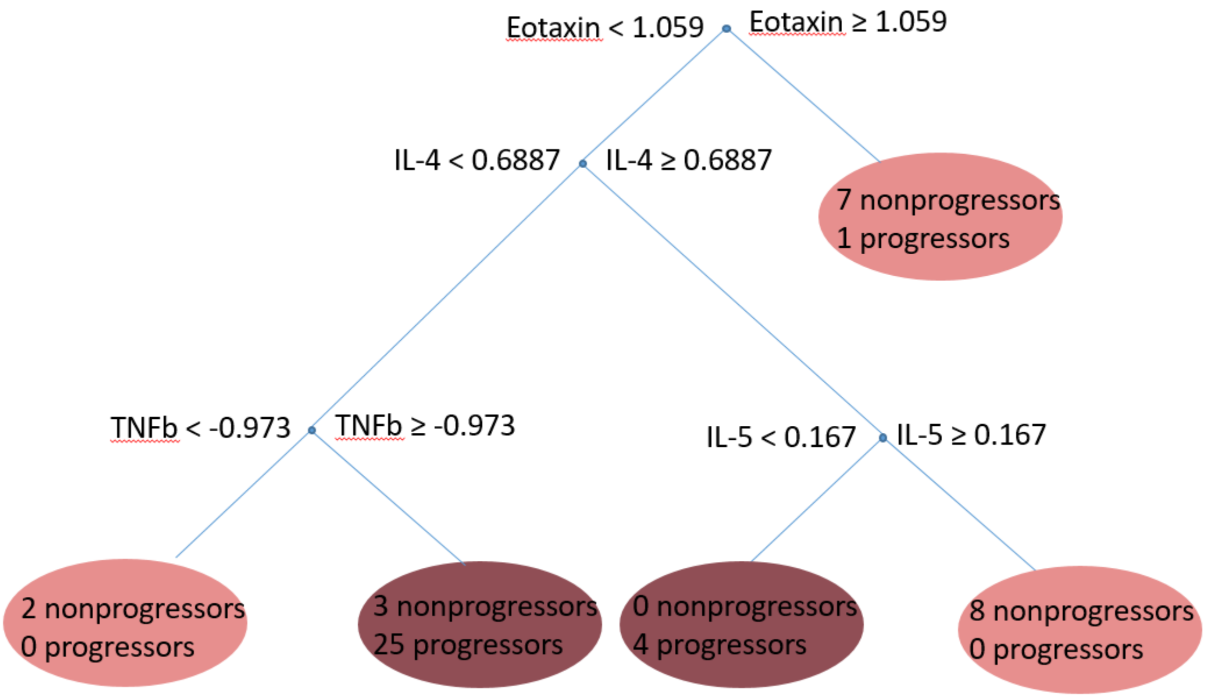


Figure 4.3. Decision tree analysis based on gradients of protein concentrations across tissue compartments highlights hierarchical importance of gradient concentrations in differentiating progressors (P) and non-progressors (NP), with eotaxin being the most hierarchically important blood-lung gradient involved in differentiating the two groups, followed by IL-4.
All gradients were calculated by logarithmically transforming the protein expression data, normalizing each protein to have a mean of zero and a standard deviation of one, and then subtracting BAL protein – blood proteins. A positive gradient is indicative of a higher concentration of the protein in the BAL sample.

values, a positive gradient value indicates a protein is higher in DTA highlighted eotaxin, followed by IL-4, as being the two most hierarchically important gradients involved in classifying IPF progressors and non-progressors (**Figure 4.3**). This model separated progressors and non-progressors with 76% cross-validation and 92% calibration accuracy. Interestingly, the majority of all progressors were found in one leaf, described by having a lower BAL eotaxin concentration compared to plasma, a lower BAL IL-4 concentration compared to plasma, and a higher BAL TNF-β concentration compared to plasma. Although we found that the DTA model based on the gradient concentration across the lung and blood tissue compartments nominally outperformed DTA models based on only blood or only BAL protein expression, the calibration and cross-validation accuracies of these models were very close to each other and were not significantly different (data not shown).

4.3 Discussion

In this work, we identified signatures of cytokines measured in BAL samples that were able to differentiate healthy and IPF patients, as well as IPF progressors and non-progressors. We were successfully able to detect and measure the concentrations of cytokines in BAL samples collected from IPF patients by doubling the volume used in our Luminex assay. We hypothesized that lung chemokines may play a role in IPF progression, and that these chemokines recruited cell types that suggested that multiple mechanisms of tissue reorganization may be at play in progression. These results illustrate the value of coupling Luminex measurements with data-driven modeling techniques in order to gain increased insight into proteins and potential mechanisms associated with IPF disease state and progression.

To our knowledge, this was the first time that a signature of BAL cytokines had been identified that could differentiate IPF from healthy patients. This model outperformed all analyses based on differentially expressed proteins except for models based on IL-15, for which our VIP-selected signature was less accurate than by only 1%. We hypothesize this occurred because IL-15 was highly significantly different across the healthy and IPF groups ($p = 2.66*10^{-11}$, two-sample t-test). Overall, using signatures of lung proteins to differentiate disease state in IPF may be able to serve as a complementary tool and confirm co-variation between proteins that were already identified in univariate analysis, which could allow us to gain increased insight into the pulmonary environment of IPF.

We hypothesized that chemokines are important in IPF progression from our results of protein correlation coefficient networks based on the BAL protein signature that differentiated

IPF progressors and non-progressors in **Chapter 3**[197]. In the progressor correlation network based on the VIP-selected BAL protein signature, the hub proteins included MCP-1, IL-8, GM-CSF, G-CSF, and EGF. These cytokines attract and support the growth of neutrophils (GM-CSF and IL-8), are chemoattractive for and stimulate the growth of monocytes (MCP-1), and increase fibronectin secretion in IPF fibroblasts (EGF). The interactions between these hub proteins and cell types are intriguing given current hypotheses surrounding IPF pathogenesis and progression: when recruited to the lung tissue, monocytes secrete pro-fibrotic inflammatory cytokines[16] and can differentiate into macrophages[198], which are associated with IPF pathogenesis[199]; neutrophils may be involved in regulating lung fibrosis levels through their role in ECM regulation via secretion of neutrophil elastase and in balancing levels of matrix metalloproteinases (MMPs) and tissue inhibitors of metalloproteinases (particularly MMP-8[200]), although their exact contribution to IPF fibrosis remain unclear[201]; EGF has been shown to cause IPF fibroblasts to secrete increased levels of fibronectin[202]. Taking these functions and the correlation network together, this suggests that progressors may undergo tissue reorganization through multiple pathways, and that each pathway is potentially affected by each other. Follow-up on these results with in vivo models of fibrosis will be key to see if these mechanisms are affected by each other, and if all of them are associated with fibrosis.

When applying decision tree analysis to the IPF progression data, we saw that gradients of cytokines were only slightly better at differentiating IPF progressors and non-progressors compared to expression data from single tissue compartments alone. We found the IL-4 gradient to be somewhat surprising, as it would be expected based on the literature that alveolar macrophages secrete higher levels of IL-4 than compared to smokers and controls[203], and that IL-4 is increased in the BAL of IPF patients compared to controls[204]. This DTA result could be due to the preprocessing that had to be performed before the creation of the DTA model because the blood and BAL proteins were measured using different platforms – the aptamer-based SomaLogic platform that reported blood protein concentrations in relative fluorescence units (RFUs), and the albumin-normalized antibody-based Luminex platform that reported BAL protein concentrations in pg protein/mg albumin. Overall, this result does not suggest that gradients of cytokines across the blood and lung tissue compartments are significantly better at differentiating IPF progression status.

Limitations associated with this work come from the small sample size (especially in the case of the healthy patients), the nature of the COMET cohort, and the variability associated with obtaining BAL measurements. Healthy BAL samples are difficult to come by due to the invasiveness of the procedure, which is why so few healthy samples were included in this model. The low number of healthy samples is the reason why we did not perform any follow-up analyses on the signature, as the 5 healthy patients we were able to include may not be a complete representation of the healthy population at large. As stated in **Chapter 3**, all COMET IPF subjects lived through the end of the study, so hypotheses presented here may only apply to mild- to moderate-IPF and not end-stage IPF. We did not have access to new samples for model validation, but we did cross-validate our models whenever possible. Lastly, the BAL sample collection procedure is a variable process. We have done our best to account for this variability by normalizing protein concentrations measured by Luminex to the total protein albumin concentration measured by the BCA assay in each sample. We found that our models based on albumin-normalized protein concentrations performed better than models based on non-normalized protein concentrations (data not shown), but it should be mentioned that there is no consensus as to which BAL normalization technique best reflects the physiological concentration of the proteins in the lung lining fluid.

4.4 Methods

4.4.1 Human sample collection and protein measurements

IPF BAL samples were collected from patients enrolled in the Correlating Outcomes with biochemical Markers to Estimate Time-progression in IPF (COMET) study (clinicaltrials.gov, clinical trials ID no. NCT01071707). Although the COMET study recruited 60 IPF patients, only 51 IPF BAL samples were available when measuring protein concentrations (20 non-progressors and 31 progressors). Inclusion criteria and the definition of disease progression employed in this study have previously been described[136,174]. Informed consent was obtained from all participating centers, which included University of California Los Angeles. Los Angeles, CA, United States–University of California, San Francisco. San Francisco, CA, United States–National Jewish Medical and Research Center, Denver, CO, United States–University of Chicago, Chicago, IL, United States–University of Michigan Ann Arbor, MI, United States–Cleveland Clinic Foundation, Cleveland, OH, United States–Temple University, Philadelphia, PA, United States–Brown University, Providence, RI, United States–Vanderbilt University, Nashville, TN, United States. The study protocol was approved by the institutional review board of all participating centers and methods were carried out in accordance with the relevant guidelines and regulations. Bronchoscopy was performed at enrollment in patients who were healthy enough to undergo the procedure. BAL samples were collected and pooled from 4 installations of 50 mL sterile isotonic saline aliquots. Cell-free fluid was stored at -80°C.

Four healthy BAL samples were collected at the Veteran’s Association Ann Arbor Healthcare System (VAAAHS), with the collection protocol approved by internal review boards

(IRBs) at the VAAAHS and at the University of Michigan Health System (UMHS). BAL was performed through 5 installations of 30 mL of sterile saline into each side of the lung, with all installations being then pooled at the end. The fifth healthy BAL sample was also collected at the UMHS. For this sample, 2-3 installations of 60 mL of sterile saline were flushed into the right lung, and all installations were later pooled. For all five healthy BAL samples, cell-free fluid was stored at -80°C until protein measurement occurred.

All BAL samples were then collected and Luminex FlexMAP 3D technology (Luminex Corporation, Austin, TX) was used to measure 29 cytokines/chemokines in all BAL samples. For protein measurements in Luminex, we used a protocol that used ¼ of the recommended number of beads and sample to minimize bead and sample volume for the assay, which was inspired by Arnold et al.[205]. Due to low cytokine concentrations present in BAL samples[206], we also ran BAL samples at 2X the normal volume for this protocol, which was 30 µL per well. Samples were run in duplicate, and those that were below the lower limit of detection were set to be ½ the lowest minimum detectable concentration across the standard curves of all analytes. Before inclusion in any analyses, all BAL protein concentrations were normalized to total protein concentration as quantified by a Pierce BCA Protein Assay Kit (Pierce Protein Biology, Rockford, IL).

4.4.2 Quantitative modeling approaches

The first step in the data-driven analysis was to determine if any samples negatively drove the creation of the data-driven models. All data were normalized by mean centering and variance scaling before any PCA models were built. Negative drivers were samples which disproportionally drove the final models of disease state or of disease progression such that model parameters solely explained the driver's variance, and were characterized as samples with a Hotelling's Reduced T^2 statistic value > 5. The sample with the highest Hotelling's Reduced T^2

statistic that was greater than 5 were subsequently removed and another PCA model was generated based on the remaining data. This process was iteratively implemented until all samples produced Hotelling's Reduced T^2 statistics < 5.

Once all negative drivers had been identified, we used PLSDA in conjunction with VIP scores to determine the protein signatures that best differentiated the healthy and IPF patients, and IPF progressors and non-progressors. Proteins that had a VIP score ≥ 1 were said to be important, and another PLSDA model was then built based only on the VIP-selected features. All data were normalized by mean centering and variance scaling before any PLSDA models were built. All PLSDA models were built using K-fold cross-validation (k = 10), and models were orthogonalized after VIP-selection to improve interpretability. The model of BAL proteins that differentiated healthy and IPF patients is discussed in depth in **Chapter 3**[197].

Protein correlation coefficient networks were constructed using pairwise Pearson's correlation coefficient based on expression of the BAL proteins in the VIP-selected signature in progressors and non-progressors separately. A brighter and thicker line connecting two protein nodes indicates a stronger and more significant correlation, respectively, with only significant ($p < 0.05$) correlations being shown. Node size is proportional to its degree of connectedness.

Cytokine gradients were calculated by first log10 transforming the 23 proteins that were measured both by Luminex technology in the BCA normalized BAL samples and by SOMAmers in the blood samples. The log10 transformed values were then standardized, and the gradient was calculated by subtracting blood values from BAL values such that a positive gradient indicated higher concentration in the BAL. A classification decision tree algorithm predicted the hierarchy of importance in gradient or raw concentration from single tissue compartments that were best at differentiating IPF progressors and non-progressors, with Gini Diversity Index being used as the split criterion. Each tree was cross-validated using k-fold cross-validation with 10 folds. Trees were pruned to the level that exhibited the lowest calibration and cross-validation error.

A two-tailed, two-sample t-test was used to determine significant differences in expression across the healthy and IPF groups. All quantitative models, decision trees, and statistical analyses were created using Matlab (v2016b, Matlab, Natick, MA). PCA, PLSDA, and VIP scores were calculated using the PLS toolbox available in Matlab (v8.2.1, Eigenvector, Mason, WA).

Chapter 5 Inference of Cellular Immune Environments in Sputum and Peripheral Blood Associated with Acute Exacerbations of COPD

Katy C. Norman[1] *, Christine M. Freeman[2, 3, 4] *, Neha S. Bidthanapally[1], MeiLan K. Han[2], Fernando J. Martinez[5], Jeffrey L. Curtis[2, 3, 6 #], and Kelly B. Arnold[1 #]

*co-first authors, #co-corresponding authors

[1] Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI 48109

[2] Department of Internal Medicine, Division of Pulmonary & Critical Care, University of Michigan, Ann Arbor, MI 48109

[3] Research Service, VA Ann Arbor Healthcare System, Ann Arbor, MI 48105

[4] Graduate Program in Immunology, Rackham Graduate School, University of Michigan, Ann Arbor, MI 48109

[5] Joan & Sanford I. Weill Department of Medicine, Division of Pulmonary & Critical Care Medicine, Weill Cornell Medical College, New York, NY 10065

[6] Medicine Service, Pulmonary & Critical Care Section, VA Ann Arbor Healthcare System, Ann Arbor, MI 48105

Cellular and Molecular Bioengineering. 2019; 12(2): 165-177.

5.1 Abstract

Introduction—Chronic obstructive pulmonary disease (COPD) is the fourth leading cause of death in the United States, with high associated costs. Most of the cost burden results from acute exacerbations of COPD (AE-COPD), events associated with heightened symptoms and mortality. Cellular mechanisms underlying AE-COPD are poorly understood, likely because they arise from dysregulation of complex immune networks across multiple tissue compartments.

Methods—To gain systems-level insight into cellular environments relevant to exacerbation, we applied data-driven modeling approaches to measurements of immune factors (cytokines and flow cytometry) measured previously in two different human tissue environments (sputum and peripheral blood) during the stable and exacerbated state.

Results—Using partial least squares discriminant analysis (PLSDA), we identified a unique signature of cytokines in serum that differentiated stable and AE-COPD better than individual measurements. Furthermore, we found that models integrating data across tissue compartments (serum and sputum) trended towards being more accurate. The resulting paracrine signature defining AE-COPD events combined elevations of proteins associated with cell adhesion (sVCAM-1, sICAM-1) and increased levels of neutrophils and dendritic cells in blood with elevated chemoattractants (IP-10 and MCP-2) in sputum.

Conclusions—Our results supported a new hypothesis that AE-COPD is driven by immune cell trafficking into the lung, which requires expression of cell adhesion molecules and raised levels of innate immune cells in blood, with parallel upregulated expression of specific chemokines in pulmonary tissue. Overall, this work serves as a proof-of-concept for using data-driven modeling approaches to generate new insights into cellular processes involved in complex pulmonary diseases.

5.2 Introduction

Chronic obstructive pulmonary disease (COPD) is a progressive and heterogeneous lung disease that is the fourth leading cause of death in the United States[207], with yearly U.S. medical costs expected to increase to nearly $50 billion in 2020[57]. A large portion of these costs is attributed to acute exacerbations of COPD (AE-COPD), characterized by increased symptoms (dyspnea, coughing, sputum production, and fatigue) beyond day-to-day variation that require treatment with antibiotics or corticosteroids[208]. Severe exacerbations (that require hospitalization) have an in-hospital all-cause mortality rate of 5-7%[71,72], and account for most of the financial burden of COPD[209]. Accordingly, the prediction and treatment of AE-COPD events are top priorities.

Nonetheless, pathogenic cellular mechanisms underpinning AE-COPD are largely undefined. Local tissue and systemic inflammatory pathways are hallmarks of COPD, and are further increased during AE-COPD. Most AE-COPD are also associated with evidence of viral or bacterial infections or both[92,210,211], with upregulation of IL-8, TNF-α and reactive oxygen species in cells and tissue environments[211]. Some AE-COPD are also highly eosinophilic[92]. COPD patients with persistent systemic inflammation have higher mortality and exacerbation rates compared to non-inflamed patients[212]. AE-COPD frequency is reduced by several types of therapies, including inhaled corticosteroids (ICS), long-acting muscarinic antagonists, scheduled azithromycin, and roflumilast[81,213–215]. The success of these treatments, which share immunomodulatory effects, support acutely increased inflammation as contributing to AE-COPD, though fundamental mechanisms driving AE-COPD remain elusive.

Despite identification of individual cell types and cytokines that are differentially expressed between stable and exacerbated COPD[216–218], no single factor entirely accounts for

AE-COPD, and therapies based on single targets have been unsuccessful. In the past 25 years, only one new class of medicine has been accepted for COPD treatment[219]. Plasma fibrinogen was recently qualified by the Food and Drug Administration as a prognostic biomarker, but only for subject enrichment in clinical trials of exacerbation and mortality[96]. Both serum C-reactive protein (CRP)[89,220] and IL-6[220,221] are upregulated in the secreted systemic environment during AE-COPD, but CRP alone is insufficiently sensitive as an AE-COPD biomarker[222], and IL-6 elevations are inconsistently associated with exacerbations[223]. New approaches to understanding cellular mechanisms underpinning AE-COPD pathogenesis are clearly required.

As COPD is a complex condition exhibiting evidence of immunological involvement[224,225], it is plausible that AE-COPD events result from disrupted networks of immune cells and cytokine communication, rather than from individual mediators. Data-driven modeling approaches offer the opportunity to infer these systems-level relationships by identifying small signatures of proteins or other cellular immune factors that co-vary with each other and are associated with disease state. These signatures can then be linked to mechanisms or cell types involved in phenotypes or pathogenic states, providing insight into specific disease biology and potential targets for follow-up experiments and therapeutic intervention. Partial Least Squares Discriminant Analysis (PLSDA) is a useful tool for highlighting covariance among variables that best classify groups of interest, which could lead to the identification of potential proteomic and cellular networks associated with AE-COPD. We have previously illustrated that PLSDA is able to identify and aid in visualizing biologically relevant proteomic and cellular signatures that may give insight into inflammatory pathways. We have used it to evaluate inflammatory signatures in the female reproductive tract mucosa[102] and the blood of interstitial pulmonary fibrosis (IPF) patients[174], in both cases identifying new biomarkers and generating novel insight into key cellular mechanisms.

In this study we apply data-driven modeling approaches to gain insight into the proteomic networks and cellular mechanisms in blood and lung environments that underpin AE-COPD using a prospective cohort study[91], which collected paired sputum and peripheral blood samples from COPD subjects when clinically stable and again before treatment for an AE-COPD. We show that data-driven modeling approaches are able to 1) identify cytokine networks that may be better for classifying AE-COPD than individual cytokines, 2) determine key relationships between cytokines in different tissue compartments, and 3) integrate information measured in different assays to provide a more complete picture of pathogenic processes involved in AE-COPD.

5.3 Methods

5.3.1 Study design, ethics and subject populations

All samples and data in this analysis derived from a published prospective observational trial (ClinicalTrials.gov NCT00281216)[91], which followed subjects at increased risk of AE-COPD for up to three years. Patients were recruited at the VA Ann Arbor Healthcare System (VAAAHS) and the University of Michigan Health System (UMHS). All parts of the study adhered to the Declaration of Helsinki and obtained approval of each site’s Institutional Review Board, with all subjects giving written consent to the study before any procedures occurred. At enrollment and quarterly, participants underwent spirometry, pulmonologist clinical evaluations, collection of peripheral blood and spontaneously expectorated sputum, and a post visit questionnaire. An exacerbation of COPD was said to occur if the subject reported an increase in dyspnea, cough or sputum production, and if the study physician ordered antibiotics or oral steroid for the patient after a physical examination and chest radiographs to rule out pneumonia. Only if a diagnosis of AE-COPD was made were sputum and peripheral blood samples collected at these unscheduled visits. After all data and sample collection occurred, then each subject began treatment for AE-COPD.

5.3.2 Sample collection, processing, and measurements

Peripheral blood was used for both leukocyte immunophenotyping and to measure 40 analytes in serum, which was stored at -80°C until analysis. Spontaneously expectorated sputum was immediately processed in a 9:1 mixture of distilled water to Sputolysin® (EMD Millipore, Billercia, MA) as described[91], and the resulting supernatant was stored at -80°C until used to measure 36 analytes. Serum and sputum samples were unfrozen and protein concentrations were measured simultaneously either using a Luminex 200 System® (Luminex Corporation, Austin TX) or ELISA (GDF-15, IL-18, IL-23p19 and IFN-β)[91].

Whole blood was stained with directly conjugated monoclonal antibodies on the day of the visit as described in the text and supplemental information of Freeman et al.[91] Cells were analyzed using a LSR II flow cytometer (BD Bioscience, San Jose, CA) as reported in McCubbrey et al.[226], using FACSDiva software (BD Biosciences) data with automatic compensation and FlowJo software (Tree Star, Ashland, OR).

5.3.3 Data processing and systems analysis

Samples with multiple missing measurements were removed from analysis if missing values were recorded for more than 25% of the proteins that were measured in each assay (serum protein, sputum protein or blood cell marker); proteins were then removed if more than two measurements were missing for any one protein. We identified and illustrated individual proteins that were differentially expressed in stable and exacerbated states using a volcano plot. First, a non-parametric, two-sided Wilcoxon paired signed rank test was used to determine significance in the non-normalized proteomic or cell marker expression during the stable and exacerbation states, with significance being defined as $p < 0.05$. Then, the relative fold change in protein or cell marker level was calculated by dividing the average concentration during exacerbation by the average concentration during stability. Each protein or cell marker was then plotted in one figure, with fold change on the x-axis and the p-value on the y- axis. Minor differences between these results and the previously published univariate results (Freeman et al.) can be attributed to variation in which subset of patients were included in each analysis[91].

PLSDA, which was performed using the Eigenvector PLS Toolbox in MATLAB, was used to identify and visualize signatures of multivariate cytokine and cellular markers that differentiated stable and AE-COPD[176]. Taking a supervised approach, PLSDA assigns a loading to each variable and selects a linear combination of all variables (a latent variable) that best separates pre-defined groups. A higher value of a protein loading on a latent variable indicates the protein is of more importance in differentiating the groups of interest. Each sample is then scored based on its protein expression and are visualized in the scores plot. The loadings can be used for hypothesis generation based on how the subsets of the protein signature are associated with each of the groups in the scores plot. Each PLSDA model was cross-validated as a measure of model accuracy. Cross-validation was performed by iteratively excluding ~10% of the data for all models based on serum proteins only, ~17% of the data from the serum and sputum protein PLSDA model, and ~20% from the serum and sputum protein and blood cell marker PLSDA, which in each case resulted in 3-4 samples being excluded. The excluded data was then used to test the trained model. Care was taken when designing the training and test sets to ensure that no test set had more than one measurement from a unique patient. All missing data points included in the PLSDA models were filled in by the Eigenvector software's "best guess." All models were orthogonalized to enable clear visualization of the results, and all data were mean centered and variance scaled before being used to create the model. Variable Importance in Projection (VIP) scores were used to reduce model dimensionality by determining the importance of each variable in differentiating the groups of interest[227]. Proteins with a VIP score < 1 were removed from the model, and a new PLSDA model was then built based on the remaining proteins or cellular factors.

In order to facilitate a more quantitative comparison across PLSDA analyses, we calculated the cross-validation accuracy associated with each training and test set that was created during cross-validation. We then statistically compared cross-validation accuracies across the models based on different folds by using a one-way ANOVA with Tukey's *post hoc* test. A *p* value of less than 0.05 was considered significant after application of Tukey's test.

We visualized the distinct proteomes associated with stable and AE-COPD events through unsupervised average linkage hierarchical clustering; Spearman's correlation coefficient was used as the distance metric. Correlation heat maps were constructed based on the Spearman rank correlation calculated between the difference in cell marker and protein concentration from the stable to the exacerbated state, where correlation coefficients that had a *p* value of greater than 0.05 were set to be zero for the figure. When creating hierarchical clusters or correlation heat maps, all missing data points were imputed using the MATLAB function knnimpute, with the pairwise distances between patients calculated based on the Spearman rank correlation.

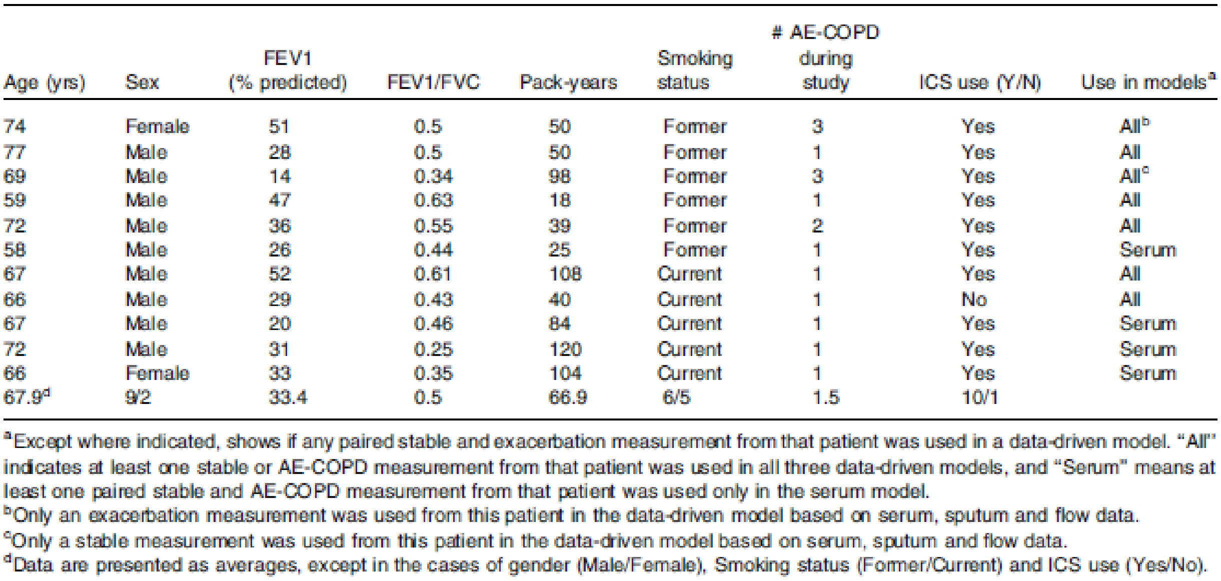
All PLSDA models, VIP scores, Wilcoxon signed rank tests, hierarchical clusters, heat maps, and Spearman correlation testing were created or calculated using MATLAB (MATLAB, Natick, MA); PLSDA models and VIP scores were specifically generated using the PLS toolbox in MATLAB (Eigenvector, Manson, WA). ANOVA and Tukey's tests were performed using Prism version 7.00 (GraphPad Software, San Diego, CA).

5.4 Results

5.4.1 Patient enrollment and demographics

We analyzed data from 13 COPD subjects who completed both the baseline visit and at least one AE-COPD visit. They were a predominantly middle-aged (mean age 67.9 years), male (9 of 11) group with advanced COPD (mean FEV1 33.4% predicted) comprised of both current and former smokers. Specifics of their demographics, clinical characteristics and in which data-driven models their data were used is shown in **Table 5.1**. In summation, this study captured 18 total paired stable and AE-COPD events among the 13 subjects, with some subjects experiencing more than one AE-COPD during the course of the study.

Table 5.1. Summary of demographic, smoking, and spirometry and model inclusion information.



Age (yrs)	Sex	FEV1 (% predicted)	FEV1/FVC	Pack-years	Smoking status	# AE-COPD during study	ICS use (Y/N)	Use in models[a]
74	Female	51	0.5	50	Former	3	Yes	All[b]
77	Male	28	0.5	50	Former	1	Yes	All
69	Male	14	0.34	98	Former	3	Yes	All[c]
59	Male	47	0.63	18	Former	1	Yes	All
72	Male	36	0.55	39	Former	2	Yes	All
58	Male	26	0.44	25	Former	1	Yes	Serum
67	Male	52	0.61	108	Current	1	Yes	All
66	Male	29	0.43	40	Current	1	No	All
67	Male	20	0.46	84	Current	1	Yes	Serum
72	Male	31	0.25	120	Current	1	Yes	Serum
66	Female	33	0.35	104	Current	1	Yes	Serum
67.9[d]	9/2	33.4	0.5	66.9	6/5	1.5	10/1	

[a] Except where indicated, shows if any paired stable and exacerbation measurement from that patient was used in a data-driven model. "All" indicates at least one stable or AE-COPD measurement from that patient was used in all three data-driven models, and "Serum" means at least one paired stable and AE-COPD measurement from that patient was used only in the serum model.
[b] Only an exacerbation measurement was used from this patient in the data-driven model based on serum, sputum and flow data.
[c] Only a stable measurement was used from this patient in the data-driven model based on serum, sputum and flow data.
[d] Data are presented as averages, except in the cases of gender (Male/Female), Smoking status (Former/Current) and ICS use (Yes/No).

5.4.2 Evaluation of individual immune factors associated with AE-COPD

We first identified individual cellular immune factors and receptors that differed significantly between stable and AE-COPD, similar to our previously published work[91]. Out of 35 serum proteins (see Materials and Methods in **Section 5.2**), five were found to be

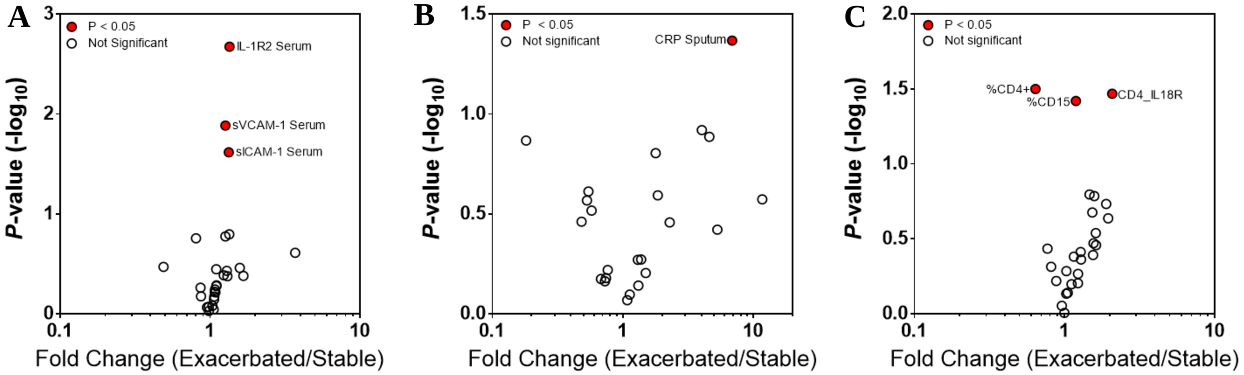


Figure 5.1. Individual proteins and cell populations measured in stable and exacerbated states.
(**A**) Volcano plot illustrates serum proteins that are both differentially expressed (x axis) and significantly different (y axis) between the stable and exacerbated state. Significance was determined using non-normalized data (**Supplemental Figures C.S1, C.S2** and **C.S3**), and points in red indicate significantly different expression between the stable and exacerbated state via paired Wilcoxon signed rank test, with significance being defined as $p < 0.05$. (**B**) Volcano plot highlighting significantly different sputum proteins across the stable and exacerbated state. Significance was determined as described above ($p < 0.05$). (**C**) Volcano plot illustrating blood cell marker measurements that were significantly different between stable and AE-COPD. Significance was determined as described above ($p < 0.05$).

significantly different ($p < 0.05$): interleukin 1 receptor 2 (IL-1R2; fold change 1.35), soluble intercellular adhesion molecule 1 (sICAM-1; fold change 1.33), soluble vascular cellular adhesion molecule 1 (sVCAM-1; fold change 1.27), growth differentiation factor (GDF-15; fold change 1.29) and interleukin 10 (IL-10; fold change 1.66) (**Figure 5.1A**). From 30 proteins measured in sputum, only CRP was significantly different between stable and AE-COPD (fold change 5.56) (**Figure 5.1B**). Three of 26 cellular markers measured by flow cytometry were differentially expressed: percent of CD4+ cells (%CD4+; fold change 0.61), CD4+ CD62L cells (CD4_CD62L, fold change 1.03), and CD4+ IL-18R cells (CD4_IL18; fold change 2.08) (**Figure 5.1C**). The expression of both CD62L and IL-18R indicate activation of CD4+ T cells. While the significance levels indicated in the volcano plots are based on average concentration data, the grouped scatter plots in **Supplemental Figures C.S1**, **C.S2**, and **C.S3** track individual changes across the two COPD states in specific patients. All immune factors were significantly elevated during exacerbation with the exception of %CD4+ cells. Overall, these results reflect observations in the original study[91], in which only a small number of proteins and individual blood cell types and activation markers were significantly different between stable and exacerbation. None of the proteins or cell markers in the three volcano plots were found to be significant after application of the Bonferroni correction and many of the fold changes measured were small (close to 1).

In our data there were three patients who had more than one exacerbation event. We explored the effects of this by additionally analyzing the data after averaging multiple stable and multiple exacerbation measurements within the same patient. Overall, we found that our results were similar, both in individual significant proteins identified and in fold change in the exacerbated state (**Supplemental Figure C.S4**).

Additionally, we also constructed a model of exacerbation based only on protein measurements in sputum samples. This VIP-selected PLSDA model performed with 91.67% calibration and 78.33% cross-validation accuracy and can be found in **Supplemental Figure E.9.**

5.4.3 PLSDA identified a signature of serum proteins that differentiated stable and exacerbated COPD

To obtain new insight into key systems-level relationships between networks of immune factors in sputum and blood that associated with AE-COPD, we next employed data-driven modeling approaches to integrate matched stable and exacerbation data in both blood and pulmonary immune environments from the same COPD patients. We first examined serum protein measurements alone with PLSDA[176]. PLSDA is a useful tool due to its ability to highlight covariance among variables that best classify groups of interest, which could lead to the identification of potential proteomic networks associated with AE-COPD. Calibration accuracy and k-fold cross-validation were used to assess model accuracy (see Materials and Methods in **Section 5.2**). To focus on the cytokines that were best at differentiating stable and

AE-COPD, we used variable importance in projection (VIP) scores[227] as a feature selection technique. The value of using PLSDA with VIP feature selection is the identification of small protein “signatures” that differentiate groups of interest and are potentially biologically meaningful, which helps with generating new mechanistic hypotheses.

We found that a two-latent variable PLSDA model based on the serum VIP-selected protein signature best classified stable and exacerbation points with 81.25% cross-validation accuracy and an 84.38% calibration accuracy (**Figure 5.2A**). Latent variable 1 (LV1) differentiated most stable visits (negative scores on LV1) from AE-COPD (positive scores on LV1; **Figure 5.2B**). Six of the seven proteins were loaded positively on LV1, indicating positive association with AE-COPD, while only tissue inhibitor of metalloproteinases (TIMP4) was loaded negatively on LV1, indicating negative association with AE-COPD. The six positively associated proteins were IL-1R2, sVCAM-1, sICAM-1, matrix metalloproteinase 9 (MMP-9), interferon gamma-induced protein 10 (IP-10, the chemokine also known as CXCL10), and IL-6.

We next compared the classification ability of this signature to the classification ability of the top individual factors identified in univariate analysis of these data[91]. The univariate model indicated that IL-10, IL-15, GDF-15, sICAM-1, and sVCAM-1 were individual factors that were significantly increased during exacerbation[91]. For the purpose of comparing multivariate with univariate results, we took each of the top significant individual mediators in previous analysis (sICAM-1, sVCAM-1, and IL-15) and assessed their individual ability to classify stable and AE-COPD. We then made a PLSDA model where we combined all five significant proteins previously identified through univariate analysis. We compared the performance of these four analyses to our VIP-selected PLSDA model described above, using the cross-validation accuracy and the calibration accuracy as comparison metrics. The cross-validation accuracy of the VIP-

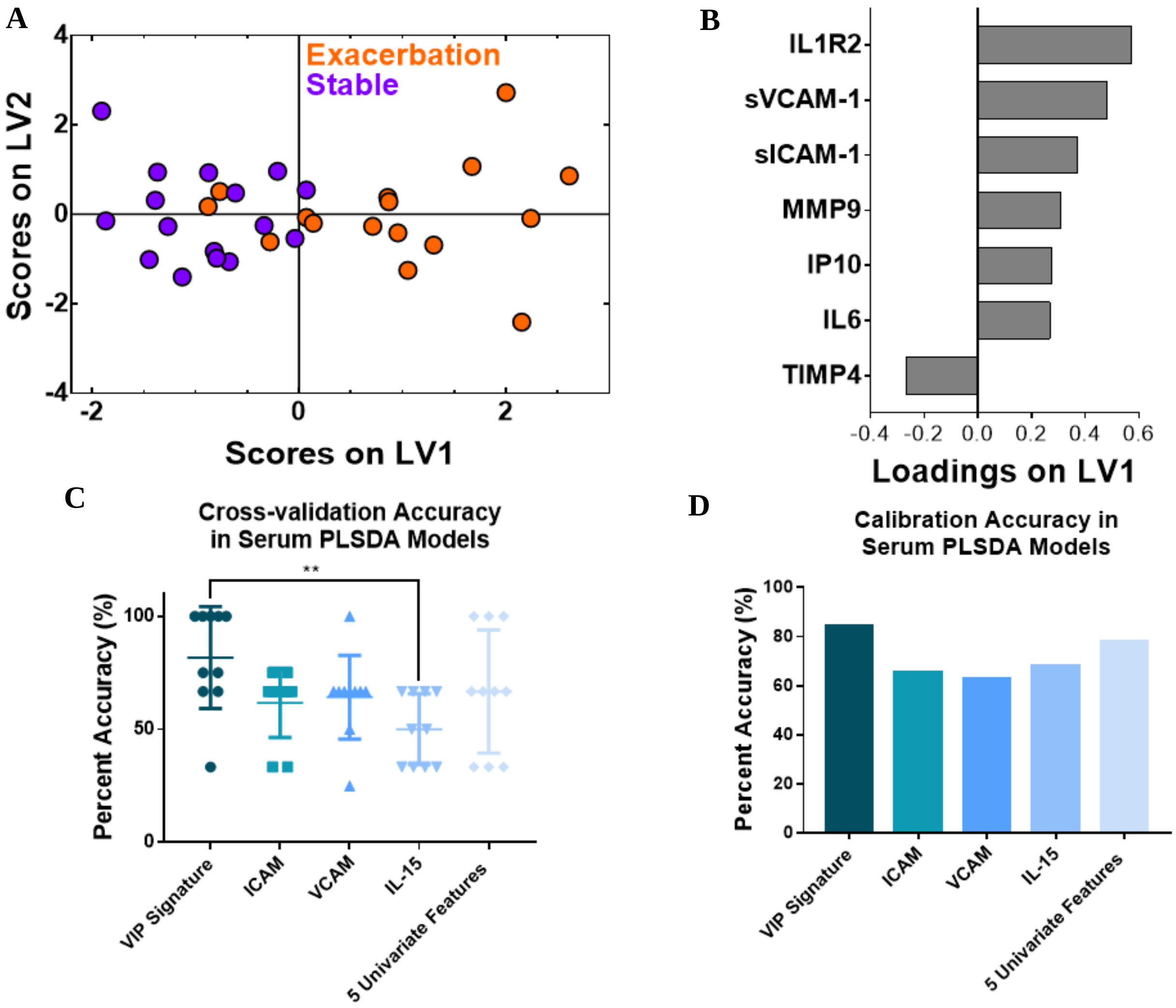


Figure 5.2. VIP scores and PLSDA identified a signature of 7 serum proteins that differentiated a stable from exacerbation measurement in 16 paired stable and AE-COPD events experienced by 11 unique patients.
(**A**) VIP scores identified a 7-protein serum signature that differentiated stable (purple) and exacerbation (orange) events with 81.25% cross-validation accuracy and 84.38% calibration accuracy. Latent variable 1 (LV1) accounted for 25.00% of the variance in the data, and latent variable 2 accounted for 16.75% of the variance in the data. (**B**) The loadings plot shows how much each protein contributes to the signature, with positive loadings associated with exacerbation events, and negative loadings comparatively reduced in exacerbation. (**C**) Comparison of the differentiation between stable and exacerbated states based on individual factors vs. multivariate signatures. The VIP signature identified by the PLSDA models trended towards higher cross-validation accuracy than individual factors that were most significantly different. A one-way ANOVA determined that this signature was significantly better than IL-15 alone, with ** indicating a p-value less than 0.01 after Tukey's test for multiple comparisons. (**D**) Comparison of the calibration accuracies for individual factors vs. the VIP signature identified by the PLSDA model.

selected PLSDA model trended towards being higher than all analyses based on single significant proteins, but was only significantly better than the cross-validation based on IL-15 alone ($p < 0.01$, one-way ANOVA with Tukey's HSD) (**Figure 5.2C**). The VIP-selected PLSDA model did have the highest calibration accuracy out of all five accuracies that were compared

(**Figure 5.2D**). Overall, these figures serve to highlight the use of co-varying features, or "signatures," in differentiating exacerbation events.

5.4.4 Insight into cross-tissue compartment proteomic interactions associated with AE-COPD

To gain deeper insight into relationships between immune factors in lung and serum tissue compartments involved in AE-COPD, we used PLSDA to integrate data from serum and sputum measurements in stable and exacerbated states. We first evaluated proteins for which both paired sputum and serum results were available (n=9 matched stable and AE measurements), creating a PLSDA model based on 60 total analytes and employing VIP feature selection to eliminate those not contributing to differentiation. A one-latent variable PLSDA model separated exacerbation and stable measurements with a cross-validation and calibration accuracy of 88.89%, though a two-latent variable PLSDA model scores plot is presented to facilitate interpretation of group clustering (**Figure 5.3A**). LV1 largely differentiated the stable

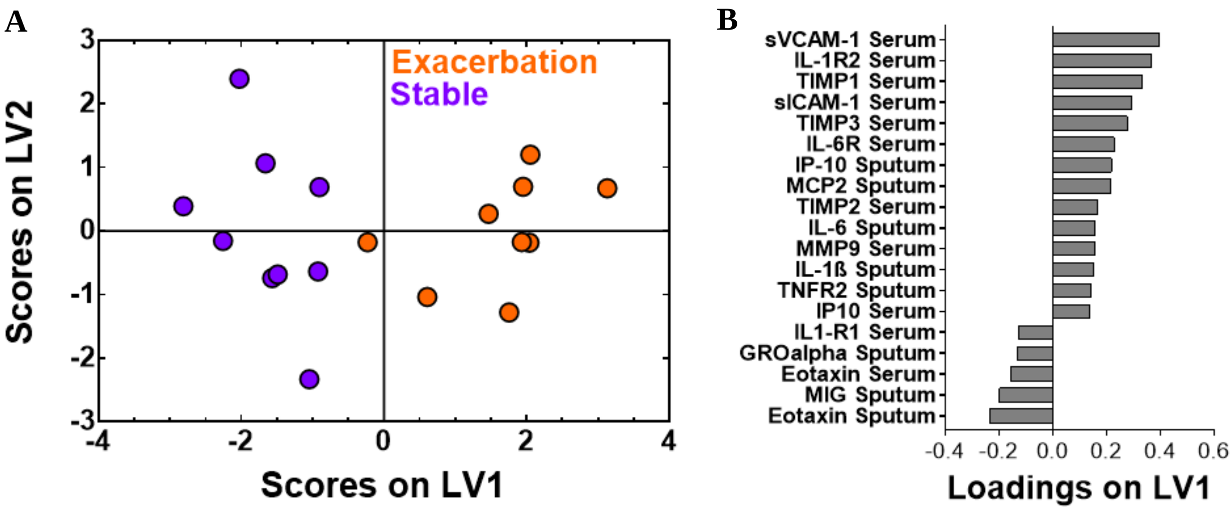


Figure 5.3. A one latent variable PLSDA model of VIP-selected proteins from the serum and sputum samples combined resulted in clear differentiation between stable and exacerbation measurements across 9 paired stable and AE-COPD events experienced by 7 unique patients.

(**A**) PLSDA and VIP scores identified a signature of 19 proteins that differentiated the stable (purple) from exacerbation (orange) states with 88.89% cross-validation and calibration accuracy. Latent variable 1 accounted for 21.73% of the variance in the data. The scores plot shown is based on a two latent variable model to enable better visualization of group separation. (**B**) The loadings plot illustrates the protein contributions to the VIP-selected signature, with positive loadings positively associated with the exacerbation measurements, and negative loadings comparatively reduced during exacerbation.

state (negative scores on LV1) from AE-COPD (**Figure 5.3B**). Fourteen of the nineteen proteins were loaded positively on LV1, indicating positive association with AE-COPD, whereas five proteins were associated with stable COPD. Of the fourteen proteins that were positively associated with exacerbation, many of the serum proteins have been established as adhesion factors or chemokines (sICAM-1[228], sVCAM-1[229], IP-10[230], MCP-2[231]), while most of the sputum proteins were known inflammatory factors (IL-6[232], IL-1β[233], TNFR-2[234]). Similar to the serum-only model, this signature suggests migration and activation of innate immune cells in the serum during exacerbation, yet the addition of sputum data to the model demonstrates the corresponding importance of lung inflammation and chemokine secretion. As classification accuracy of the combined serum-sputum model was better than either separately, these results highlight the importance of the parallel relationship between chemokine secretion in lung and innate immune cell activation in serum.

5.4.5 Integration of data across experimental assays gives additional insight into the cellular and proteomic mechanisms associated with AE-COPD

We also used our systems approach to integrate data across experimental assays by adding flow cytometry measurements, which were performed only on whole blood samples. We specifically explored whether PLSDA might help us integrate measurements made in different experimental assays. PLSDA and two rounds of VIP selection identified a one-latent variable model and a signature of eleven cell markers and proteins that differentiated stable COPD from AE-COPD with a cross-validation accuracy and a calibration accuracy of 87.5%. Differentiation between states (**Figure 5.4A**) was driven by the loadings on LV1, which separated most individuals by exacerbation status (**Figure 5.4B**). Nine of the cytokines and cell markers were loaded positively on LV1, indicating positive association with exacerbation, and two were loaded

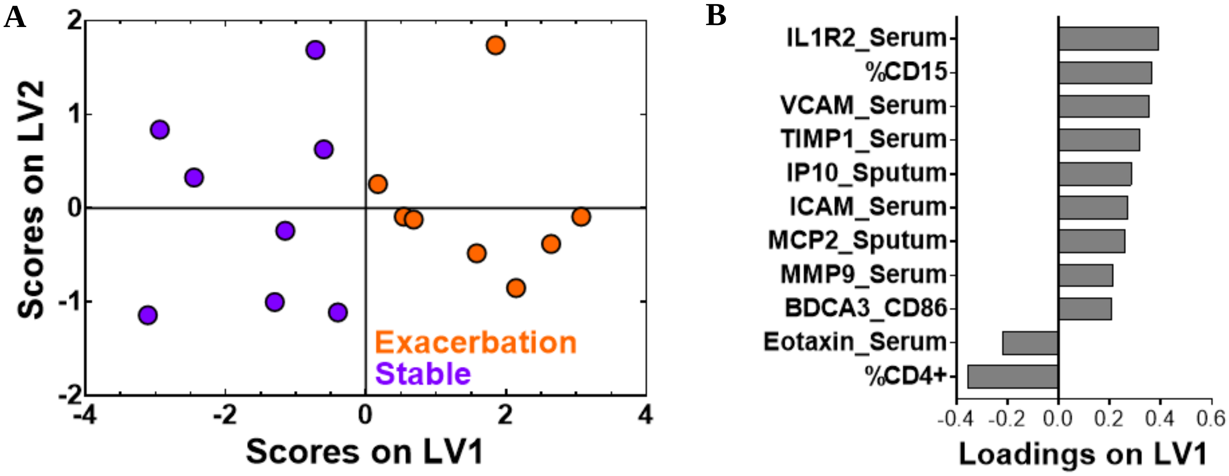


Figure 5.4. A one latent variable PLSDA model based on two rounds of VIP selection from serum and sputum proteins and blood flow markers shows clear differentiation between stable and exacerbation events across 8 pairs of patient samples, which included 7 paired stable and AE-COPD events experienced by 6 unique patients and one stable and one exacerbation measurement that were not patient matched.
(**A**) PLSDA and two rounds of VIP analysis identified a signature of eleven factors that differentiated the stable (purple) from the exacerbation (orange) events, with 87.5% calibration and cross-validation accuracy. Latent variable 1 (LV1) accounted for 41.51% of the variance in the data. The scores plot shown is based on a two latent variable model to enable better visualization of group separation. (**B**) The loadings plot highlights factor contributions to the VIP-selected signature, with positive loadings positively associated with AE-COPD, and negative loadings comparatively reduced during an exacerbation event.

negatively on LV1, indicating negative association with exacerbation. Cellular factors associated with exacerbation in the integrated PLSDA model included CD86 expression by BDCA-3+ dendritic cells (DC) and the percentage of CD15+ granulocytes (reported in the original study to be neutrophils)[91]. In contrast, the percent of CD4+ T-cells was found to be associated with the stable measurements in this model.

We next compared the cross-validation accuracies across all three of the VIP-selected models that consisted of varying amounts of tissue compartment and assay data. Although none of these three models were significantly different from each other according to Tukey's *post hoc* test (one-way ANOVA), inclusion of data from more tissues and assays in the model trended toward a tighter and higher range of cross-validation accuracies (**Supplemental Figure C.S5**).

To visualize the unbiased classification ability of this signature, we also employed hierarchical clustering and created a heat map (**Supplemental Figure C.S6**). We found this clustering algorithm based on distance metrics was not as useful for classification, with three stable and four exacerbation samples misclassified out of sixteen total samples (56.25% classification accuracy). As our data contained measurements from three individuals with more than one exacerbation event, we also examined our scores plot after labeling the points with the patient's exacerbation status and visit number. The resulting scores plot (**Supplemental Figure C.S7**) indicates no clear intra-patient clustering, though this study was not powered for a thorough statistical analysis in this direction.

We further explored potential relationships between cell numbers and protein concentrations across the stable and exacerbated states in our identified signatures using Spearman rank correlation coefficients and a heat map. Overall we found that MMP-9 in the serum was positively correlated with CD4+ cells expressing the IL-18 receptor, and TIMP1 in the serum was positively correlated with CD4+ cells expressing the CD122 activation marker. The BDCA3+ CD86+ and the %CD15 neutrophils were not correlated with the other proteins in the signature, but were correlated with other measured proteins (**Supplemental Figure C.S8**). Overall, this suggests that changes in cell number from the stable to the exacerbated state may be related to simultaneous increases in concentration of some inflammatory proteins across the two states.

5.5 Discussion

Using systems analysis of paired data points from cellular factors measured in blood and sputum in exacerbated and stable COPD states, we identified a signature that differentiated AE-COPD with >87% cross-validation accuracy. This signature trended towards being better than any previously identified individual cellular factors for differentiating stable and exacerbated COPD states, though more measurements would be needed to determine statistical significance. Biologically, the signature indicated that parallel increases in inflammatory cytokines and chemokines in sputum environments, adhesion/chemoattractive cytokines in serum environments, and greater numbers of BDCA-3+ DC and an increased percent of CD15+ neutrophils in the blood were all associated with AE-COPD. These results highlight the value of computational approaches when integrating measurements across tissue compartments and from different experimental assays, and motivate use of these approaches to gain new perspective into cellular systems involved in this prevalent, lethal, but understudied disorder.

One important strength of our approach is the ability to define parsimonious cellular signatures by selecting the most significant co-varying cellular immune factors. This approach may be valuable as a means of defining key cellular systems involved in disease progression, and using these to efficiently choose end-points in clinical trials and guide future experimental endeavors. This approach is especially useful for integrating cellular measurements made in multiple tissue compartments, which is important given the central role of sputum production in AE-COPD. Based on these findings, we propose a model of key networks in AE-COPD (**Figure 5.5**) involving specific immune cell types, metalloproteinases (MMPs) and tissue inhibitors of metalloproteinases (TIMPs), and chemokines. We discuss our findings in that framework.

In terms of peripheral blood leukocyte participation in AE-COPD, we extend the observation from univariate analysis of these data[91] that CD4+ T cells decreased in blood during exacerbation, which is compatible with trafficking to lung or regional lymph nodes (or both), by showing the importance of simultaneous increase in blood of BDCA-3+ DC. We have previously demonstrated the physical interaction of this DC subset with CD4+ T cells in lung tissue from COPD patients[235]. BDCA-3+ DC were previously termed mDC2, but are now designated as cDC1[236]; they are the counterpart of murine CD103+ DC, which are essential for cross-presentation of viral antigens to CD8+ T cells. Our model suggests recruitment to the lungs of

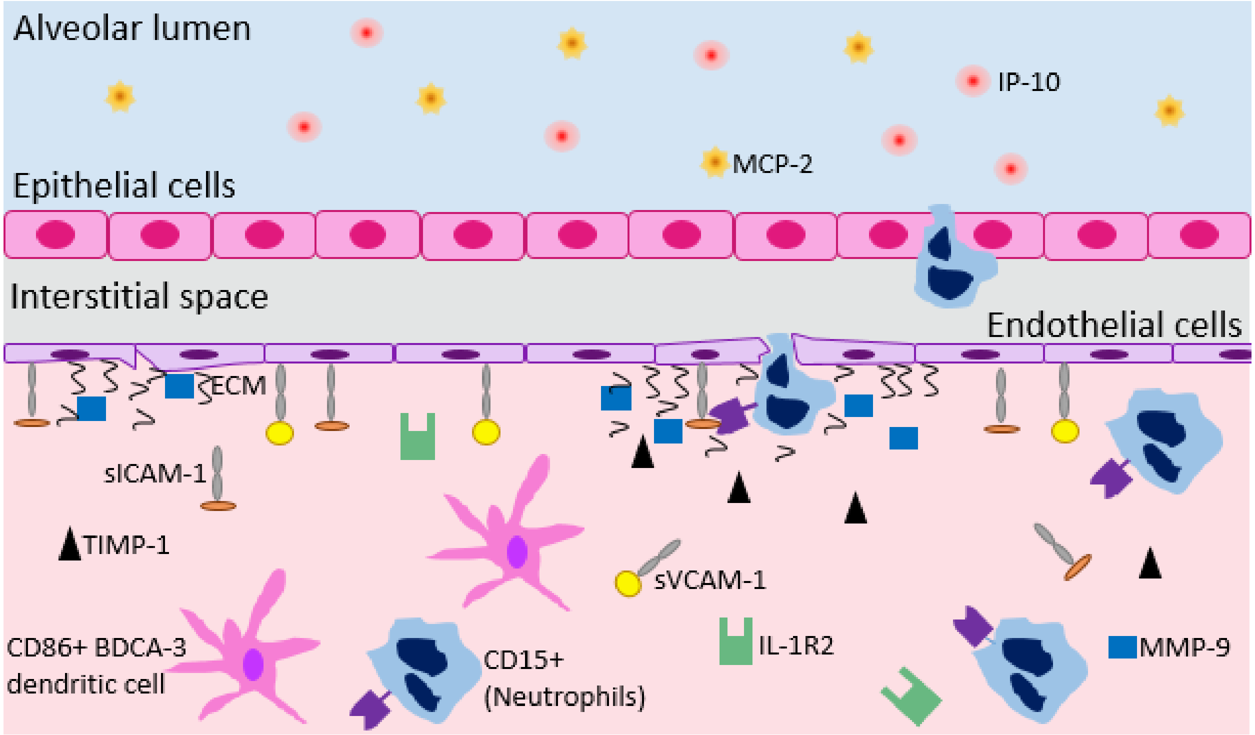


Figure 5.5. A hypothesis of cross-tissue mechanisms of action in the lungs and blood of patients experiencing an AE-COPD.
Adhesion molecules aid in moving immune cells from the blood to the lung, which is further promoted by the presence of the chemokine interferon gamma-induced protein 10 (IP-10) and monocyte chemoattractive protein 2 (MCP-2) in the sputum. sICAM: soluble intercellular adhesion molecule. sVCAM: vascular cell adhesion molecule. TIMP: tissue inhibitor of metalloproteinases. MMP: matrix metalloproteinase. R2: receptor 2. ECM: extracellular matrix. CD: cluster of differentiation. BDCA: blood dendritic cell antigen.

cDC1, likely from the bone marrow, as a crucial step driving lung inflammation during AE-COPD. The other type of leukocyte in our signature, neutrophils, has been shown by other studies to be linked to AE-COPD[237], one of which related their numbers to exacerbation severity[210].

Key soluble factors in our signature agree with and extend previous individual associations of inflammatory mediators with AE-COPD. These not only include the anticipated agreement with previous univariate analysis of these data[91], but also several serum proteins involved in adhesion and chemoattraction of inflammatory cells. Chief among these is the neutrophil chemoattractant IP-10/CXCL10, also found to be elevated in AE-COPD in two studies[89,92]. Our signature also included IL-6, a pro-inflammatory cytokine[232] that has been vigorously investigated as a possible biomarker for AE-COPD. Increased IL-6 in serum and sputum during AE-COPD was reported by several large studies using longitudinal design[92,238]; this association was questioned in a systematic review which, however, included many studies of cross-sectional design[239]. Our results illustrate the superior power of comparing paired results from the same subjects across stable and exacerbated states. We also identified elevations in levels of sICAM-1 and sVCAM-1, truncated forms of transmembrane adhesion molecules that interact with leukocyte integrins. sVCAM is chemotactic for murine neutrophils in vitro[240]. sICAM-1 is expressed both by leukocytes and by activated endothelial cells, and levels of sICAM-1 correlate to endothelial cell ICAM expression in vitro[230]. Each of these proteins are elevated in stable COPD[241,242], though to our knowledge, no study (other than our original data) has linked it to AE-COPD in longitudinal data. sICAM has been reported to be elevated in subjects admitted for AE-COPD compared with healthy control subjects[243]. Higher plasma sICAM-1 levels were also independently associated with emphysema progression in the Multi-Ethnic Study of Atherosclerosis (MESA) Lung cohort, a general population sample[244].

Our signature identified elevated serum MMP-9 as a crucial feature of AE-COPD, in agreement with a previous study[245]. Also known as gelatinase B, MMP-9 is released by activated neutrophils[246]. It has an unique ability to induce self-perpetuating lung inflammation by degrading extracellular matrix, thus liberating the neutrophil chemoattractant tripeptide N-acetyl Proline-Glycine-Proline[231]. Along with IL-6, MMP-9 was one of 34 serum analytes found to be highly reproducible over a 6 week period of clinical stability in COPD patients[247], further supporting our findings. Our MMP-9 finding is interesting in light of the disparity between the association with exacerbation of TIMP1, TIMP2, and TIMP3, which stoichiometrically inhibit MMP activity[248,249], and TIMP4, which associated with the stable state in the VIP signature. Unlike the other three TIMP family members, which act as soluble inhibitors, TIMP3 is typically bound to matrix sulfated glycosaminoglycans[248], suggesting that its presence in the serum during AE-COPD might reflect matrix degradation.

All of our models identified IL-1R2 as a crucial serum factor increased during AE-COPD, in agreement with two studies from the group in Maastricht of patients admitted for AE-COPD[250,251]. IL-1R2 (Gene ID: 7850) is an early response gene[252] whose product is a decoy receptor that inhibits activity of its three ligands: IL-1α, IL-1β, and the type I IL-1 receptor. Together with associations for TIMP1-3, our results highlight the importance of counter-regulatory factors during AE-COPD. Although all the subjects in the original dataset were successfully treated as outpatients with resolution, not all patients regain lung function following AE-COPD; an intriguing possibility is that those who do not recover entirely might exhibit relatively deficient up-regulation of IL-1R2 and TIMPs during AE-COPD.

There are several limitations to this analysis. Although our original study[91] recruited a larger group of subjects, many sought treatment for AE-COPD locally, rather than returning when acutely ill. Additionally, some measurements had to be excluded from this analysis due to missing data. Collectively, these factors reduced our sample size, making it all the more noteworthy that our approach identified AE-COPD cellular signatures that could be used to gain biological insight. However, the small sample size did limit our ability to find signatures that could be used in diagnostic contexts. Even though our identified signature trended towards being better than individual factors, it was only statistically significant in one case. Furthermore, additional unknown test data in different patient cohorts would be needed to truly assess signature classification ability for diagnostic purposes. A second limitation is the necessary dependence on proteins measured in the original study, which used a “candidate gene” approach based in part on prior knowledge, and not an unbiased screen of the entire proteome. Because our original study involved flow cytometric analysis of peripheral blood leukocytes collected in part during AE-COPD, there is, to our knowledge, no current exacerbation cohort available for validation testing. However, to prevent model overfitting as much as possible, we did employ internal cross-validation.

Results of this work support exciting future research in several directions. First, if similar data from other cohorts of paired stable and exacerbation measurements were to become available, generated models could be tested and validated. Data-driven approaches such as these could be applied as a classification tool to identify differences in exacerbation endotypes or in AE-COPD events resulting from different upstream causes (including viruses, bacteria, etc.), thus providing insight into systems-level mechanisms of action that could result in personalized treatment options. Unbiased data-driven models applied to multiplex COPD data from across tissue compartments may also prove useful to characterize COPD endotypes.

Chapter 6 Proteomic Signatures and Immune Cell-Cell Communication Patterns in a Large Clinical Cohort Associated With COPD Disease State and Severity

Contributions. *The SPIROMICS investigators were involved in collecting the blood and BAL samples from the SPIROMICS smokers, never smokers, and COPD subjects. Drs. Curtis and Freeman collected whole blood from subjects for PBMC isolation, stored these samples until stimulation, and, along with lab manager Valerie Stolberg, assisted in the PBMC stimulation experiments planning and execution.*

6.1 Introduction

Following work presented in **Chapter 5**, we generated additional models of COPD cell-cell communication networks, disease state, and disease severity that have not been published. For this work we had access to data and samples from smokers, never smokers, and COPD subjects enrolled in the Subpopulations and Intermediate Outcomes in COPD Study (SPIROMICS)[253], which was extremely valuable due to the cohort's large size (2,981 subjects recruited to the overall study), the variety of clinical data and matched biological samples available from some patients, and the 5 year follow up visit for these patients associated with the SPIROMICS II study which are currently underway. The SPIROMICS II visits are of key interest to us because they will involve the collection of another set of clinical measurements (FEV_1, CT scans, etc.) and biological samples (blood, BAL, etc.) from many of the original subjects, which would allow for investigations into progression. Although 2,981 subjects participated in SPIROMICS overall, a subset of these subjects also qualified for enrollment in the Bronchoscopy study and had BAL samples collected on top of blood, sputum, and CT scan measurements (n = 215)[254]. In the end, around 190 smokers, never smokers, and COPD subjects (mostly classified with mild to moderate COPD) were enrolled in the Bronchoscopy substudy. These matched blood and BAL samples from subjects allowed us to fulfill Aims 2B and 3B by creating data-driven models based on BAL proteins alone and based on blood and BAL proteins combined. Once identified, we were able to further investigate the signatures' potential biological meaning in the context of COPD disease state and severity. Additionally, with help from Drs. Curtis and Freeman at the Veteran's Affairs Ann Arbor Healthcare System (VAAAHS), we were also able to collect whole blood samples from another group of smokers, never smokers, and COPD subjects visiting the VAAAHS and isolate peripheral blood mononuclear cells (PBMCs) from these samples. We then compared differences in PBMC communication networks of the three groups in response to various immune stimuli by creating data-driven models of the secreted proteins during these stimulations, which satisfied Aim 1B.

As discussed in **Chapter 1**, the mechanisms that underlie COPD disease state and progression are complex and not well understood, and thus new approaches must be taken in order to gain insight into this area. Some of the difficulty in studying COPD comes from the heterogeneity associated with the disease. Examples of this heterogeneity include the lack of concrete biomarkers for COPD or exacerbations[94,255,256]; the current conversations about defining an early COPD state, which focus on how it is still unknown why some people (especially smokers) develop COPD when others do not[257,258]; and the research into different underlying endotypes that could result in the same COPD phenotype[94,259,260]. The conversations about early COPD come from the fact that smokers and COPD subjects may be more similar than we currently realize, as it has been seen that smokers without airway obstruction are still more likely to experience negative respiratory events when compared to never smokers[261,262]. This could mean that models of COPD disease state that contain smokers, never smokers, and COPD subjects may hold more diagnostic or prognostic value than our model of healthy and IPF subjects (**Chapter 2** and **4**). To investigate these topical questions about COPD further, other large longitudinal cohorts of smokers, never smokers, and COPD subjects have been created in addition to SPIROMICS, such as the Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE) study[263] and COPDGene[264].

The SPIROMICS cohort enabled us to apply data-driven modeling approaches to integrate matched lung and blood protein data collected from enrolled subjects, similar to our approach in IPF (**Chapters 2-4**). With the SPIROMICS samples, as we did in IPF, we wanted to create models that took advantage of proteins that were measured in bronchoalveolar lavage (BAL) samples to gain insight into lung-specific signatures associated with disease state or progression. There has been more work with BAL protein data in COPD disease state and progression than in IPF[265–273], but few studies have measured more than 10 proteins in each sample[274–277], and fewer still that have focused on the use of multiple BAL proteins to differentiate clinical groups[278,279]. To our knowledge, although single cytokines across blood and BAL samples have been investigated in the context of COPD disease state or severity[271,280–284], this work is the first time that signatures of blood and BAL proteins have been used to differentiate and hypothesize potential mechanisms associated with COPD disease state or severity. This presents the opportunity for us to generate unique insight into cross-tissue compartment mechanisms in human subjects that may potentially be involved with COPD disease state or severity.

Though it is clear that the immune system is altered in COPD[66], previous evaluation of individual immune cell types and cytokines has not yet led to definitive biomarkers or broadly effective treatment options[80,82,285]. In addition to current literature, our work which we will discuss in this chapter suggests that circulating blood cytokines alone are not useful for differentiating COPD disease state. However, previous work from the HIV field suggests that stimulation of immune cells from whole blood samples can be used to both gain insight into adverse immune responses associated with immunological diseases, and to design targeted follow-up experiments to test newly generated hypotheses[205]. This was achieved by collecting and culturing PBMCs (which include T cells, B cells, NK cells, and monocytes) in the presence of various innate and adaptive immune stimuli, such as lipopolysaccharide (LPS) or anti-CD3/CD28 beads[205]. In the second part of this chapter, we describe preliminary results in applying this approach to identify systems-level differences in immune cell communication networks of smokers, never smokers, and COPD subjects.

6.2 Results

6.2.1 Models based on signatures of proteins from multiple tissue compartments led to better classification of COPD disease state and severity

We first wanted to determine whether there were differences in individual cytokines measured in BAL samples from smokers, never smokers, and COPD subjects. We measured individual expression of cytokines in these groups and investigated if there were any significant differences in expression of any of these proteins. Using a one-way ANOVA with Tukey’s post hoc test, we saw that 11 out of 25 measured proteins were significantly differentially expressed across at least two of the clinical groups; a selection of these results can be seen in **Figure 6.1**.

Due to the variation within individual expression of measured BAL proteins, we next employed data-driven analysis techniques to try to obtain more biologically meaningful separation of the clinical groups when focusing on covariation between proteins. VIP-scores

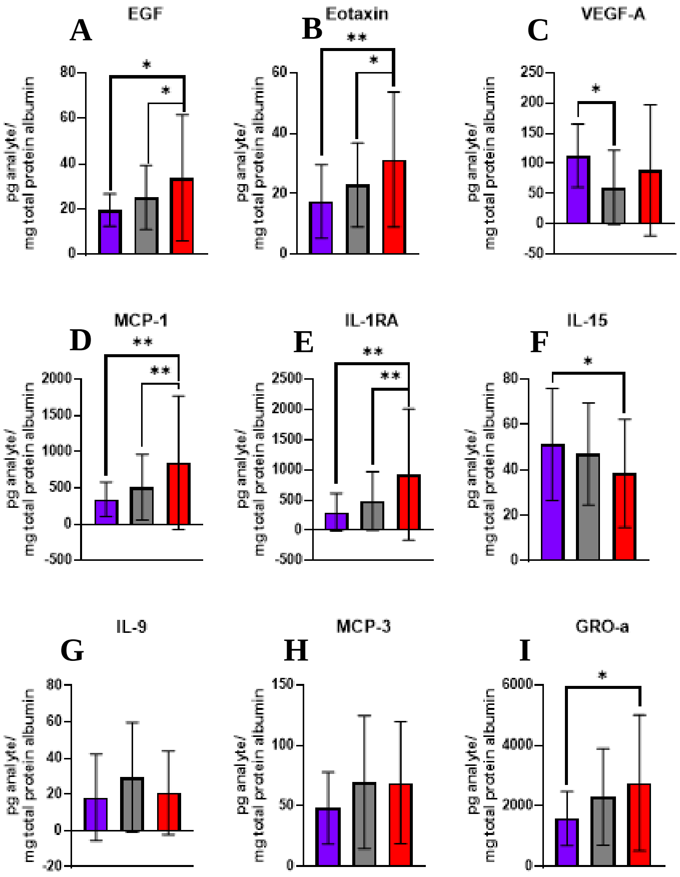


Figure 6.1. (A-I) Representative individual cytokine measurements from BAL samples from smokers with (red) and without (gray) COPD and healthy controls (purple) enrolled in the SPIROMICS study. * p < 0.05, ** p < 0.01; one-way ANOVA with Tukey's post hoc test. n = 25 never smokers, 75 smokers, and 82 COPD patients.

selected and PLSDA visualized a signature of 18 BAL cytokines that differentiated the three groups with 75.31% calibration and 71.76% cross-validation accuracy (**Figure 6.2A**). Samples showed trends towards separation on latent variable 1 (LV1), with COPD subjects having higher scores on LV1 (**Figure 6.2B**).

We then determined if the classification ability of circulating, unstimulated blood cytokines was better than that of BAL proteins in differentiating COPD disease state. VIP scores selected a signature of 22 Luminex-measured cytokines that differentiated the three groups with 65.69% calibration and 57.88% cross-validation accuracy (**Figure D.1A**). Little differentiation could be seen across LV1 or LV2 (**Figure D.1B**). Due to the poor performance of this model, we then used the feature selection technique LASSO to identify a signature of 24 SOMAmer-measured blood proteins that differentiated the groups with 74.56% calibration and 67.59% cross-validation accuracy (**Figure D.2A** and **D.2B**). While this model outperformed the one based only on plasma cytokines, the model based on BAL proteins still had the highest calibration and cross-validation accuracy. However, when comparing these models, it must be noted that confounding factors

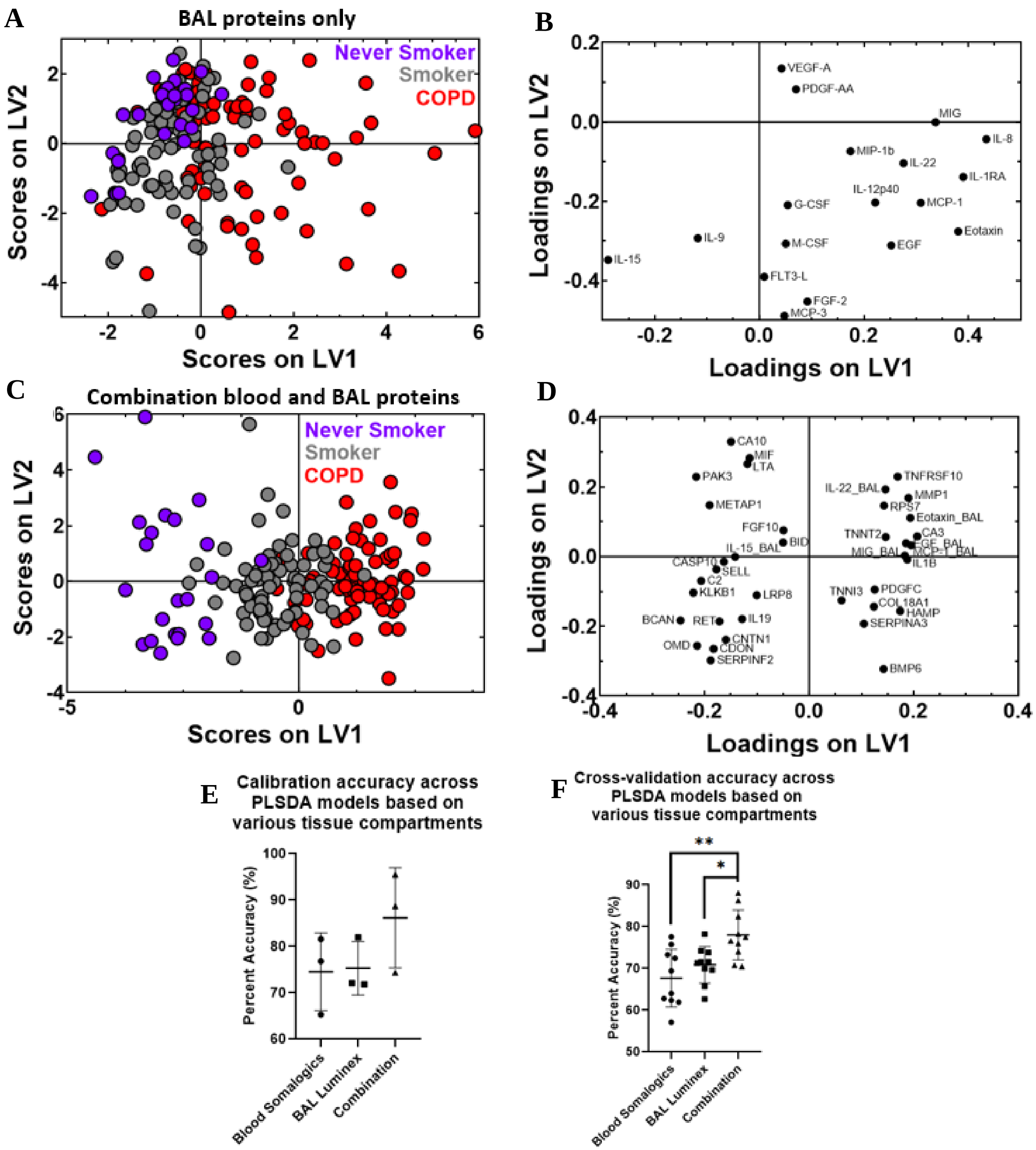


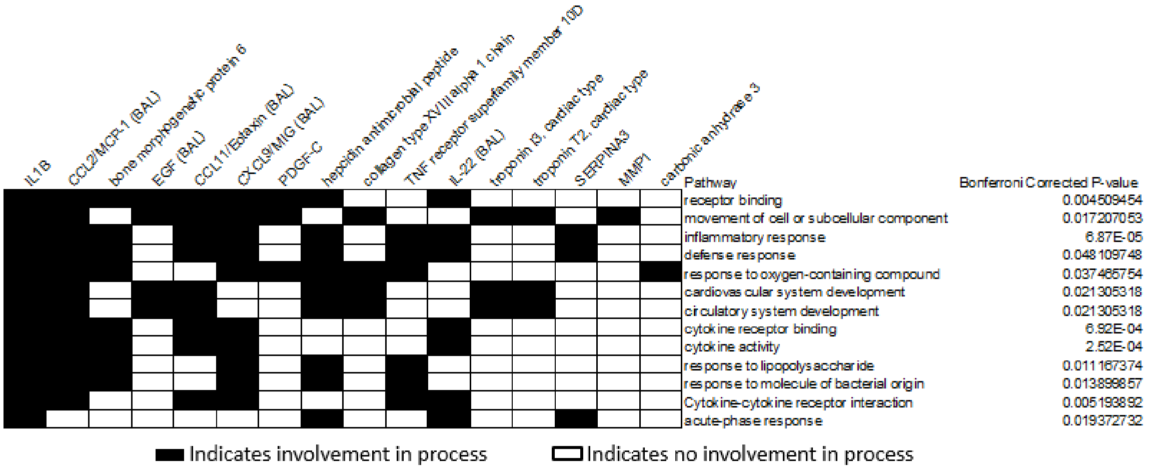
Figure 6.2. Feature selected PLSDA model based on blood and BAL protein data combined differentiates COPD disease state significantly better than models based only on protein data from a single tissue compartment. (**A**) PLSDA scores plot based on BAL proteins moderately separates smokers (grey), never smokers (purple), and COPD subjects (red) with 75.31% calibration (Cal) and 71.76% cross-validation (CV) accuracy. (**B**) PLSDA loadings plot captured 12.27% of the variance on latent variable 1 (LV1). (**C**) PLSDA scores plot based on blood and BAL proteins highlights differentiation between the three clinical groups; the model separated the groups with 86.18% Cal and 76.52% CV accuracy. (**D**) The loadings on LV1 captured 7.26% of the variance in the data. (**E, F**) The cross-tissue compartment model trended towards higher Cal accuracy (**E**) and had significantly higher CV accuracy (**F**) than models based on blood or BAL proteins alone. Significance in both cases calculated by a one-way ANOVA with Tukey's post-hoc test; ** indicates p < 0.01 and * indicates p < 0.05.

such as current smoking status were not able to be corrected for when using VIP scores as a feature selection technique, but were corrected when using LASSO for feature selection.

We discovered that a model based on protein data from multiple tissue compartments was a better classifier of COPD disease state and could be useful in gaining a deeper understanding of holistic relationships associated with COPD. We used LASSO to identify a signature of 37 proteins (31 from the blood and 6 BAL proteins) that differentiated smokers, never smokers, and COPD subjects with 86.18% calibration and 76.52% cross-validation accuracy (**Figure 6.2C**). The three groups were separated by scores on LV1 (**Figure 6.2D**), where the COPD subjects (red) had the most positive scores on LV1, the never smokers (purple) had the most negative scores on LV1, and the smokers (grey) fell in between the other two groups.

We next wanted to determine if the PLSDA model based on the combination of blood and BAL proteins was a better classifier of COPD disease state than data-driven models based on signatures of blood or BAL proteins alone. To illustrate this, we compared the calibration and cross-validation accuracies of our cross-compartment model with other feature-selected PLSDA models based on single tissue compartments: our model based only on 24 LASSO-identified, SOMAmer-measured blood proteins (**Figure D.2A** and **D.2B**), and our model based only on 18 VIP-selected, Luminex-measured BAL proteins (**Figure 6.2A** and **6.2B**). The model based on blood and BAL proteins combined trended towards having higher calibration accuracy than the model based on blood or BAL proteins alone (**Figure 6.2E**, one-way ANOVA with Tukey's post hoc test). Additionally, this combination model was found to be significantly better in terms of cross-validation accuracy than the blood model alone or the BAL model alone (**Figure 6.2F**, $p < 0.05$ for both comparisons, one-way ANOVA with Tukey's post hoc test), which indicates that the combination model might be able to handle unseen data better than the models based on proteins from one tissue compartment.

The database for annotation, visualization and integrated discovery (DAVID) determined that the positively loaded proteins on LV1 that were comparatively increased in COPD subjects and some smokers were significantly enriched for processes related to cytokine activity and the immune and defense response (**Figure 6.3**, enrichment score (ES) 1.94). Proteins that were negatively loaded on LV1 and comparatively increased in never smokers and in most of the smokers were enriched for processes involving the positive regulation of general cellular processes such as those related to metabolism, cell communication, signal transduction, and phosphorylation (**Figure D.3A**, ES 2.27), as well as regulation of the response to external stimuli (**Figure D.3B**, ES 1.92).



IL1B	CCL2/MCP-1 (BAL)	bone morphogenetic protein 6	EGF (BAL)	CCL11/Eotaxin (BAL)	CXCL9/MIG (BAL)	PDGF-C	hepcidin antimicrobial peptide	collagen type XVIII alpha 1 chain	TNF receptor superfamily member 10D	IL-22 (BAL)	troponin I3, cardiac type	troponin T2, cardiac type	SERPINA3	MMP1	carbonic anhydrase 3	Pathway	Bonferroni Corrected P-value
■	■	■	■	■	■	■	■	□	□	■	□	□	□	□	□	receptor binding	0.004509454
■	■	□	■	■	■	■	□	■	□	□	■	■	□	■	□	movement of cell or subcellular component	0.017207053
■	■	■	□	■	■	□	■	□	■	■	□	□	■	□	□	inflammatory response	6.87E-05
■	■	■	□	■	■	□	■	□	■	■	□	□	■	□	□	defense response	0.048109748
■	■	■	□	□	■	■	■	■	■	□	□	□	□	□	■	response to oxygen-containing compound	0.037465754
■	■	□	■	■	□	□	■	■	□	□	■	■	□	□	□	cardiovascular system development	0.021305318
■	■	□	■	■	□	□	■	■	□	□	■	■	□	□	□	circulatory system development	0.021305318
■	■	■	□	■	■	□	□	□	□	■	□	□	□	□	□	cytokine receptor binding	6.92E-04
■	■	■	□	■	■	□	□	□	□	■	□	□	□	□	□	cytokine activity	2.52E-04
■	■	■	□	□	■	□	■	□	■	□	□	□	□	□	□	response to lipopolysaccharide	0.011167374
■	■	■	□	□	■	□	■	□	■	□	□	□	□	□	□	response to molecule of bacterial origin	0.013899857
■	■	□	□	■	■	□	□	□	■	■	□	□	□	□	□	Cytokine-cytokine receptor interaction	0.005193892
■	□	□	□	□	□	□	■	□	□	■	□	□	■	□	□	acute-phase response	0.019372732

■ Indicates involvement in process □ Indicates no involvement in process

Figure 6.3. DAVID identified a cluster of significant pathways (Bonferroni corrected p < 0.05) involving cytokine activity and the inflammatory and defense response that was enriched in proteins that were comparatively increased in the COPD subjects in the LASSO-identified cross-tissue compartment signature. This cluster had an enrichment score (ES) of 2.57. Proteins found in the BAL in the signature are marked as so; unmarked proteins come from the blood samples.

Interestingly, when we examined the correlations between expression of the proteins in the blood and BAL combined signature within the smokers, never smokers, and COPD subjects separately, we found that the smokers and the COPD subjects had correlations that were weaker (lower Pearson's correlation coefficient value) than those seen in the never smokers but contained more hub proteins. The protein correlation network of the never smokers contained one protein that had 5 significant correlations to other proteins in the signature (**Figure 6.4A**).

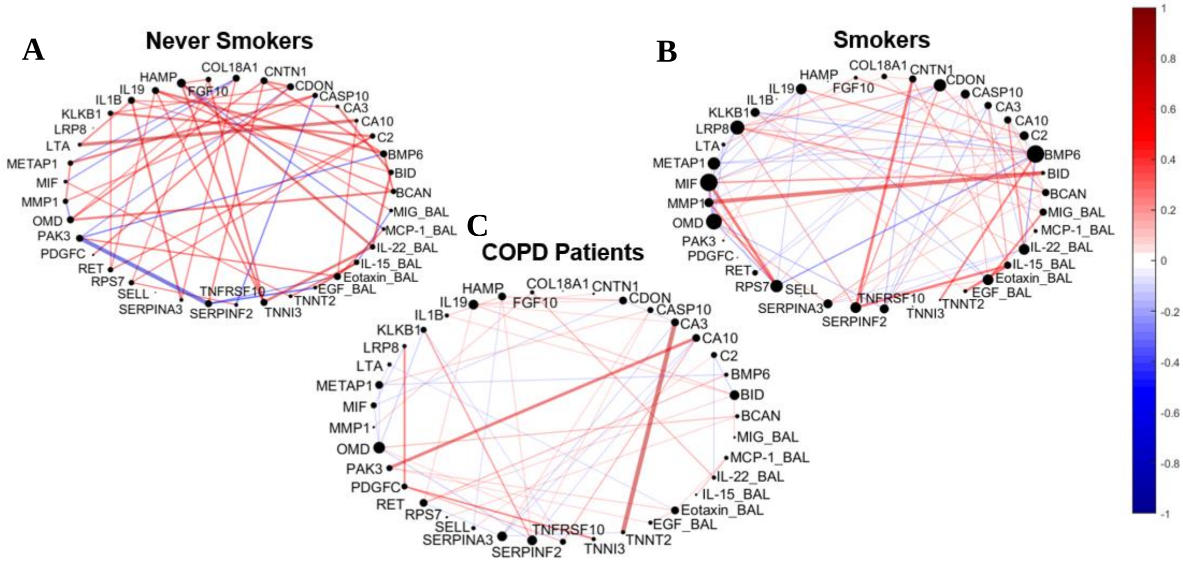


Figure 6.4. Protein correlation networks of the LASSO-identified blood and BAL protein signature present in never smokers (A), smokers (B), and COPD subjects (C) illustrate highly significant correlations and few hub proteins in the never smoker network.
Lines connecting two proteins indicate a significant correlation ($p < 0.05$) as determined by Pearson's correlation coefficient. Brighter and thicker lines indicate a stronger and more significant correlation, respectively, with color bar on the right displaying the value of the correlation coefficient. Red indicates a positive correlation.

However, the correlation network based on the signature proteins' expression in smokers had 17 proteins with 5 or more significant correlations (**Figure 6.4B**), and the network based on COPD subjects contained 5 proteins with 5 or more significant correlations (**Figure 6.4C**). Additionally, when comparing the strength of the connections in the networks, we saw that smokers had a larger number of significant correlations (77 correlations) than never smokers (53 correlations) and COPD subjects (52 correlations), but that the never smokers had correlations that were significantly stronger in terms of the absolute value of the Pearson's correlation coefficient than those present in the smoker and COPD subject networks ($p < 0.0001$ for both comparisons with the never smoker network, one-way ANOVA with Tukey's post hoc test).

We next used this approach to determine whether a combined signature of blood and BAL proteins was able to differentiate COPD patients with differing levels of disease severity according to the Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines to gain understanding into potential cross-tissue compartment proteomic relationships associated with disease severity. LASSO identified a signature of 38 SOMAmer-measured blood and 3 Luminex-measured BAL proteins that differentiated the three GOLD stages with 95.05% calibration and 78.76% cross-validation accuracy (**Figure 6.5A**). LV1 separated three stages of disease severity, with GOLD 1 subjects having negative scores on LV1, and GOLD 2 and GOLD 3 subjects having positive scores on LV1, with GOLD 3 having the most positive scores on LV1 (**Figure 6.5B**).

We next examined whether this signature based on blood and BAL proteins was better at differentiating the three GOLD stages than models based on blood or BAL proteins alone. We

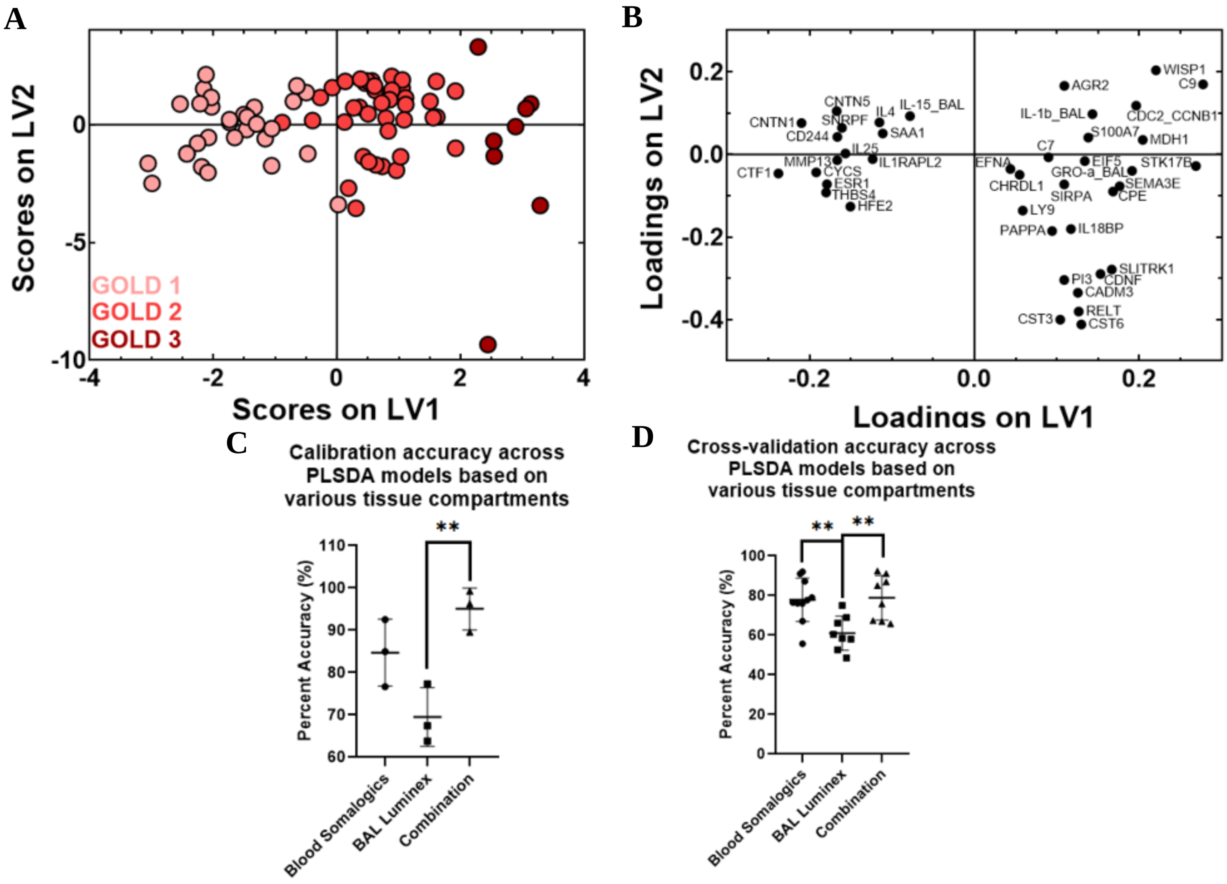


Figure 6.5. LASSO-identified signature of blood and BAL proteins differentiates GOLD status significantly better than model based on BAL proteins.

(**A**) A LASSO-identified signature of blood and BAL proteins was able to differentiate COPD subjects with GOLD 1, GOLD 2, and GOLD 3 disease severity classification with 95.05% calibration (Cal) and 78.76% cross-validation (CV) accuracy. (**B**) The loadings plot captured 7.29% of the variance in the data on latent variable 1 (LV1). (**C, D**) The PLSDA model based on a signature of blood and BAL proteins significantly outperformed a model based only on BAL proteins in terms of Cal (**C**) and CV (**D**) accuracy, and trended towards having higher accuracy than a model based on blood proteins alone. Significance in both cases calculated by a one-way ANOVA with Tukey's post hoc test; ** indicates p < 0.01

compared the calibration and cross-validation accuracies of the model based on blood and BAL proteins combined with a model based on a LASSO-identified signature of 29 SOMAmer-measured blood proteins (**Figure D.4A** and **D.4B**) and a model based on a VIP-selected signature of 13 BAL proteins (**Figure D.5A** and **D.5B**). Our model based on blood and BAL proteins combined had significantly higher calibration (**Figure 6.5C**) and cross-validation (**Figure 6.5D**) accuracy than the model based BAL proteins ($p < 0.01$, one-way ANOVA with Tukey's post hoc test for both comparisons), and trended towards having higher accuracies than the model based only on blood proteins.

6.2.2 PBMCs from smoker, never smoker, and COPD patients secrete distinct cytokine patterns in response to innate stimuli

We first looked at the differences in expression of 29 Luminex-measured proteins in response to stimulation with lipopolysaccharide (LPS) across PBMC cultures from smokers, never smokers, and COPD subjects using a one-way ANOVA with Tukey's post-hoc test. The only difference in expression that was found to be significant ($p < 0.05$) was between the never smoker and COPD subject expression of IP-10; all other comparisons were found to be not significant. This result highlights the individual differences in

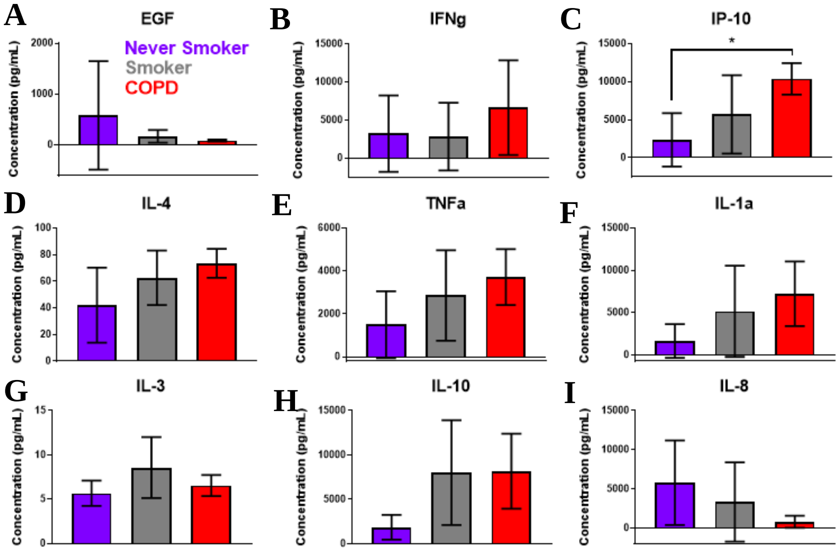


Figure 6.6. (A-I) Representative individual cytokine measurements from PBMCs of smokers with (red) and without (gray) COPD and healthy controls (purple) after in vitro stimulation with LPS. * $p < 0.05$.

response to LPS across the three disease states, which makes it difficult to generate testable hypotheses. The significant IP-10 result and a selection of some of the other non-significant results can be seen in **Figure 6.6**.

We then turned to data-driven modeling techniques to try to identify signatures of covarying proteins that could differentiate the PBMC responses to various immune stimuli across smoker, never smoker, and COPD subjects. In contrast to our univariate results, we saw a PLSDA model coupled with VIP feature selection led to a model that differentiated the PBMC responses of smokers (n = 7), never smokers (n = 6), and COPD subjects (n = 3) to a LPS stimulus with 98.72% calibration and 70.81% cross-validation accuracy (**Figure 6.7A**). The cross-validation accuracy may be slightly lower in this model due to the fact that there were so few COPD subjects, making it difficult to predict their secreted response in unseen cases. While we also saw similar results in the PLSDA models of the PBMC responses to other innate stimuli (the VIP-selected PLSDA model of stimulation with R848 had 88.19% calibration and 71.96% cross-validation accuracy and a model based on stimulation with Poly(I:C) had 88.76%

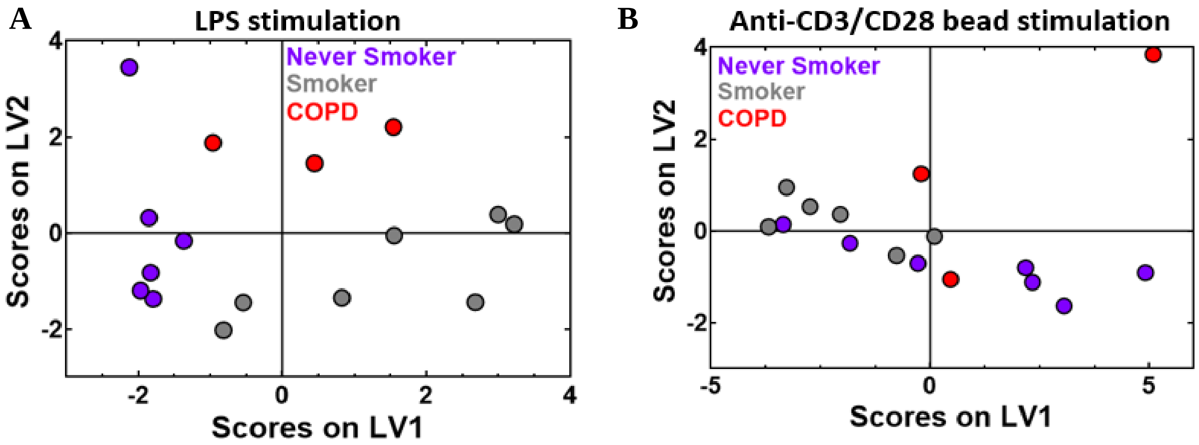


Figure 6.7. Data-driven modeling techniques are able to achieve better separation of the PBMC response of smokers, never smokers, and COPD subjects to innate immune stimuli as opposed to adaptive.
(**A**) The PLSDA scores plot shows the smokers (grey), never smokers (purple), and COPD subjects (red) clustering in different areas of the plot. The VIP-signature of the PBMC's secretome in response to an LPS stimulus performed with 98.72% calibration (Cal) and 70.81% cross-validation (CV) accuracy in a PLSDA model. (**B**) The PLSDA scores plot generated based on the PBMC's secreted response to a co-incubation with anti-CD3/CD28 beads does not separate the three clinical groups, performing only with 76.92% Cal and 59.33% CV accuracy.

calibration and 69.35% cross-validation accuracy), a VIP-selected PLSA model based on the adaptive stimulus of anti-CD3/CD28 beads only had 76.92% calibration and 59.33% cross-validation accuracy (**Figure 6.7B**).

6.3 Discussion

In this work, we used secreted proteins from *in vitro* cultures and proteomics data from a clinical trial to identify differentiating signatures and gain insight into potential mechanisms associated with COPD disease state and progression. Our preliminary analysis of immune cell communication networks in COPD suggests that the innate immune response of COPD PBMCs may be more different than that of smokers and never smokers, while the adaptive immune response of all three groups may be more similar. BAL cytokines were better at differentiating COPD disease state than models based only on plasma cytokine or blood protein data. However, we saw stronger differentiation of COPD disease state and GOLD status when signatures included data from multiple tissue compartments. Proteins that were comparatively increased in COPD subjects in the model of COPD disease state were enriched for cytokine activity and the immune and defense response. Additionally, in models of both COPD disease state and GOLD status, we reported that models based on data from multiple tissue compartments either trended towards or were significantly better in terms of calibration and/or cross-validation accuracy than models based on single tissue compartments. Overall, these results highlight the usefulness of data-driven models to take in a wide variety of clinically-relevant data and identify patterns that lead to increased insight into important factors involved in COPD disease state and progression.

We discovered that a model based on BAL cytokines was clearly better at differentiating COPD disease state than a model based on blood cytokines, and was also slightly more accurate than a model based on the blood SOMAmer measurements. This last point speaks to how much

COPD affects cytokine expression in the lung compared to protein expression in the blood, especially considering that only 25 cytokines were measured in the BAL, whereas 1305 blood proteins were measured by SOMAmers. The BAL model outperforming models based on blood proteins echoes results reported by Halper-Stromberg et al. in their analysis of the SPIROMICS cohort as well. They saw that there were a higher number of proteins measured in BAL samples than in plasma samples that were associated with variables such as FEV_1/FVC, emphysema, FEV_1 % predicted, and COPD exacerbations[274]. Additionally, we created a model based on sputum proteins that was able to differentiate stable and AE-COPD better than a model based on blood proteins (**Chapter 5**) in terms of calibration accuracy, but the two models had similar cross-validation accuracies. These results suggest that BAL cytokines may serve as better differentiators of COPD disease state than unstimulated proteins from the blood, especially when the number of measured blood proteins is low.

We identified that data-driven modeling techniques were able to identify differences in the responses of PBMCs from smokers, never smokers, and COPD subjects to innate immune stimuli, but not as much to adaptive immune stimuli. These models of cell signaling networks in PBMCs post-innate immune stimulation exhibited high calibration accuracy but only moderate cross-validation accuracy, which is most likely due to the low number of COPD subjects that were included in the analysis (n=3). These were promising results especially in comparison to the low calibration and cross-validation accuracy seen in the models of blood proteins in the SPIROMICS participants. It is especially promising that PBMC collection is a low-risk procedure for patients, but that we saw strong differentiation given our sample size. Although these preliminary results are suggestive of greater differences in the immune response of myeloid cells across smokers, never smokers, and COPD patients compared to the immune response of T cells, we could not dive deeper into the biological implications of any of our PBMC cell-cell signaling signatures due to the low sample size. However, other researchers have been able to investigate PBMC-based differences associated with COPD. Kawayama et al. stimulated PBMCs from smokers, never smokers, and COPD subjects with LPS and TNF-α, but did not report any significant differences in measured cytokines (MMP-9, TNF-α, IL-8, and IL-6) post stimulation[286]. They did notice that COPD PBMCs had the highest change in protein production post-stimulation, and suggested that PBMCs in COPD subjects are primed and ready to immediately respond to any immune stimuli[286]. Another study investigated differences in PBMC gene expression across smokers and COPD subjects and found that IL-16 mRNA levels were negatively correlated with upper lobe emphysema[287]. Taken together, this is suggestive of the usefulness of PBMC collection and study: either stimulation experiments or multi-omics analyses of these cells could help us identify, potentially understand, and, in the future, target cell signaling pathways that may be dysregulated in COPD.

Lastly, we have shown that combining data across multiple tissue compartments can lead to better differentiation when investigating both COPD disease state and progression as defined by GOLD guidelines. Both of our models either trended towards being or were significantly better in terms of calibration and cross-validation accuracy than models based on proteins from one tissue compartment. According to the prior knowledge database DAVID, the proteins comparatively increased in the COPD patients in the multiple tissue compartment signature of disease state were enriched for cytokine activity and the immune and defense response. Hogg et al. has reported that in a cohort of GOLD 0 – 4 subjects, the number of inflammatory immune cells (polymorphonuclear leukocytes, macrophages, CD4+ and CD8+ T cells, and B cells) in lung tissue samples were found at higher numbers in patients with higher GOLD classification[68].

The more inflammatory cells in the lung could possibly be an indicator for increased cytokine activity and immune response. These results highlight the importance of cytokine activity associated with COPD over smokers and never smokers that is independent of current smoking status. Based on protein correlation coefficient networks of the cross-tissue compartment signature in the smokers, never smokers, and COPD patients, we have hypothesized that the protein network in never smokers is more difficult to perturb. This is due to the never smoker network having a smaller number of hub proteins compared to the smoker and COPD network, and to it having highly significant correlations present within the network. Although there were no processes that were significantly enriched in the signature of COPD GOLD status according to DAVID, the signature did contain two complement proteins (complement 7 and 9) that were comparatively increased in GOLD 2 and 3. Complement proteins have been identified as being differentially expressed across COPD GOLD stages before: Baralla et al. reported that complement 4B was significantly decreased in GOLD 2 compared to GOLD 1 when comparing expression using two-dimensional gel electrophoresis[288]. Although these results suggest that complement system activity may be associated in some way with COPD GOLD status, our model of GOLD status was only based on GOLD 1-3, with only eight GOLD 3 patients being included. Thus these results need to be reconfirmed after the addition of more GOLD 3 patients to the model.

This analysis did not come without limitations that we have done our best to work around. The biggest limitation associated with all models was the lack of a validation cohort. However, we did perform cross-validation on all of our models and feature selection techniques whenever possible to prevent model overfitting the best that we could. So far, we have only explored BAL normalization to total protein albumin levels, but there are other options that we could explore as was discussed in **Chapter 4**. Additionally, in the PBMC communication network models and the models of AE-COPD based on sputum proteins, the sample size that we had access to was very small due to the former being a preliminary study and the latter experiencing difficulty in getting subjects to return to the VAAAHS or UMHS during an exacerbation. We recognize that the data we worked with was of a small size and have tried to focus our results and discussion of these results more on general trends within the models as opposed to generating specific hypotheses for mechanistic ways the signature features could be involved in differences between clinical groups. In the case of the PBMC network models, we are planning on following up on these results in a larger scale cohort of SPIROMICS II subjects with new funding that was recently obtained. Based on the success we have had with the PBMC models of cell-cell signaling so far and the ease of collection of these samples, it may be of interest to translate this system to study other diseases where chronic and dysregulated immune pathways may be at play.

6.4 Methods

6.4.1 Collection of biological samples from the SPIROMICS cohort

The Subpopulations and Intermediate Outcomes in COPD Study (SPIROMICS, ClinialTrials.gov Identifier: NCR01969344) is a multi-center, longitudinal study that was designed with overall goal of better understanding the disease in order to help inform the development of future treatment options. The study ended up enrolling subjects, a mix of smokers, never smokers, and mild/moderate and severe COPD subjects between the ages of 40-80 who met the lung function criteria for each recruited group without a diagnosis of non-COPD obstructive lung disease or unstable cardiovascular disease. Couper et al. described the complete details on the inclusion and exclusion criteria for the SPIROMICS study[253]. Enrolled subjects visited a participating center four times over three years for the collection of biological samples (blood, urine), lung function measurements, and questionnaires, and also received quarterly follow-up calls for the recording of exacerbation events and other health status updates. Portions of blood samples from one of these visits were sent to SomaLogic for the measurement of 1,305 proteins using their modified aptamer (SOMAmer) technology. Plasma from blood samples that were collected closest to the bronchoscopy visits were sent to the SPIROMICS Genomics and Informatics Coordinating Center (GIC) for storage at -80°C.

A subset of SPIROMICS participants (n = 215)[254] were also enrolled in the Bronchoscopy substudy. Inclusion criteria is described in detail by Freeman et al., and involved lung function spirometry results and smoking history[289]. Out of the COPD subjects recruited to the Bronchoscopy substudy, a much larger number had mild to moderate COPD as opposed to severe disease in an attempt to lessen the chance of an adverse event due to the bronchoscopy procedure. Enrollment in this study required two extra visits: one where induced sputum was collected, and a second visit where peripheral blood and BAL samples were collected. BAL was performed with sterile saline in the right middle lobe or lingula regions of the lung with 2 installations of 40 mL followed by 1 installation of 50 mL. Installations were pooled and spun down. Cell-free supernatants were aliquoted and stored until sent to the Arnold lab.

6.4.2 Measurement of proteins in BAL and plasma samples from the SPIROMICS cohort

Once collected, cell-free BAL samples were sent to the SPIROMICS Genetics and Informatics Coordinating Center (GIC) for storage at -80°C. The SPIROMICS GIC then sent matched plasma and BAL samples to the Arnold lab at the University of Michigan for protein measurements. Luminex FLEXMAP 3D technology was used to measure the concentration of 48 cytokines and chemokines in the BAL samples, and 47 cytokines and chemokines in the plasma samples. For protein measurements made by Luminex, we used a protocol that used ¼ of the recommended number of beads and sample volume to minimize the volumes necessasry for the assay, which was inspired by Arnold et al.[205] For the BAL samples, we ran them at twice the recommended volume due to low cytokine concentrations in these samples. Samples were run in duplicate, and the concentration of wells that were below the lower limit of detection were set to be equal to that of half of the lowest limit of detection of all cytokines.

6.4.3 Computational models of SPIROMICS patients

The first step in our computational analysis of the plasma and BAL proteins measured by us and the blood proteins measured by the SPIROMICS investigators was to identify proteins whose measurements were not different than the lower limit of detection, as well as samples that acted as negative drivers in each data-driven model. For the proteins, we removed proteins from analysis if more than 25% of the measurements were found to be below the lower limit of detection. None of the SOMAmer$^{©}$-measured proteins fell into this category, but 23 proteins measured in the BAL samples by Luminex and 9 proteins measured in the plasma samples by Luminex were removed before continuing on with our computational analysis. Overall, this meant that 1,305 SOMAmer-measured proteins, 25 BAL proteins, and 39 plasma proteins were initially included in models. We defined samples as being negative drivers of the model if they disproportionally drove our data-driven models such that the algorithm derived model parameters solely to account for that one sample. We quantitatively characterized these samples as those with a Hotelling's Reduced T^2 statistic value > 5 within a principal components analysis model (PCA) based on all measured proteins as calculated by the Eigenvector PLS Toolbox (Eigenvector, Mason, WA) software within MATLAB (MATLAB, Natick, MA). All protein data was mean centered and variance scaled before being used to build the PCA model. The sample with the highest Hotelling's Reduced T^2 statistic value that was > 5 was then removed, and a new PCA model was built based on the remaining samples. This process was repeated iteratively until all samples met that criteria. This resulted in the following models of COPD disease state: (1) A model based on SOMAmer-measured blood proteins alone that contained 47 never smokers, 102 smokers, and 121 COPD subjects; (2) A model based on BAL proteins alone that contained 25 never smokers, 75 smokers, and 82 COPD subjects; (3) A model based on Luminex-measured plasma protein alone that contained 25 never smokers, 74 smokers, and 84 COPD subjects; and (4) A model based on the combination of SOMAmer-measured blood and Luminex-measured BAL proteins together that contained 23 never smokers, 71 smokers, and 78 COPD subjects. When exploring differences in proteomic expression across GOLD status within the COPD subjects, this resulted in a model based on blood proteins which contained 45 GOLD 1, 56 GOLD 2, and 20 GOLD 3 subjects; a model based on BAL proteins alone which contained 32 GOLD 1, 44 GOLD 2, and 8 GOLD 3 subjects; and blood and BAL proteins combined which contained 30 GOLD 1, 40 GOLD 2 and 8 GOLD 3 subjects.

Once the negative drivers were removed, we moved onto identifying and visualizing proteomic signatures that could differentiate COPD disease state and GOLD status using feature selection techniques and partial least squares discriminant analysis (PLSDA). Again, all data were normalized via mean centering and variance scaling before any models were built or any feature selection was performed. Two different feature selection techniques were used: the least absolute shrinkage and selection operator (LASSO) was used for models that contained the blood protein measurements by SOMAmers$^{©}$ due to the large number of proteins that were measured, and VIP scores were used for models based on BAL or plasma proteins alone. For the LASSO models, k-fold cross-validation (k=10) was performed to generate the model with the lowest possible mean-squared error for prediction by iteratively excluding 10% of the samples during model training and then using this excluded data to test the model later. The batch number associated with the SomaLogic assay and current smoking status at enrollment of the SPIROMICS clinical trial were also included in each LASSO model with no penalty to correct for these differences across all clinical groups. The VIP score feature selection technique was not able to correct for confounding demographic or patient history factors. Proteins with VIP scores ≥ 1 were included in the final PLSDA model. All PLSDA models were additionally cross-validated using k-fold cross-validation (k=10). All final PLSDA models were also orthogonalized in order to improve interpretability.

Once the models were created, we then compared PLSDA model performance parameters to statistically say if one model was indeed better than others. To compare the calibration accuracy of multiple PLSDA models with each other, we took our final, cross-validated model and calculated the calibration accuracy for each defined class (e.g. smoker, never smoker, and COPD subject) by averaging the true positive rate and the true negative rate. We then took the calibration accuracies associated with each of these three classes from one model and statistically compared them to the accuracies present within the two other PLSDA models using a one-way ANOVA with Tukey's post hoc test, where $p < 0.05$ was deemed significant.

To compare the cross-validation accuracy of multiple PLSDA models with each other, we split the data into ten groups. We iteratively excluded one group and trained a PLSDA model on the remaining 9 groups, for a total of 10 PLSDA models. We tested the model using samples from the remaining group, and quantitatively defined the accuracy of the model again by judging how accurate the model was at classifying this unseen data. Specifically, we averaged the true positive rate and the true negative rate for each of the test set samples for each of the clinical groups (e.g. smoker, never smoker, and COPD subject). We averaged the accuracies of the three clinical groups within each of the ten PLSDA models to define the cross-validation accuracy associated with that particular test set. We then compared all ten calculated cross-validation accuracies from each of the PLSDA models, so all samples would serve within the test set once. We performed a similar calculation of cross-validation accuracies of PLSDA models based on other types of data, and finally used a one-way ANOVA with Tukey's post hoc test to compare calculated cross-validation accuracies across multiple models. P-values < 0.05 were deemed significant.

The database for annotation, visualization and integrated discovery (DAVID[165]) was used to help identify biological pathways that were significantly enriched among subsets of proteins in the LASSO-identified signature. Proteins in the signature were split into two groups based on the sign of their loading on LV1, and then run separately in DAVID. The resulting clustering and enrichment diagrams from DAVID were created by searching through Gene Ontology (GO) biological processes (BP FAT), GO molecular function (MF FAT), and the Kyoto Encyclopedia of Genes and Genomes (KEGG). For all analyses, only the clusters and pathways that were significant after the application of the Bonferroni correction were reported.

Protein correlation networks were created for smokers, never smokers, and COPD subjects separately. Pairwise Pearson's correlation coefficient was used to calculate the edges connecting the expression of two proteins in the LASSO-identified signature. The brightness and thickness of each edge indicate the value of the coefficient and the statistical significance of that correlation, respectively. Only significant ($p < 0.05$) correlations were plotted. Node size is proportional to its degree of connectedness.

6.4.4 Stimulation of peripheral blood mononuclear cells and measurement of secreted proteins

Practicing clinicians in Dr. Jeff Curtis's lab at the VA Ann Arbor Healthcare System (VAAAHS) collected 24 mL of whole blood from healthy subjects, smokers without airway obstruction, and smokers with airway obstruction (COPD subjects). Informed consent was obtained from each subject, and the blood collection protocol was approved by the VAAAHS IRB. PBMC isolation and stimulation was performed according to methods outlined by Arnold et al.[205] PBMCs were isolated from whole blood samples within one hour of collection via density centrifugation in Ficoll solution. Once isolated, cells were first counted before being stored at -80°C until stimulation.

On the day of stimulation experiments, PBMCs were thawed and resuspended in R10 media at a concentration of 20 million cells/mL. Cells were plated at a final concentration of 2 million cells/well in a 96-well U-bottom plate, in the presence of either a negative control (R10 media) or one immune stimulus. Investigated stimuli included R848 (stimulates TLR7 and TLR8; replicates a viral infection), LPS (stimulates TLR2 and TLR4; replicates a bacterial infection), CD3/CD28 dynabeads (stimulates the adaptive immune response), and Poly(I:C) (stimulates TLR3; replicates a viral infection). Cells were incubated for either 72 hours with the immune stimulus. Afterwards, adherent and nonadherent cells were collected and separated from the culture supernatant. Supernatant and cells were stored separately at -80°C until further analysis.

The concentrations of 29 cytokines and chemokines in the collected supernatants were measured using Luminex FLEXMAP 3D technology. For protein measurements in Luminex, we used a protocol that used ¼ of the recommended number of beads and sample volume to minimize the required volume for the assay, which was inspired by a protocol detailed in Arnold et al.[205] Samples were run in duplicate, and the concentration of wells that were below the lower limit of detection were set to be equal to half of the lowest limit of detection of all cytokines. Wells that were above the highest limit of detection were set to be the highest detectable concentration for that particular cytokine.

6.4.5 Computational models of cytokine secretions of stimulated PBMCs

Univariate analysis was performed by using a one-way ANOVA with Tukey's post hoc test to statistically compare the expression levels of the cytokines in the PBMC cultures from healthy subjects, smokers, and COPD subjects. Significance was defined as $p < 0.05$. Partial least squares discriminant analysis (PLSDA) was used to visualize signatures of covarying secreted cytokines that differentiated the healthy, smoker, and COPD PBMC response to various immune stimuli. Variable importance in projection (VIP) scores were used to identify protein signatures that were most important in differentiating the groups of interest. The final PLSDA models shown for these results are based only on proteins with VIP scores that were ≥ 1. All PLSDA models were cross-validated to prevent major model overfitting. K-fold cross-validation was performed by iteratively excluding ~8% of the samples from each model; this excluded data was then used to train the model. All PLSDA models based on VIP-selected features were orthogonalized to improve interpretability, and all data were mean centered and variance scaled before being used in PLSDA models.

The one-way ANOVA with Tukey's post hoc test was performed using GraphPad Prism (GraphPad Software, San Diego, CA). All PLSDA models and VIP score calculations were performed by the PLS toolbox in MATLAB (Eigenvector, Manson, WA).

Chapter 7 Overall Discussion

In this thesis, we have applied data-driven, systems biology-focused computational models to identify and explore signatures of covarying cells and proteins from multiple tissue compartments that successfully differentiated IPF and COPD disease state and progression. The specific conclusions and discussion of results will be presented according to the three aims of this work: using data-driven modeling tools to identify blood protein signatures, lung protein signatures, and multi-compartment and multi-assay signatures that could differentiate IPF and COPD disease state and disease progression.**Blood protein models of IPF and COPD disease state and progression introduce potential differentiating signatures in peripheral blood**

In work to support this aim, we were able to illustrate how data-driven approaches were effective for identifying blood protein signatures to differentiate individuals based on IPF disease state and progression status, as well as COPD exacerbation state.

One key result of this Aim illustrated that protein signatures were more useful than individual cytokines in differentiating clinical groups of interest. For example, we identified a signature of 61 blood proteins that outperformed previously published single markers of IPF progression. This signature also performed better than a previously published index of 6 proteins[136]. Likewise, in COPD a signature of 7 serum proteins was able to moderately differentiate stable and AE-COPD. This signature trended towards significantly outperforming the cross-validation of models based on single proteins that were differentially expressed across the two disease states, as well as a model based on all five proteins that were differentially expressed across the two groups. While these protein signatures hold promise for the development of prognostic assays, additional investigation and analysis will be required in the future to 1) reduce the number of proteins in the signature; 2) validate in new, larger cohorts; and 3) develop appropriate technology, as discussed below.

Results of this Aim also illustrated that in general, circulating cytokines were not useful for classification of groups (e.g. smokers and COPD subjects, IPF progressors and non-progressors), but that stronger differentiation could be achieved when either looking at a larger panel of proteins (e.g. SOMAscan assay) or when looking at distinctly different disease states. For example, in analysis of SPRIOMICS samples, when attempting to differentiate smokers, never smokers, and COPD subjects, plasma cytokines only performed with 65.69% calibration and 57.88% cross-validation accuracy, compared to the larger panel of SOMAmer-measured blood proteins, which performed with 74.56% calibration and 67.59% cross-validation accuracy. This indicated that it can be difficult for our modeling techniques to differentiate the systemic, unstimulated differences between a healthy group (never smokers) and two groups with worsening physiology who are much more similar to each other (smokers and COPD subjects), especially when only measuring a small number of proteins. Another example of the strength of a large panel of proteins can be seen in our models of the COMET IPF subjects: we saw strong separation both when the clinical groups being modeled were very distinct, like our model of 8 SOMAmer-measured proteins that differentiated healthy and IPF subjects, as well as when the clinical groups were similar, like our model of the 61 blood protein signature that differentiated IPF progressors and non-progressors.

Though results here suggest it may be difficult to differentiate clinical groups based on circulating cytokines, we have shown that stimulated systems of immune blood cell communication networks may hold more promise for a blood diagnostic. Our preliminary analysis of peripheral blood mononuclear cells (PBMCs) from small numbers (all n < 8) of smokers, never smokers, and COPD subjects illustrated that cytokine signatures secreted in response to innate immune stimuli (LPS, R848, Poly(I:C)) may be effective for differentiating these groups. Biologically, these results emphasize how much COPD and its associated airway obstruction affects the immune system, even at sites peripheral to the tissue compartment of injury. Due to low sample number (n = 3 COPD subjects), we were not able to explore these communication networks in depth, although the results suggested that differences associated with myeloid cell stimulation were greater across smokers, never smokers, and COPD subjects than those seen after T cell stimulation. However, we were encouraged to see that we were able to achieve high calibration accuracy and were able to visually separate all three patient groups while working with limited data. Additionally, studying these cell-cell communication networks can give more information about adverse immune responses present in COPD subjects that could inspire new ideas for experimental follow-up with potential therapeutic goals in mind. Overall, PBMC simulation experiments suggest a new paradigm for studying network level events that are able to differentiate clinical groups with progressive, immunological diseases.

Lastly, we also saw that data-driven modeling techniques were able to capture temporal changes in proteomic expression that were associated with disease progression. We identified a signature that was significantly different across three time points of IPF progressors using an unsupervised PCA model, but a similar signature was not identified in non-progressors. This result highlights the importance of measuring a wide variety of circulating, unstimulated proteins in order to obtain signatures that identify significant differences across clinical groups. Overall, these models showed that there are temporal changes in the peripheral blood proteome of IPF progressors that is not seen in non-progressors, and that our modeling techniques are able to capture these changes. In the future, it will be important to obtain access to separate validation cohorts to truly validate our signatures and PLSDA models, as discussed below.

Another challenge in using the IPF progression signature identified here as a prognostic signature is the large size of the signature (61 blood proteins). Proposing to create a prognostic test from this many factors would be difficult, as the test might take a longer time to process and may cost more, making it somewhat unattractive as a product. A general limitation of this work with the COMET cohort can be attributed to this study's demographics. All 60 IPF subjects that were recruited to COMET lived through the end of the 80-week study, which means that our reported trends and signatures are associated with mild to moderate IPF, but may not describe end-stage disease.

7.2 Lung protein models of IPF and COPD disease state and progression aid in hypothesis generation

Overall, results from this aim suggest that PLSDA models based on signatures of BAL lung cytokines are useful for gaining mechanistic insight and are better for classification than circulating, unstimulated blood cytokines or individual proteins. When we measured the same set of cytokines in matched plasma and BAL samples collected from subjects enrolled in the SPIROMICS study, a signature of BAL cytokines was better at differentiating COPD disease state (75.31% calibration and 71.76% cross-validation accuracy) than a signature of plasma cytokines (65.69% calibration and 57.88% cross-validation accuracy). We speculate this results from increased cytokine and chemokine activity in the COPD lung that is not apparent in blood measurements. In support of this hypothesis, another group has also reported that more BAL proteins that were associated with COPD-related variables than plasma proteins measured in the same subjects[274], although these protein measurements were made using untargeted liquid chromatography-mass spectrometry (LC-MS) technology, and not Luminex technology like we employed. To our knowledge, results from Aim 2 also represent the first time signatures of BAL cytokines were useful for differentiating IPF disease state. The identified signature was a better classifier than IFNa2 and IL-7, and classified with the same accuracy as analyses based on IL-15 and a combination model of all three proteins together. Overall, this suggests that signatures of cytokines may be better than individual cytokines in differentiating clinical groups, due to high inherent variability in the expression of single cytokines.

In this aim, we were able to gain mechanistic insight and generate hypotheses into lung-specific proteomic relationships associated with IPF and COPD disease state and progression. Generating biological insight into these lung-associated network relationships may be more useful than classification, as the invasive nature of BAL sample collection prevents widespread use for diagnostic or prognostic purposes. In IPF, a protein correlation coefficient network based on progressors' expression data indicated the proteins that had the most significant correlations to other proteins were chemokines (MCP-1, IL-8, GM-CSF), which suggested that cell trafficking into the lung may be associated with IPF progression. Additional examination of the other two hub proteins (EGF and G-CSF) also suggested that together these factors may be involved in immune cell recruitment to the lung and the associated tissue reorganization and fibrosis. In COPD exacerbation, we found that sputum IL-1β, IL-6, and C-reactive protein (CRP) were comparatively increased during exacerbation, suggesting pro-inflammatory functions. Based on these specific factors, we speculate that this exacerbation-associated inflammation may arise from macrophages[290]. The identified exacerbation signature also suggested a comparative increase IL-10, which could indicate attempted suppression of the inflammation present.

Lastly, we have also gained mechanistic insight by comparing the models of IPF and COPD progression with each other. When we compared the models of disease progression based on inflammatory cytokine and chemokine measurements from the lung environment, we saw that the PLSDA model of AE-COPD had more than 10% greater cross-validation and calibration accuracy than the model of IPF progression. Based on these results, we hypothesize that inflammatory cytokines may play a greater role in COPD progression than in IPF, which also aligns with current IPF pathogenesis hypotheses[291] and clinical trial results[47].

As in Aim 1, we also did not have access to a true validation cohort for these samples. Another limitation on nearly all the models of BAL proteins is that most were only moderately accurate at differentiating disease state or disease progression, with the exception of the healthy and IPF model based on BAL proteins. This could indicate a number of things: that BAL cytokines alone are not a strong classifier of lung disease state or progression and more proteins with a wider array of functions may need to be measured in each sample, that a different sample normalization technique should be used, or that more data need to be included to increase model accuracy. In the work presented here, all BAL protein data used as inputs into models and feature selection algorithms were first normalized to the total protein albumin concentration in the samples as calculated by a bicinchoninic acid (BCA) assay. We did see that models built on the BCA-normalized BAL data outperformed models based on raw data output from the Luminex assay (non-normalized models not shown). There is no field standard for BAL sample normalization, but we have not yet had the chance to explore other normalization techniques. However, we have already taken steps towards investigating some of these points. In the future, we plan on comparing BCA normalization of BAL samples collected in the SPIROMICS study with that of urea normalization[292], where the actual volume of the epithelial lining fluid (ELF) collected during the BAL procedure is estimated by assuming equal urea concentration in the ELF portion of the BAL and in the plasma of matched samples from the subjects. The limitation with the urea normalization method is that both BAL and plasma samples need to be collected during the bronchoscopy procedure, which requires foresight when designing the study and writing the methods. To investigate if BAL proteins are useful when combined with other data types, we have integrated data from multiple tissue compartments and assays into the same models; results from these models will be discussed in the following section.

7.3 Integrated blood and lung protein and cellular models of IPF and COPD disease state and progression lead to better classification and increased insight to mechanism

Results from Aim 3 across both IPF and COPD disease state and disease progression suggest that models based on protein data from multiple tissue compartments trend toward being or are significantly better at classification than models based on data from single tissue compartments. We identified a signature of 51 blood and 3 BAL proteins that differentiated IPF progressors and non-progressors with high accuracy. In addition to significantly outperforming nearly all models based on single or combination of proteins identified in univariate analyses, this model also had significantly better calibration and cross-validation accuracy than a model based on a signature of BAL cytokines. Likewise, we saw that signatures of blood and BAL proteins that differentiated COPD disease state had significantly better cross-validation accuracy than models based on blood or BAL proteins alone, and that a cross-tissue compartment model differentiating COPD GOLD status (a measure of disease severity) was significantly better in terms of calibration and cross-validation accuracy than models based on BAL proteins alone. These results suggest that the systemic and the pulmonary environments are both important to consider when trying to obtain the best differentiation between clinical groups, especially between those with subtle differences.

On top of performing with high accuracy, our models based on integrated signatures of blood and lung proteins and cellular markers have the potential to give new biological insight. Our approach is unique because by investigating and emphasizing the covariation between expressed proteins in clinical groups using computational data-driven modeling techniques, we can potentially begin to piece together larger networks of interactions that are present throughout the human body during disease. In our cross-tissue compartment model of IPF progression, we used a prior knowledge database (DAVID) and discovered that proteins that were comparatively increased in non-progressors were enriched for the regulation of the immune and defense system response. Additionally, we speculated that IPF non-progressors have greater control over their proteomic processes, and that this results in a network with few drivers that is difficult to perturb based on the low number of hub proteins in the non-progressor protein correlation network. We have hypothesized that the IPF progressors are a heterogeneous group, as seen by the correlation network with many hub proteins and less significant correlations, and that potential subgroups or endotypes of progressors may be identified by differences in proteomic expression.

We generated hypotheses for mechanisms associated with COPD disease state and progression with our cross-tissue compartment models as well. For our model of COPD disease state, we again used DAVID and discovered that the proteins that were comparatively increased in COPD subjects were enriched for cytokine activity and the immune and defense response, which could be related to the high levels of inflammation reported in COPD subjects in other studies[212,293,294]. Additionally, we created correlation networks based on the signature protein expression in smokers, never smokers, and COPD subjects separately. Similar to IPF progression results, we saw that the never smokers had stronger correlations compared to smokers and COPD subjects, and that the never smokers had fewer hub proteins than the other groups. This led to the hypothesis that never smokers have proteomic signaling networks that are more stable and difficult to perturb. A VIP-selected signature of serum and sputum proteins and blood cell markers was able to differentiate stable and AE-COPD, and also led to potential mechanistic insight into exacerbations. Based on the proteins and cells that were comparatively increased during exacerbation, we hypothesized that parallel increases in serum adhesion cytokines (sICAM-1 and sVCAM-1) and sputum inflammatory chemokines (MCP-2 and IP-10) are both critical to help inflammatory immune cells (such as CD15+ neutrophils) traffic into the lung during exacerbation. The percentage of CD4+ T cells was found to be comparatively increased in the stable state in this signature, and based on this and results from a previous study of this data[91], we also hypothesize that CD4+ T cells are some of the first cells to traffic to the lung during exacerbation. However, this result is then curious when compared to our preliminary analysis of cytokine secretions from stimulated PBMCs from smokers, never smokers, and COPD subjects in which we did not report strong differentiation after stimulating the T cells *in vitro*. This could speak to a potential change in T cell function associated with exacerbation, or this could be related to only looking at COPD as opposed to the two non-diseased groups. Overall, we speculate that this influx of immune cells to the lung during exacerbation helps create the inflammatory environment that is characteristic of AE-COPD events[295,296].

By comparing results from models based on IPF and COPD, we have reported some differences which should help us in planning experiments when moving forward with this work. Due to the larger number of SOMAmer-measured proteins, we expected blood proteins would dominate each cross-tissue compartment signature, even though IPF and COPD are lung-focused diseases. However, we noticed that when we compared the number of BAL proteins chosen in IPF and COPD models that there was a higher percentage of BAL proteins chosen in the models of COPD disease state (16.22% of the cross-tissue compartment signature was made up of BAL proteins) and GOLD status (7.31%) than in the model of IPF progression (5.56%). These results indicate that BAL cytokines may be more important in differentiating COPD disease state than IPF progression. We thus hypothesize that lung cytokines play a larger role in COPD than in IPF, a conclusion which is similar to that seen in Aim 2, and that in the future we should look into measuring more non-cytokine proteins in IPF BAL samples to see if this helps improve differentiation and mechanistic insight.

Unexpectedly, we did see some similar biological results in results across the COPD and IPF analyses. We reported similar trends in our protein correlation networks of IPF progression and COPD disease state, where the “sicker” groups (e.g. IPF progressors or the SPIROMICS smokers and COPD subjects) exhibited a large number of weakly significant correlations, whereas the “healthier” group (either IPF non-progressors or non-smoking controls for COPD) had protein networks that were characterized by correlations that were more significant. Additionally, in our models where BAL proteins were combined with SOMAmer-measured blood proteins, the LASSO feature selection technique almost always chose at least one complement protein as being one of the most important differentiating factors across IPF or COPD disease state or progression. This is curious because only 23 complement proteins were measured out of 1129 proteins total in the blood in the IPF samples, and 26 complement proteins out of 1305 were measured in the blood in the COPD samples. Specifically, inactivated complement component 3b was chosen in the blood protein signature that differentiated IPF disease state and was also involved in differentiating both IPF progressors and non-progressors in the trajectory PCA models. The complement 5b and 6 complex was chosen as part of the cross-tissue compartment signature that differentiated IPF progressors and non-progressors. In our models of COPD, complement 2 was comparatively increased in the never smokers in the cross-tissue compartment model of disease state, and complements 7 and 9 were associated with higher GOLD stage in the model of COPD progression. Various complement proteins (complement 4b[111,297], complement C1R[111]) have previously been reported as differentially expressed in healthy and IPF, although these studies show that higher expression of these proteins is not always consistently associated with one group. Higher complement 3 expression has been linked to the MUC5B promoter variant rs3570590 in humans with IPF[182]. Complement 3 (C3) and 4 (C4) have been reported to be decreased in the blood of COPD subjects compared to controls[298,299]. Sun et al. confirmed that lower C3 levels were associated with COPD and emphysema through models of protein quantitative trait loci (pQTL) and expression QTL (eQTL) SNPs in the SPIROMICS and COPDGene cohort, and suggested that the relationship between C3 protein levels and disease state may be mediated by genetic variants[300]. Overall, our data-driven modeling techniques have helped us generate biological hypotheses common to both IPF and COPD that deserve to be explored in the future.

Limitations associated with this analysis are similar to Aims 1 and 2 and involve the lack of a true validation cohort to test our models. Additionally, validation cohorts become difficult obtain when multiple omics analyses are performed on multiple samples from the same subjects. For example, to our knowledge, there is currently no cohort available that we could use to validate our model of AE-COPD events due to the need to have measured serum and sputum proteins (preferably by Luminex) and blood cell markers by flow cytometry. Additionally, many models presented in this section rely on proteomic data from BAL measurements, which requires an invasive procedure to obtain, though BAL is still a relatively low risk procedure[188]. Patient safety must and will always be considered before enrolling subjects in studies or collecting any BAL samples, but we have shown in this analysis that BAL samples do provide value. Our models that combined protein data from multiple tissue compartments performed with the highest accuracy, and integrated signatures have the potential to give new biological insight. By investigating the covariation between expressed proteins in clinical groups using computational data-driven modeling techniques, we can potentially begin to piece together larger interactions of networks between multiple organ systems that are present in the human body during disease, which could lead to a deeper understanding of disease state and disease progression.

7.4 Future work

As discussed above, one of the most important steps in the future of this work involves model validation in a separate cohort, especially if this approach is to be used to develop prognostic signatures. We did perform cross-validation during feature selection and model building whenever it was possible, but this does not replace a true validation cohort. This may be difficult due to the nature of the SomaLogic data: to our knowledge, currently the SomaLogic platform is not available for academic use, and we and other researchers have reported that SOMAmer-based measurements sometimes[136,301,302], but not always[130,197,301,302], correlate with antibody-based measurement techniques. Looking at cohorts that are currently available that could potentially be used to validate these models, Todd et al. recently published a study where they used multiple models to differentiate the SOMAmer-measured blood proteome of healthy and IPF subjects [111]. Data from this study could potentially be used for validation of our model of healthy and IPF, as long as it is confirmed that the IPF diagnosis process was the same in each study. Recently, COPDGene[264] investigators were able to send blood samples for measurement with the SOMAscan assay[301], making it a promising cohort for validation of our models of COPD disease state. However, models that differentiate disease state are not the most clinically relevant, as these groups are normally easy to tell apart without performing tests. To create a clinically relevant diagnostic signature, we would need access to SOMAmer-measured proteins from subjects with other lung diseases that are commonly misdiagnosed as IPF, such as other idiopathic interstitial pneumonias such as nonspecific interstitial pneumonia (NSIP)[291], or the immune disorder chronic hypersensitivity pneumonitis (HP)[303]. In IPF, it would be useful to obtain protein data from cohorts of HP or NSIP subjects for comparison.

To validate the IPF progression signature, we would need other IPF cohorts with SomaLogic data where progression could be tracked similarly as in the COMET study. The IPF-PRO study[304], which also had blood SOMAmer protein measurements collected[111], may be able to serve as a validation cohort; otherwise the PROFILE cohort study[170], which had blood Myriad RBM protein measurements collected[109], may also be able to serve as a validation cohort. For our models of cell-cell communication in COPD disease state, our lab will be able to perform the PBMC studies on a larger scale due to new collaborations with the SPIROMICS II visits, though validation of the original signatures presented here will probably not occur because they were created based on so few samples.

Once our signatures were validated in a separate cohort, there would be additional challenges associated with developing the assays that would be used in making diagnostic or prognostic decisions. This would include determining how sensitive the models are to the method of protein measurement (e.g. Aptamer vs. antibody), and if other systems besides the SomaLogic platform and Luminex technology could be used and still result in the same level of differentiation. This also involves identifying a technology that could be used to measure protein signatures vs. absolute cut-offs of individual cytokines. Additionally, we would need to determine how the test results would be used in decision-making: either directly by the clinician as an index with a cut-off for group classification[136], or by an outside company that would convey classification results to the clinicians. Steps following these decisions would then involve much more validation and eventually working with governmental agencies for approval.

In terms of moving forward in gaining insight into mechanism, one potential option is to move into animal models of disease. For IPF disease state research, the most common model of pulmonary fibrosis used is the bleomycin (BLM) murine model[305], and some researchers have developed a multi-BLM dose murine model that better models progression of pulmonary fibrosis[306]. Though these animal models do not capture all aspects of human IPF, they have been used extensively in the past to gain basic insight into IPF disease state and progression. The most common animal model used to study COPD involves exposing animals (dogs, guinea pigs, rats, mice, etc.) to high levels of cigarette smoke over a period of at least 3 to 6 months[307], but like the animal models used to study IPF, this model does not recapitulate all aspects of COPD. Researchers commonly administer bacterial (such as nontypeable *Hemophilus influenzae*[308] or LPS[309]) or viral infections[310] to model AE-COPD events in these animals. When moving into animal models of disease state and progression, first we would need to identify murine homologs of the human proteins in our signature and confirm the differentiating ability of these proteins in our groups of animals. Once we either reconfirmed the human signature in the animal models or identified animal-specific differentiating signatures, we could perform new experiments testing the importance of some of the higher loaded proteins in the PLSDA loadings plot or the hub proteins in the correlation networks by blocking signaling pathways downstream of that cytokine. We could then explore if these changes that were made caused the animals to cluster in a different area of the PLSDA scores plot or not. We could also use animal models to investigate cellular signatures that are associated with clinical groups due to the ease of sample collection. Moving into animal models of pulmonary fibrosis and COPD to investigate omics differences between corresponding clinical groups would allow for a more specific level of mechanistic exploration than what is possible in humans, although this comes at the price of then having to validate results in humans again later on.

There is still much to be done with samples that have currently been collected. To our knowledge, BAL samples collected during the COMET and SPIROMICS studies still exist. It could be useful to measure additional proteins in BAL samples using Luminex technology, as the SomaLogic platform is currently not available for academic use. For example, existing pre-mixed Luminex kits focused on the Th17 response could be intriguing to explore based on COPD endotype research that has been published[84,88]. Pre-mixed Luminex assays could also be useful for measurements of complement proteins including complement 3b/iC3b and complement 4, which would be of interest based on the large number of blood complement proteins that were chosen in our models. We were only able to measure 29 cytokines in the COMET samples, so gaining information about the concentration of a wider variety of signaling molecules could help increase the classification ability of our identified signatures. Measuring proteins with growth factor or tissue reorganization functions could be valuable in further evaluation of IPF, as evolving evidence from failed anti-inflammatory drug trials suggests that other factors may play a more central a role in disease natural history than cytokines alone[47,311,312]. However, multiple freeze-thaw cycles could make future measurements from these samples problematic, and must be taken into account before moving forward.

Additional analysis of data that has already been collected could also be of high value. One potential direction could be in identifying novel subgroups ("endotypes") within a disease state, and relying more on unsupervised analytical approaches that are focused on the diseased or progressing subpopulation alone. For example, it could be useful to create additional unsupervised models of IPF progressors to explore how these patients cluster without the non-progressors being present. I would want to explore if we continue to see the same three groups of progressors that we saw in the hierarchical cluster when using other unsupervised clustering algorithms, as well as applying supervised approaches to explore the biological mechanisms associated with each of the proteins increased in these groups using prior knowledge databases. Currently, COPD clinicians and researchers are focusing on two areas where I believe that unsupervised approaches could be of help: 1. The identification of potential COPD endotypes, and 2. Exploration and definition of potential differences in clustering of smokers without airway obstruction and COPD subjects. Based on interest in the early COPD disease state[262], it would be interesting to explore how smokers and COPD subjects cluster together in an unbiased way, and if any of these identified clusters are associated with clinical variables, such as number of pack years smoked, history of asthma or respiratory symptoms, or spirometry measurement ranges.

There are also new approaches, both computationally and in experimental design/sample collection, that could be taken in this area of pulmonary signature identification and should be explored. It was promising for us to have been able to build these cross-tissue compartment models of disease state and disease progression: we were able to computationally explore and define human proteomic and cellular relationships that are otherwise difficult to construct and study. While in this thesis we mostly focused on proteomic relationships that differentiated clinical groups, we have recently gained access to more omics data that were collected during the

SPIROMICS study that would be fascinating to integrate with our proteomics data to study more multi-omics mechanisms associated with disease state and progression. Specifically, there is transcriptomic data collected from epithelial cell brushings in the SPIROMICS participants, and in the future there will also be flow cytometry and microbiome measurements of the BAL samples as well. Incorporating all of these different types of data into a single model could give us a deeper understanding of COPD disease state and would be of great interest. We could approach all of these data types using the methods detailed in this thesis, or we could also explore other multi-omics integration and analysis tools. One methodology of interest includes multi-omics factor analysis (MOFA), which can be described as a versatile and generalized PCA analysis built to handle multi-omics data[313], and is better able to include a larger number of patient omic samples into models than PCA. When we performed PCA or PLSDA on our multi-omics or cross-tissue compartment data in this work, we were only able to include subjects who had successful measurements of all omics samples included in the model, which accounted for changes in the sample size used in our models of AE-COPD (**Chapter 3** and **5**). A MOFA model would have been able to include any subject that had at least one of the omics sample measurements. Another new approach that we could explore involves the type of samples we collect from subjects. We emphasized throughout this work the importance of collecting samples and proteomic data from the tissue compartment of interest, but we also recognize that these procedures are invasive to some extent, and are not always in the subject's best interest. However, some researchers have reported differences in the concentrations of cytokines measured in the exhaled breath condensate (EBC) that related to disease state in IPF[314] and COPD[315,316]. We would be curious to see what sort of proteomic measurements we could make from EBC samples using Luminex technology, and if these measurements translate into strong differentiating signatures. If we discover that cytokines are not easily detectable in EBC, switching our protein measurement approaches to those based on mass spectrometry or focusing on volatile organic compounds instead of cytokines[317] may lead to the generation of omics data that we could still model using our data-driven techniques. If so, that could mean that a much less invasive procedure could be performed that still allows for surveying of the lung environment. However, previous studies that focused on fatty acid[318] and 16S rRNA measurements[319] of EBC samples have shown how difficult it is to ensure that omics measurements of EBC samples are actually associated with a true biological signal, and thus it might be a safer move to first analyze previously collected EBC protein data before collecting new samples.

7.5 Conclusion

In conclusion, we have shown that we are able to identify proteomic and cellular signatures that can differentiate disease state and progression of IPF and COPD, and that these signatures are biologically relevant starting points for generating new hypotheses for mechanisms of action associated with disease and inspiring new directions for follow-up experiments. Our data-driven methods of signature identification may prove to be useful tools in identifying differentiating diagnostic and prognostic signatures that could hold clinical value if they are validated in separate cohorts. Ultimately, we hope that these signatures can be validated in human or murine models of IPF and COPD, enabling us to employ mechanistic models of the most important pathways and binding events to quantitatively investigate system perturbations and mechanistic hypotheses *in silico* in order to increase our understanding of IPF and COPD.

Appendices

APPENDIX A. Supplement to: The Peripheral Blood Proteome Signature of Idiopathic Pulmonary Fibrosis Is Distinct From Normal and Is Associated With Novel Immunological Processes

*David N. O'Dwyer[1†], Katy C. Norman[2†], Meng Xia[3], Stephen J. Gurczynski[1], Shanna L. Ashley[4], Eric S. White[1], Kevin J. Flaherty[1], Fernando J. Martinez[5], Susan Murray[3], Kelly B. Arnold[2#], and Bethany B. Moore[1,6#]

[1]Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA.

[2]Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI, USA.

[3]Biostatistics Department, University of Michigan School of Public Health, Ann Arbor, MI, USA.

[4]Immunology Graduate Program, University of Michigan, Ann Arbor, MI, USA.

[5]Department of Internal Medicine, Weill Cornell Medical College, New York, NY, USA.

[6]Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI, USA.

†Authors contributed equally; #Shared senior authorship

Scientific Reports. 2017; 7: 46560.

Table A.S1. Study population demographics

	IPF	Normal controls	P value*
Mean age in yrs. (SD)	64.56 (7.74)	69.97 (8.78)	0.0037
Male No. (%)	41 (68.33)	20 (66.67)	0.8733
Smoking status			
Never	19 (31.66)	N/A	_
Ex	40 (66.66)	N/A	_
Current	1 (1.66)	N/A	_

*Students t-test and Pearson χ2 squared test respectively. SD: standard deviation

Table A.S2. List of relevant co-morbidities and their frequencies with the COMET IPF patient cohort.

Co-morbidity	**Freq (N=60) (%)**
CAD	
Yes	7 (11.67)
No	53 (88.33)
MI	
Yes	2 (3.33)
No	58 (96.67)
Lung Cancer	
Yes	1 (1.67)
No	59 (98.33)
Other Cancer	
Yes	5 (8.33)
No	55 (91.67)
GERD	
Yes	34 (56.67)
No	25 (41.67)
Unknown	1 (1.67)
OSA	
Yes	12 (20.00)
No	48 (80.00)
Pulm HTN	
Yes	4 (6.67)
No	55 (91.67)
Unknown	1 (1.67)
Emphysema/Bronchitis	
Yes	1 (1.67)
No	58 (96.67)
Unknown	1 (1.67)

CAD-coronary artery disease: MI – myocardial infarction: GERD – gastroesophageal reflux disease: OSA – obstructive sleep apnea: Pulm HTN – pulmonary hypertension.

Table A.S3. List of upregulated proteins in the IPF peripheral proteome compared to control

Protein	UniProt ID	Gene ID
Afamin	P43652	AFM
Aflatoxin B1 aldehyde reductase member 2	O43488	AKR7A2
AH receptor-interacting protein	O00170	AIP
Alpha-soluble NSF attachment protein	P54920	NAPA
Aminoacylase-1	Q03154	ACY1
Apolipoprotein A-I	P02647	APOA1
Beta-Ala-His dipeptidase	Q96KN2	CNDP1
Bone morphogenetic protein 1	P13497	BMP1
C5a anaphylatoxin	P01031	C5
Cathepsin B	P07858	CTSB
cGMP-specific 3',5'-cyclic phosphodiesterase	O76074	PDE5A
Chloride intracellular channel protein 1	O00299	CLIC1
Coagulation Factor V	P12259	F5
Complement C1r subcomponent	P00736	C1R
Complement C4	P0C0L4	C4A
Cyclin-dependent kinase 8:Cyclin-C complex	P49336, P24863	CDK8 CCNC
Dual 3',5'-cyclic-AMP and -GMP phosphodiesterase 11A	Q9HCR9	PDE11A
Dual specificity mitogen-activated protein kinase kinase 4	P45985	MAP2K4
Endothelin-converting enzyme 1	P42892	ECE1
Fibronectin	P02751	FN1
Glyceraldehyde-3-phosphate dehydrogenase	P04406	GAPDH
Glycogen synthase kinase-3 alpha/beta	P49840, P49841	GSK3A GSK3B
Growth hormone receptor	P10912	GHR
Growth/differentiation factor 11	O95390	GDF11
GTP-binding nuclear protein Ran	P62826	RAN

Intercellular adhesion molecule 5	Q9UMF0	ICAM5
MAP kinase-activated protein kinase 2	P49137	MAPKAPK2
Matrilysin	P09237	MMP7
Methionine aminopeptidase 2	P50579	METAP2
Nascent polypeptide-associated complex subunit alpha	Q13765	NACA
Peptidyl-prolyl cis-trans isomerase D	Q08752	PPID
Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit alpha isoform: Phosphatidylinositol 3-kinase regulatory subunit alpha complex	P42336, P27986	PIK3CA PIK3R1
Plasma serine protease inhibitor	P05154	SERPINA5
Proprotein convertase subtilisin/kexin type 7	Q16549	PCSK7
Protein kinase C alpha type	P17252	PRKCA
Protein kinase C beta type (splice variant beta-II)	P05771	PRKCB
Proto-oncogene tyrosine-protein kinase Src	P12931	SRC
P-Selectin	P16109	SELP
Pyruvate kinase PKM	P14618	PKM2
Ras-related C3 botulinum toxin substrate 1	P63000	RAC1
Ribosome maturation protein SBDS	Q9Y3A5	SBDS
Small glutamine-rich tetratricopeptide repeat-containing protein alpha	O43765	SGTA
Sorting nexin-4	O95219	SNX4
Sphingosine kinase 1	Q9NYA1	SPHK1
Tumor necrosis factor ligand superfamily member 13B	Q9Y275	TNFSF13B
Tyrosine-protein kinase BTK	Q06187	BTK
Tyrosine-protein kinase CSK	P41240	CSK
Tyrosine-protein kinase Tec	P42680	TEC

Uniprot Accession ID listed. Detailed descriptions of proteins available at http://www.uniprot.org/
N=48 proteins
*Measured in combination

Table A.S4. List of downregulated proteins in the IPF peripheral proteome compared to control.

Protein	UniProt ID	Gene ID
Allograft inflammatory factor 1	P55008	AIF1
Alcohol dehydrogenase [NADP(+)]	P14550	AKR1A1
Alkaline phosphatase, tissue-nonspecific isozyme	P05186	ALPL
Annexin A1	P04083	ANXA1
Annexin A2	P07355	ANXA2
Complement C3	P01024	C3
Complement C3b, inactivated	P01024	C3
Complement C4b	P0C0L5	C4B
Carbonic anhydrase 3	P07451	CA3
Calcium/calmodulin-dependent protein kinase type II subunit beta	Q13554	CAMK2B
Calcium/calmodulin-dependent protein kinase type II subunit delta	Q13557	CAMK2D
Macrophage-capping protein	P40121	CAPG
Caspase-10	Q92851	CASP10
Calpastatin	P20810	CAST
C-C motif chemokine 14	Q16627	CCL14
C-C motif chemokine 23	P55773	CCL23
Cyclin-dependent kinase inhibitor 1B	P46527	CDKN1B
Cryptic protein	P0CG37	CFC1
Cofilin-1	P23528	CFL1
Chymase	P23946	CMA1
C-reactive protein	P02741	CRP
Macrophage colony-stimulating factor 1	P09603	CSF1
Granulocyte-macrophage colony-stimulating factor	P04141	CSF2
Cystatin-C	P01034	CST3
Cathepsin S	P25774	CTSS
C-X-C motif chemokine 11	O14625	CXCL11
Interleukin-8	P10145	CXCL8

Protein	UniProt ID	Gene ID
Discoidin domain-containing receptor 2	Q16832	DDR2
Eukaryotic translation initiation factor 4 gamma 2	P78344	EIF4G2
Eukaryotic translation initiation factor 5	P55010	EIF5
Eukaryotic translation initiation factor 5A-1	P63241	EIF5A
Ephrin type-A receptor 2	P29317	EPHA2
Tissue Factor	P13726	F3
Ficolin-1	O00602	FCN1
Tyrosine-protein kinase Fyn	P06241	FYN
Growth/differentiation factor 5	P43026	GDF5
Aspartate aminotransferase, cytoplasmic	P17174	GOT1
Glucose-6-phosphate isomerase	P06744	GPI
Glutathione S-transferase P	P09211	GSTP1
Histone H2A.z	P0C0S5	H2AFZ
Hepatitis A virus cellular receptor 2	Q8TDQ0	HAVCR2
Hepatoma-derived growth factor-related protein 2	Q7Z4V5	HDGFRP2
Histone H1.2	P16403	HIST1H1C
High mobility group protein B1	P09429	HMGB1
Heme oxygenase 2	P30519	HMOX2
Heterogeneous nuclear ribonucleoproteins A2/B1	P22626	HNRNPA2B1
Heterogeneous nuclear ribonucleoprotein A/B	Q99729	HNRNPAB
Estradiol 17-beta-dehydrogenase 1	P14061	HSD17B1
Heat shock 70 kDa protein 1A/1B	P08107	HSPA1A
Serine protease HTRA2, mitochondrial	O43464	HTRA2
ICOS ligand	O75144	ICOSLG
Insulin-like growth factor-binding protein 1	P08833	IGFBP1
Insulin-like growth factor-binding protein 2	P18065	IGFBP2
Interleukin-16	Q14005	IL16
Interleukin-2	P60568	IL2

Protein	UniProt ID	Gene ID
Interleukin-3	P08700	IL3
Integrin alpha-I: beta-1 complex	P56199, P05556	ITGA1 ITGB1
Killer cell immunoglobulin-like receptor 2DL4	Q99706	KIR2DL4
Importin subunit alpha-1	P52292	KPNA2
Lipopolysaccharide-binding protein	P18428	LBP
Neutrophil gelatinase-associated lipocalin	P80188	LCN2
Lactotransferrin	P02788	LTF
Dual specificity mitogen-activated protein kinase kinase 1	Q02750	MAP2K1
Dual specificity mitogen-activated protein kinase kinase 2	P36507	MAP2K2
Mitogen-activated protein kinase 13	O15264	MAPK13
Myoglobin	P02144	MB
Matrix metalloproteinase-9	P14780	MMP9
Myeloperoxidase	P05164	MPO
Moesin	P26038	MSN
Nicotinamide phosphoribosyltransferase	P43490	NAMPT
NudC domain-containing protein 3	Q8IVD9	NUDCD3
Oxidized low-density lipoprotein receptor 1	P78380	OLR1
Protein DJ-1	Q99497	PARK7
Phosphatidylethanolamine-binding protein 1	P30086	PEBP1
Prefoldin subunit 5	Q99471	PFDN5
Phosphoglycerate mutase 1	P18669	PGAM1
Peptidoglycan recognition protein 1	O75594	PGLYRP1
Elafin	P19957	PI3
Phospholipase A2, membrane associated	P14555	PLA2G2A
Urokinase plasminogen activator surface receptor	Q03405	PLAUR
NADPH--cytochrome P450 reductase	P16435	POR
Myeloblastin	P24158	PRTN3
Proteasome subunit alpha type-2	P25787	PSMA2

Protein	UniProt ID	Gene ID
Prostaglandin G/H synthase 2	P35354	PTGS2
Tyrosine-protein phosphatase non-receptor type 1	P18031	PTPN1
Tyrosine-protein phosphatase non-receptor type 11	Q06124	PTPN11
Tyrosine-protein phosphatase non-receptor type 6	P29350	PTPN6
RNA-binding protein 39	Q14498	RBM39
Resistin	Q9HD89	RETN
Ubiquitin	P62979	RPS27A
Ubiquitin+1, truncated mutation for UbB	P62979	RPS27A
40S ribosomal protein S7	P62081	RPS7
Protein S100-A9	P06702	S100A9
Serum amyloid A-1 protein	P0DJI8	SAA1
Scavenger receptor class F member 1	Q14162	SCARF1
alpha-1-antichymotrypsin complex	P01011	SERPINA3
Plasma protease C1 inhibitor	P05155	SERPING1
Pulmonary surfactant-associated protein D	P35247	SFTPD
SHC-transforming protein 1	P29353	SHC1
Sialic acid-binding Ig-like lectin 14	Q08ET2	SIGLEC14
Small nuclear ribonucleoprotein F	P62306	SNRPF
FACT complex subunit SSRP1	Q08945	SSRP1
Heterogeneous nuclear ribonucleoprotein Q	O60506	SYNCRIP
Trefoil factor 3	Q07654	TFF3
Metalloproteinase inhibitor 1	P01033	TIMP1
Tumor necrosis factor receptor superfamily member 1B	P20333	TNFRSF1B
Tumor necrosis factor ligand superfamily member 14	O43557	TNFSF14

DNA topoisomerase 1	P11387	TOP1
Triosephosphate isomerase	P60174	TPI1
SUMO-conjugating enzyme UBC9	P63279	UBE2I
Ubiquitin-conjugating enzyme E2 N	P61088	UBE2N
Ubiquitin-fold modifier 1	P61960	UFM1
Vacuolar protein sorting-associated protein VTA1 homolog	Q9NP79	VTA1
X-ray repair cross-complementing protein 6	P12956	XRCC6
Tyrosine-protein kinase Yes	P07947	YES1

* P08107 updated as secondary accession to P0DMV8/P0DMV9 (HSPA1A/HSPA1B).
N= 116

Table A.S5. List of all significant proteins in analysis of IPF proteome versus healthy. Color code information at bottom.

Uniprot ID	Color Code	Gene ID	Fold expression	Predicted Linear Model Ratios (IPF/Healthy)	P-values significant after Bonferroni correction	FDR	Age-adjusted P-values significant after Bonferroni Correction	Age-adjusted FDR
P17252	🟥	PRKCA	2.844150745	2.7443155	1.36E-14	0.000194363	4.3293E-14	0.001000972
P02647	🟥	APOA1	1.542341061	1.544871222	1.451E-06	0.001467444	4.2806E-09	0.001282799
O95390	🟥	GDF11	1.628565596	1.608156622	7.663E-12	0.000359572	1.1845E-07	0.000855199
P09237	🟥	MMP7	2.654602809	2.747872619	2.054E-18	0.000126336	2.5766E-10	0.001720117
P01031	🟥	C5	1.654531862	1.675076486	5.828E-07	0.001282799	7.3906E-06	0.001681244
P63000	🟥	RAC1	1.610302077	1.574044262	1.098E-06	0.001389699	6.2295E-06	0.00191448
Q9Y275	🟥	TNFSF13B	2.129450761	2.191299147	3.817E-19	9.71817E-05	4.4978E-11	0.000272109
P07858	🟥	CTSB	1.665079978	1.742647874	3.293E-13	0.000242954	3.113E-11	0.000437318
P50579	🟥	METAP2	2.400193949	2.369123805	4.876E-10	0.000515063	2.655E-08	0.002147716
P00736	🟥	C1R	4.387258094	4.463494883	7.042E-22	5.8309E-05	6.7295E-17	0.001574344
P13497	🟥	BMP1	2.462088383	2.418442708	4.891E-16	0.000165209	4.7633E-14	0.002196307
P49336, P24863	🟥	CDK8 CCNC	1.642587894	1.64618749	5.658E-19	0.000116618	8.7935E-11	0.00292517
P41240	🟥	CSK	1.955228256	1.939577175	5.987E-10	0.000524781	2.9947E-06	0.001137026
P05154	🟥	SERPINA5	2.493816792	2.438316738	4.487E-18	0.000136054	2.0621E-20	0.001525753
P49840, P49841	🟥	GSK3A GSK3B	3.734712707	3.634606001	3.89E-27	9.71817E-06	6.23E-18	0.001389699
P42680	🟥	TEC	2.558122735	2.478830736	2.333E-16	0.000145773	2.3177E-08	0.001156463
P49137	🟥	MAPKAPK2	1.868501241	1.883819574	4.897E-08	0.00090379	1.7658E-06	0.002400389
O00170	🟥	AIP	1.715314482	1.675073096	9.679E-15	0.000174927	6.676E-09	0.000719145
P04406	🟥	GAPDH	1.669919932	1.639169965	5.773E-06	0.001856171	5.3425E-07	0.000942663
Q13765	🟥	NACA	2.276481926	2.165538354	3.917E-21	7.77454E-05	6.6077E-12	0.00047619
O43765	🟥	SGTA	1.859855557	1.803317323	1.199E-08	0.000767736	2.8969E-07	0.002001944
O95219	🟥	SNX4	2.060637667	1.939104225	7.436E-09	0.000709427	1.9616E-08	0.002137998
P02751	🟥	FN1	2.058884999	2.1387836	5.309E-12	0.000340136	2.7176E-08	0.001059281
O43488	🟥	AKR7A2	2.382612119	2.339107573	2.438E-11	0.000388727	1.7054E-07	0.00159378
P14618	🟥	PKM2	2.372734125	2.397160175	2.892E-13	0.000233236	5.4956E-09	0.002254616
P62826	🟥	RAN	10.78442484	8.501570061	9.529E-21	8.74636E-05	3.8826E-11	0.001467444
P54920	🟥	NAPA	2.259133786	2.144386113	1.622E-13	0.000204082	1.6544E-09	0.002983479
Q9NYA1	🟥	SPHK1	4.924705886	4.492447031	2.931E-21	6.80272E-05	1.0608E-13	0.000359572

Uniprot ID	Color Code	Gene ID	Fold expression	Predicted Linear Model Ratios (IPF/Healthy)	P-values significant after Bonferroni correction	FDR	Age-adjusted P-values significant after Bonferroni Correction	Age-adjusted FDR
Q16549	🟥	PCSK7	2.06801082	1.986426465	5.236E-22	4.85909E-05	6.9419E-12	0.000291545
P12259	🟥	F5	1.568462646	1.611038214	1.518E-11	0.000369291	1.0352E-06	0.000349854
O00299	🟥	CLIC1	3.059946004	2.993311375	6.309E-12	0.000349854	6.6552E-09	2.91545E-05
Q9UMF0	🟥	ICAM5	1.759344079	1.804719167	2.014E-11	0.000379009	1.3082E-07	0.000281827
Q08752	🟥	PPID	3.742581674	3.511759899	1.307E-12	0.000281827	3.2523E-06	0.0003207
P45985	🟥	MAP2K4	1.747910459	1.715622054	4.992E-09	0.000680272	9.5677E-08	0.000330418
P12931	🟥	SRC	3.854282773	3.709558785	1.198E-22	3.88727E-05	1.027E-18	0.000252672
O76074	🟥	PDE5A	4.43765169	3.978480542	9.751E-26	1.94363E-05	8.5254E-15	0.000447036
P10912	🟥	GHR	1.777326642	1.767808273	1.284E-10	0.000447036	2.1388E-10	0.000728863
Q03154	🟥	ACY1	4.904977655	4.13058665	6.462E-13	0.000272109	4.2266E-06	0.000155491
P42336, P27986	🟥	PIK3CA PIK3R1	1.595594423	1.612386386	5.423E-19	0.0001069	6.2975E-11	0.001972789
Q06187	🟥	BTK	10.44813268	9.363373991	2.252E-24	2.91545E-05	8.0073E-14	0.001068999
Q96KN2	🟥	CNDP1	2.772175242	2.868771762	5.328E-13	0.000262391	5.454E-10	0.001127308
Q9Y3A5	🟥	SBDS	2.559809181	2.525581685	1.145E-14	0.000184645	8.1077E-10	0.000612245
P16109	🟥	SELP	1.755253189	1.688921305	5.013E-11	0.000398445	4.4751E-09	0.001292517
P08649	🟥	C4A C4B	2.163849532	2.152620093	2.106E-13	0.000223518	1.7857E-12	0.000894072
P43652	🟥	AFM	1.516790465	1.509684337	4.459E-09	0.000660836	1.8907E-10	0.001613217
Q9HCR9	🟥	PDE11A	2.693926113	2.629765299	4.46E-10	0.000505345	4.4336E-07	0.000262391
P05771	🟥	PRKCB	2.549572043	2.527742077	2.072E-13	0.0002138	1.944E-07	0.000621963
P42892	🟥	ECE1	1.496922719	1.545032257	2.628E-08	0.000864917	1.3217E-07	0.002439261
Q9UHD0	▨	IL19	1.536009374	1.522027626	6.889E-08	0.000942663	2.5591E-05	0.001146744
P23280	▨	CA6	2.120863591	2.199196968	8.504E-07	0.001350826	9.6872E-05	0.000242954
Q8N1Q1	▨	CA13	2.14243776	2.090315543	3.449E-08	0.000874636	0.00011986	0.001428571
P07996	▨	THBS1	1.722022566	1.66082076	2.042E-06	0.001564626	0.00124037	0.002954325
Q9NQU5	▨	PAK6	1.78765298	1.782944743	9.87E-07	0.001379981	3.1333E-05	0.001049563
Q8N5S9	▨	CAMKK1	1.689904266	1.68676957	9.496E-07	0.001370262	0.00133575	0.001477162
Q99714	▨	HSD17B10	2.648891087	2.544580128	1.222E-07	0.001039845	0.00027562	0.001234208
Q08209, P63098	▨	PPP3CA PPP3R1	1.996632503	2.050362653	2.685E-07	0.001175899	3.7241E-05	3.88727E-05
P03956	▨	MMP1	1.880033667	1.868914014	6.063E-08	0.000932945	9.6125E-05	0.001448008
P15514	▨	AREG	1.597172106	1.537410827	1.067E-07	0.000991254	6.0986E-05	0.000699708

Uniprot ID	Color Code	Gene ID	Fold expression	Predicted Linear Model Ratios (IPF/Healthy)	P-values significant after Bonferroni correction	FDR	Age-adjusted P-values significant after Bonferroni Correction	Age-adjusted FDR
P36888	🟥	FLT3	1.633132296	1.62729439	1.988E-07	0.00111759	0.00026356	0.000602527
P12277, P06732	🟥	CKB CKM	3.424539483	2.772971246	9.364E-09	0.000748299	0.00031418	0.001341108
O75636	🟥	FCN3	1.486624762	1.505689722	4.105E-06	0.001739553	0.00219711	0.00877551
O43557	🟦	TNFSF14	0.437387768	0.433556681	2.631E-09	0.000612245	3.9442E-15	0.001098154
P62306	🟦	SNRPF	0.589209938	0.585044542	3.073E-09	0.000631681	2.0066E-16	0.001671526
P18065	🟦	IGFBP2	0.561250316	0.560919529	1.127E-06	0.001399417	8.7262E-10	0.000641399
P14780	🟦	MMP9	0.580646533	0.581799239	8.175E-06	0.001963071	6.1719E-07	0.000340136
P05164	🟦	MPO	0.66653812	0.664052177	3.288E-07	0.001195335	3.5821E-06	0.000116618
P55010	🟦	EIF5	0.264516494	0.257394258	1.575E-10	0.000466472	2.9209E-26	0.00068999
P30519	🟦	HMOX2	0.302673799	0.309421198	1.115E-08	0.000758017	2.4311E-15	0.00079689
Q03405	🟦	PLAUR	0.714503307	0.711668768	3.972E-06	0.001729835	6.9874E-07	0.000553936
P01024	🟦	C3	0.531153018	0.521491138	3.962E-12	0.0003207	6.326E-28	0.000631681
P16435	🟦	POR	0.193507136	0.203838807	1.288E-07	0.001049563	1.2994E-16	0.001554908
P43026	🟦	GDF5	0.527925041	0.526377553	1.399E-07	0.001098154	5.7904E-16	0.000369291
P01024	🟦	C3	0.594138865	0.588274349	2.357E-08	0.000855199	7.0592E-16	0.000631681
Q14005	🟦	IL16	0.23033519	0.237400985	4.271E-09	0.000651118	2.2301E-17	0.000660836
P12956	🟦	XRCC6	0.179825644	0.192845513	9.608E-08	0.000971817	4.8595E-14	0.001758989
P55008	🟦	AIF1	0.344025162	0.34967201	5.447E-08	0.000913508	7.8303E-12	6.80272E-05
P52292	🟦	KPNA2	0.576832085	0.585079807	5.707E-09	0.00068999	1.2065E-17	0.001166181
Q02750	🟦	MAP2K1	0.359077237	0.362158859	4.702E-07	0.001234208	8.6888E-13	8.74636E-05
P11387	🟦	TOP1	0.161052587	0.172640307	7.554E-06	0.001943635	2.8135E-13	0.000544218
P63279	🟦	UBE2I	0.402202186	0.406545614	1.118E-07	0.001000972	9.1848E-12	0.000913508
P07947	🟦	YES1	0.530683894	0.535492527	5.545E-08	0.000923226	1.0937E-15	0.000515063
P16403	🟦	HIST1H1C	0.18202697	0.179435862	2.033E-12	0.000301263	1.1917E-21	0.000991254
P18031	🟦	PTPN1	0.513454609	0.518122474	4.335E-06	0.001749271	2.7091E-11	0.000145773
P20810	🟦	CAST	0.625591416	0.647812276	1.392E-10	0.000456754	1.1152E-10	0.000466472
P02144	🟦	MB	0.444987418	0.457379403	1.781E-06	0.001516035	1.3409E-12	0.000680272
P60568	🟦	IL2	0.602649926	0.607874718	8.102E-07	0.001341108	8.832E-12	7.77454E-05
P25774	🟦	CTSS	0.691366271	0.70449585	1.438E-06	0.001457726	2.5524E-06	1.94363E-05
P0CG37	🟦	CFC1	0.480568112	0.480547449	1.329E-07	0.001078717	6.6844E-15	9.71817E-05

Uniprot ID	Color Code	Gene ID	Fold expression	Predicted Linear Model Ratios (IPF/Healthy)	P-values significant after Bonferroni correction	FDR	Age-adjusted P-values significant after Bonferroni Correction	Age-adjusted FDR
O43464	🟦	HTRA2	0.727295489	0.719331748	7.34E-07	0.001321672	2.8688E-09	0.000485909
O75594	🟦	PGLYRP1	0.385704603	0.406150492	1.661E-06	0.001506317	1.0747E-11	0.000126336
Q13554	🟦	CAMK2B	0.448493259	0.46201085	3.863E-06	0.001720117	8.3106E-06	0.000651118
Q06124	🟦	PTPN11	0.325170474	0.320946853	7.107E-11	0.000417881	7.5569E-24	0.00133139
Q13557	🟦	CAMK2D	0.448876214	0.457971263	1.144E-06	0.001409135	3.2017E-08	0.000923226
P10145	🟦	CXCL8	0.034225488	0.036256866	3.956E-08	0.000894072	7.9872E-23	0.000194363
P56199, P05556	🟦	ITGA1 ITGB1	0.298316266	0.315384647	5.57E-07	0.001263362	1.0694E-12	5.8309E-05
P24158	🟦	PRTN3	0.241989732	0.250856356	2.183E-06	0.001584062	1.1768E-09	0.00058309
Q16832	🟦	DDR2	0.59816339	0.593148426	1.398E-06	0.001448008	8.3643E-13	0.001088435
O00602	🟦	FCN1	0.487691807	0.491252641	4.684E-06	0.001788144	2.947E-09	0.000388727
P36507	🟦	MAP2K2	0.399325508	0.409209093	3.71E-09	0.000641399	1.0395E-20	0.000301263
P46527	🟦	CDKN1B	0.390461304	0.389275226	1.256E-09	0.000592809	8.6472E-14	0.001205053
P09603	🟦	CSF1	0.349185705	0.345595517	5.652E-06	0.001846453	2.4947E-11	0.001253644
P35354	🟦	PTGS2	0.300945517	0.304112877	8.668E-10	0.000563654	3.2731E-23	4.85909E-05
P06241	🟦	FYN	0.15968612	0.158580485	9.413E-11	0.0004276	3.0673E-28	0.001039845
P62081	🟦	RPS7	0.185649796	0.187803316	6.968E-10	0.000553936	4.3763E-22	0.0005345
P18669	🟦	PGAM1	0.380200082	0.367406052	2.737E-10	0.000485909	2.1317E-13	9.71817E-06
P08107	🟦	HSPA1A	0.329230499	0.325625349	3.616E-13	0.000252672	5.5044E-21	0.000184645
P01011	🟦	SERPINA3	0.457669471	0.44651438	5.712E-07	0.001273081	9.634E-12	0.000670554
P14550	🟦	AKR1A1	0.591448395	0.596031761	3.696E-06	0.001690962	1.6211E-07	0.000505345
P23528	🟦	CFL1	0.669498441	0.658275648	7.5E-06	0.001933916	1.3489E-09	0.000204082
O60506	🟦	SYNCRIP	0.291191938	0.297780019	9.354E-08	0.000962099	4.8968E-20	0.0001069
Q99471	🟦	PFDN5	0.254728054	0.25020284	6.318E-09	0.000699708	1.7807E-24	0.000563654
P06744	🟦	GPI	0.529843615	0.523601892	1.48E-08	0.00079689	2.6058E-12	0.001933916
P30086	🟦	PEBP1	0.628052496	0.616272435	4.374E-06	0.001758989	1.2629E-08	0.000738581
Q14498	🟦	RBM39	0.17003682	0.17844853	1.313E-07	0.001059281	1.2501E-14	0.000592809
P29350	🟦	PTPN6	0.339639104	0.338090817	9.057E-10	0.000573372	2.0333E-16	0.000408163
P02741	🟦	CRP	0.625041071	0.602656298	1.204E-06	0.001418853	5.8331E-08	0.000233236
Q9UIK4	🟦	DAPK2	0.400809219	0.401459963	2.019E-08	0.000826045	1.0825E-07	0.000952381
P35247	🟦	SFTPD	0.226428987	0.220331959	4.803E-12	0.000330418	1.8116E-25	0.000767736

Uniprot ID	Color Code	Gene ID	Fold expression	Predicted Linear Model Ratios (IPF/Healthy)	P-values significant after Bonferroni correction	FDR	Age-adjusted P-values significant after Bonferroni Correction	Age-adjusted FDR
Q99729	🟦	HNRNPAB	0.345921247	0.35846041	1.945E-06	0.001554908	4.6961E-09	0.001953353
P62979	🟦	RPS27A	0.470375766	0.461535957	5.172E-07	0.001243926	9.9226E-13	0.0002138
Q7Z4V5	🟦	HDGFRP2	0.060924947	0.055653979	6.293E-11	0.000408163	4.0153E-34	0.001243926
P14061	🟦	HSD17B1	0.176376578	0.186493411	1.353E-07	0.001088435	4.8554E-20	0.002410107
P08700	🟦	IL3	0.62221433	0.627986419	3.837E-06	0.001710398	1.9802E-11	0.000223518
P09211	🟦	GSTP1	0.544764141	0.534249412	3.899E-07	0.001205053	3.0047E-13	0.000165209
P17174	🟦	GOT1	0.524455222	0.517574344	1.944E-09	0.000602527	5.2867E-19	0.000310982
P07355	🟦	ANXA2	0.298075338	0.296106201	3.332E-10	0.000495627	4.6371E-22	0.001788144
P40121	🟦	CAPG	0.231802025	0.233433389	1.208E-09	0.00058309	1.4771E-12	0.00244898
O15264	🟦	MAPK13	0.299068474	0.301080845	1.696E-08	0.000806608	5.1201E-11	0.001875607
P26038	🟦	MSN	0.351191486	0.349889679	2.743E-09	0.000621963	3.0007E-12	0.002478134
P43490	🟦	NAMPT	0.240390948	0.256498616	3.815E-08	0.000884354	2.7028E-14	0.001584062
Q99497	🟦	PARK7	0.310701688	0.322106754	1.307E-08	0.000777454	1.2199E-18	0.00324587
O75144	🟦	ICOSLG	0.219547354	0.223111787	1.566E-07	0.001107872	2.0022E-09	0.00180758
Q99706	🟦	KIR2DL4	0.478349129	0.479635576	1.145E-07	0.001030126	7.8372E-11	0.002099125
Q14162	🟦	SCARF1	0.516331552	0.508016155	1.754E-08	0.000816327	6.8641E-19	0.003313897
P06702	🟦	S100A9	0.591697735	0.586341331	1.214E-06	0.001428571	1.0594E-07	0.001195335
P0C0L5	🟦	C4A C4B	0.358507662	0.369051932	1.363E-12	0.000291545	3.8228E-20	0.000524781
P09429	🟦	HMGB1	0.379602216	0.375656719	2.211E-08	0.000835763	4.4323E-16	0.000495627
P29353	🟦	SHC1	0.39845923	0.390325715	6.392E-10	0.000544218	3.9594E-20	0.00170068
P04083	🟦	ANXA1	0.491332759	0.484731275	2.239E-08	0.000845481	7.6468E-10	0.000174927
Q08945	🟦	SSRP1	0.064935707	0.074339255	1.314E-07	0.001068999	7.0746E-21	0.00234208
Q92851	🟦	CASP10	0.733486915	0.746854294	4.988E-06	0.00180758	1.6671E-06	0.001457726
P22626	🟦	HNRNPA2B1	0.304653849	0.304147983	1.089E-10	0.000437318	3.7583E-14	0.001321672
Q05397	🟦	PTK2	0.361679962	0.378753188	8.674E-09	0.000738581	2.6631E-05	0.000379009
P01033	■	TIMP1	0.679269263	0.66674616	0.000391	0.002905734	6.6026E-08	0.008357629
P01034	■	CST3	0.63498025	0.648738355	0.0021138	0.003488824	3.7008E-06	0.00808552
P02788	■	LTF	0.646580172	0.660962703	0.0002031	0.002633625	3.8306E-06	0.004723032
P05186	■	ALPL	0.730903722	0.736706819	7.331E-05	0.002303207	1.5228E-06	0.007609329
P55773	■	CCL23	0.582518753	0.619691865	0.0005164	0.003061224	3.935E-07	0.003226433

Uniprot ID	Color Code	Gene ID	Fold expression	Predicted Linear Model Ratios (IPF/Healthy)	P-values significant after Bonferroni correction	FDR	Age-adjusted P-values significant after Bonferroni Correction	Age-adjusted FDR
P07451	🟦	CA3	0.595832363	0.560995319	0.0029284	0.003654033	2.8427E-06	0.001350826
P63241	🟦	EIF5A	0.639598496	0.626260436	4.906E-05	0.002264334	1.829E-08	0.004752187
P61088	🟦	UBE2N	0.679152378	0.668632317	9.424E-05	0.002380952	3.7967E-06	0.002779397
P0C0S5	🟦	H2AFZ	0.085008123	0.094716988	1.043E-05	0.002001944	2.6092E-14	0.004344023
Q9NP79	🟦	VTA1	0.68463685	0.665210265	1.045E-05	0.002011662	5.6174E-11	0.005344995
P25787	🟦	PSMA2	0.390509178	0.357843285	0.0047528	0.003906706	1.1165E-07	0.005004859
P05155	🟦	SERPING1	0.407843012	0.377329305	0.002597	0.003586006	1.7868E-08	0.001078717
P04141	🟦	CSF2	0.716695601	0.70886671	0.0024597	0.003566569	1.5111E-06	0.005218659
P08833	🟦	IGFBP1	0.393702902	0.408439401	0.000134	0.002478134	1.192E-06	0.009339164
P80188	🟦	LCN2	0.192644603	0.212966828	0.0001084	0.002419825	2.3708E-10	0.004003887
P62979	🟦	RPS27A	0.748787706	0.725982264	0.0001239	0.002458698	1.3893E-10	0.0002138
Q16627	🟦	CCL14	0.730237589	0.725946814	0.0004162	0.002944606	7.5672E-06	0.006297376
O14625	🟦	CXCL11	0.480301473	0.471102974	0.0003025	0.00281827	1.7786E-07	0.004266278
Q9HD89	🟦	RETN	0.53666193	0.538277838	0.0002682	0.002750243	1.1095E-07	0.007832847
P18428	🟦	LBP	0.601598528	0.621936449	0.0001787	0.00260447	4.7373E-08	0.010447036
P20333	🟦	TNFRSF1B	0.528556645	0.530989059	0.0001396	0.00249757	2.0522E-09	0.004596696
P23946	🟦	CMA1	0.422400525	0.449911378	0.000103	0.002410107	6.0735E-08	0.008551992
P78380	🟦	OLR1	0.376747719	0.387146365	1.569E-05	0.002060253	4.2928E-09	0.001729835
P61960	🟦	UFM1	0.61686735	0.606344498	1.583E-05	0.002069971	1.0983E-09	0.00281827
P78344	🟦	EIF4G2	0.517757776	0.509215443	5.586E-05	0.002274052	1.092E-08	0.004013605
Q8IVD9	🟦	NUDCD3	0.555472185	0.548221386	1.009E-05	0.001992225	1.1417E-10	0.003119534
P60174	🟦	TPI1	0.693640403	0.683253702	8.747E-05	0.002361516	3.7814E-06	0.006287658
P0DJI8	🟦	SAA1	0.173531112	0.180295633	0.0004928	0.00303207	3.3565E-08	0.00090379
Q07654	🟦	TFF3	0.478602321	0.492382656	0.0051497	0.003974733	8.1428E-06	0.005228377
P13726	🟦	F3	0.592836215	0.609717233	0.0001829	0.002614189	2.7012E-07	0.002827988
P19957	🟦	PI3	0.359454756	0.367402763	8.107E-05	0.002332362	1.1063E-10	0.00526725
P14555	🟦	PLA2G2A	0.164508441	0.178116603	0.0004127	0.002934888	2.7444E-09	0.002322643
P29317	🟦	EPHA2	0.543657766	0.539472754	0.0029097	0.003644315	2.3118E-06	0.007152575
Q08ET2	🟦	SIGLEC14	0.653055432	0.645056946	0.0003386	0.002857143	8.2826E-06	0.004897959
Q8TDQ0	🟦	HAVCR2	0.738663283	0.733196887	0.0002157	0.002691934	8.6442E-06	0.008678328

Uniprot ID	Color Code	Gene ID	Fold expression	Predicted Linear Model Ratios (IPF/Healthy)	P-values significant after Bonferroni correction	FDR	Age-adjusted P-values significant after Bonferroni Correction	Age-adjusted FDR
P13686		ACP5	1.504767826	1.489334666	4.532E-09	0.000670554	9.3819E-07	0.000777454
P20273		CD22	1.184673275	1.176234949	8.388E-09	0.000728863	3.2038E-06	0.003605442
P51665		PSMD7	1.227734879	1.221843412	6.226E-10	0.0005345	7.0956E-06	0.002050534
Q07817		BCL2L1	1.490277914	1.479585533	2.427E-12	0.000310982	5.7369E-08	0.000884354
P06396		GSN	1.285528406	1.306069247	3.753E-06	0.00170068	7.1603E-07	0.001418853
P04196		HRG	1.418870563	1.442868446	2.642E-07	0.001166181	1.6043E-07	0.001360544
Q9HCK4		ROBO2	1.407713095	1.395479599	1.126E-07	0.00101069	1.6629E-08	0.001516035
P12268		IMPDH2	1.338326362	1.339255747	2.562E-10	0.00047619	5.2757E-06	0.00313897
P22223		CDH3	1.385373136	1.399310188	1.063E-07	0.000981535	2.2775E-07	0.003634597
P07225		PROS1	1.272634163	1.275797707	2.912E-07	0.001185617	4.6845E-09	0.002176871
O43291		SPINT2	1.485466338	1.458185513	1.299E-06	0.00143829	1.8954E-06	0.001982507
Q9BY41		HDAC8	1.424686176	1.425858407	2.406E-16	0.000155491	1.5161E-09	0.006598639
P08697		SERPINF2	1.31047173	1.30667218	1.367E-08	0.000787172	3.6237E-11	0.002934888
P02748		C9	0.781050273	0.782880587	1.145E-07	0.001020408	1.3663E-07	0.002332362
P29622		SERPINA4	1.269072953	1.257079381	8.1E-06	0.001953353	3.5722E-06	0.001496599
Q96IY4		CPB2	0.819875158	0.824769315	4.543E-06	0.001778426	4.0513E-07	0.003449951
P31785		IL2RG	1.545659909	1.496723647	6.757E-07	0.001311953	0.00018474	0.0004276
P26951		IL3RA	1.508301381	1.468159905	2.326E-06	0.00159378	0.00142495	0.00143829
O76036		NCR1	0.741569652	0.755075196	5.621E-06	0.001836735	2.8209E-05	0.00212828
P02649		APOE	1.325399966	1.282131162	2.048E-07	0.001137026	0.000283	0.001622935
P10721		KIT	1.489125587	1.416727689	7.463E-09	0.000719145	9.22E-06	0.001652089
P09758		TACSTD2	1.420541839	1.416748142	7.103E-08	0.000952381	0.00034927	0.001506317
P00533		EGFR	1.30263973	1.283543266	6.012E-06	0.001895044	1.9617E-05	0.001399417
Q9BYF1		ACE2	1.343109036	1.351129252	7.681E-07	0.00133139	0.00036686	0.006180758
P02649		APOE	1.244447984	1.206073828	7.474E-06	0.001924198	0.0024964	0.001622935
P02649		APOE	1.286022867	1.239419842	2.083E-06	0.001574344	0.00025421	0.001622935
P17931		LGALS3	0.752839331	0.772008464	2.628E-07	0.001156463	1.0971E-05	0.000398445
P35475		IDUA	1.411818675	1.445766652	5.303E-07	0.001253644	7.8693E-05	0.000748299
P01374, Q06643		LTA LTB	1.393305269	1.389026313	1.467E-06	0.001477162	9.2873E-05	0.004664723
P29965		CD40LG	1.474153245	1.453863625	1.945E-06	0.00154519	0.00291881	0.003790087

Uniprot ID	Color Code	Gene ID	Fold expression	Predicted Linear Model Ratios (IPF/Healthy)	P-values significant after Bonferroni correction	FDR	Age-adjusted P-values significant after Bonferroni Correction	Age-adjusted FDR
Q9Y4X3		CCL27	1.173722536	1.162169851	3.663E-06	0.001681244	6.6273E-05	0.00122449
P07585		DCN	1.20880279	1.189205245	5.211E-06	0.001817298	0.00010196	0.001885326
Q9NP95		FGF20	1.438391531	1.407985353	5.916E-06	0.001875607	0.00576751	0.002390671
P08684		CYP3A4	1.445797436	1.410315476	3.527E-06	0.001661808	5.9022E-05	0.00111759
P29279		CTGF	1.32159639	1.299436776	5.526E-06	0.001827017	0.00012742	0.002585034
P56470		LGALS4	1.363080837	1.386548727	2.305E-07	0.001146744	0.0004329	0.001642371
Q15582		TGFBI	1.377537653	1.384270242	1.785E-06	0.001525753	0.00011283	0.001273081
Q8IWV2		CNTN4	1.266326604	1.257256862	4.96E-06	0.001797862	1.4765E-05	0.001107872
O94779		CNTN5	1.258655438	1.254045937	1.594E-06	0.001496599	8.4023E-05	0.002837707
P06493, P14635		CDC2 CCNB1	1.198173718	1.181337678	3.301E-06	0.001613217	0.00089827	0.005578231
Q92876		KLK6	0.864513016	0.866147653	2.492E-06	0.001603499	0.00035015	0.006375121
O00626		CCL22	1.424081215	1.429480123	3.309E-06	0.001622935	0.00020088	0.004635569
P01282		VIP	1.424445332	1.413234053	5.966E-06	0.001885326	0.00344855	0.002089407
Q6UXD5		SEZ6L2	1.407069992	1.369992374	8.816E-06	0.001972789	0.0005441	0.009193392
P68036		UBE2L3	1.467831972	1.427595832	3.606E-06	0.001671526	0.00063943	0.009115646
Q96GD0		PDXP	1.490673622	1.447598284	7.077E-06	0.00191448	0.00047606	0.000835763
P08620		FGF4	1.23892107	1.229893786	6.193E-07	0.001302235	0.00118849	0.001661808
Q99075		HBEGF	1.293217012	1.291360122	4.647E-07	0.00122449	0.00106179	0.000758017
P20783		NTF3	1.414871631	1.405958216	5.997E-07	0.001292517	0.00121293	0.000456754
P32004		L1CAM	1.422955729	1.402037028	4.477E-07	0.001214772	1.0997E-05	0.002361516
O43323		DHH	1.427120242	1.403360929	1.827E-06	0.001535471	0.00075255	0.000864917
O43320		FGF16	1.401626175	1.370275971	1.549E-06	0.00148688	0.00060477	0.002419825
O75356		ENTPD5	1.27036547	1.282597991	3.47E-06	0.001652089	1.6184E-05	0.001020408
Q4KMG0		CDON	1.438856529	1.395132851	6.493E-06	0.001904762	1.3063E-05	0.000573372
P10909		CLU	1.268085243	1.273973382	4.497E-06	0.001768707	9.0924E-05	0.000826045
Q9NZU1		FLRT1	1.438468618	1.404044634	9.319E-07	0.001360544	0.00032809	0.002069971
P21217		FUT3	1.457634251	1.441771268	2.029E-07	0.001127308	0.00057164	0.000874636
Q12884		FAP	1.355855033	1.36028669	3.434E-06	0.001642371	0.00025698	0.000136054
Q02241		KIF23	1.439636723	1.387282427	3.31E-06	0.001632653	0.00061253	0.000417881
P45984		MAPK9	1.366287937	1.385188001	5.896E-06	0.001865889	0.00016857	0.001302235

Uniprot ID	Color Code	Gene ID	Fold expression	Predicted Linear Model Ratios (IPF/Healthy)	P-values significant after Bonferroni correction	FDR	Age-adjusted P-values significant after Bonferroni Correction	Age-adjusted FDR
P48061		CXCL12	0.779402388	0.782938513	9.425E-06	0.001982507	6.6164E-06	0.00447036
P29401		TKT	0.76346032	0.750233644	0.0004592	0.003002915	7.2265E-06	0.009737609

Color Code

🟥 Upregulated in IPF AND Age-adjusted AND non-age-adjusted significant

🟥 Upregulated in IPF AND non-age significant

Age-adjusted AND/OR non-age-adjusted significant but not biologically relevant

🟦 Downregulated in IPF AND Age-adjusted AND non-age-adjusted significant

🟦 Downregulated in IPF AND Age-adjusted significant

🟦 Downregulated in IPF AND non-age-adjusted significant

Table A.S6. ClueGO analysis of biological roles of upregulated protesin in IPF plasma.

GOTerm	Ontology Source	Term PValue	Term PValue Corrected with Bonferroni	Group PValue Corrected with Bonferroni	% Associated Genes	Nr. Genes	Associated Genes Found
regulation of cardiac muscle hypertrophy	GO_BiologicalProcess-GOA_09.02.2016_16h18	71.0E-9	2.9E-6	210.0E-9	12.82	5.00	[ECE1, GSK3A, GSK3B, PDE5A, PRKCA]
ErbB signaling pathway	KEGG_10.02.2016	4.0E-9	160.0E-9	17.0E-15	8.05	7.00	[GSK3B, MAP2K4, PIK3CA, PIK3R1, PRKCA, PRKCB, SRC]
Sphingolipid signaling pathway	KEGG_10.02.2016	950.0E-9	39.0E-6	17.0E-15	5.00	6.00	[PIK3CA, PIK3R1, PRKCA, PRKCB, RAC1, SPHK1]
VEGF signaling pathway	KEGG_10.02.2016	5.1E-12	210.0E-12	17.0E-15	13.11	8.00	[MAPKAPK2, PIK3CA, PIK3R1, PRKCA, PRKCB, RAC1, SPHK1, SRC]
B cell receptor signaling pathway	KEGG_10.02.2016	49.0E-9	2.0E-6	17.0E-15	8.22	6.00	[BTK, GSK3B, PIK3CA, PIK3R1, PRKCB, RAC1]
Fc epsilon RI signaling pathway	KEGG_10.02.2016	32.0E-9	1.3E-6	17.0E-15	8.82	6.00	[BTK, MAP2K4, PIK3CA, PIK3R1, PRKCA, RAC1]
Fc gamma R-mediated phagocytosis	KEGG_10.02.2016	210.0E-9	8.6E-6	17.0E-15	6.45	6.00	[PIK3CA, PIK3R1, PRKCA, PRKCB, RAC1, SPHK1]
Thyroid hormone signaling pathway	KEGG_10.02.2016	860.0E-9	35.0E-6	17.0E-15	5.08	6.00	[GSK3B, PIK3CA, PIK3R1, PRKCA, PRKCB, SRC]
AGE-RAGE signaling pathway in diabetic complications	KEGG_10.02.2016	340.0E-9	14.0E-6	17.0E-15	5.94	6.00	[FN1, PIK3CA, PIK3R1, PRKCA, PRKCB, RAC1]
Bacterial invasion of epithelial cells	KEGG_10.02.2016	2.4E-6	99.0E-6	17.0E-15	6.41	5.00	[FN1, PIK3CA, PIK3R1, RAC1, SRC]
T cell costimulation	GO_ImmuneSystemProcess-GOA_09.02.2016_16h18	210.0E-9	8.6E-6	17.0E-15	6.45	6.00	[CSK, PIK3CA, PIK3R1, RAC1, SRC, TNFSF13B]
platelet activation	GO_BiologicalProcess-GOA_09.02.2016_16h18	430.0E-15	17.0E-12	17.0E-15	4.69	13.00	[APOA1, CLIC1, CSK, F5, FN1, PIK3CA, PIK3R1, PRKCA, PRKCB, RAC1, SELP, SRC, TEC]
regulation of cellular response to insulin stimulus	GO_BiologicalProcess-GOA_09.02.2016_16h18	700.0E-9	28.0E-6	17.0E-15	8.20	5.00	[GSK3A, PIK3R1, PRKCA, PRKCB, SRC]
Complement and coagulation cascades	KEGG_10.02.2016	1.3E-6	53.0E-6	3.9E-6	7.25	5.00	[C1R, C4B, C5, F5, SERPINA5]

Table A.S7. ClueGo analysis of the biological roles of downregulated proteins in IPF plasma.

GOTerm	Ontology Source	Term PValue	Term PValue Corrected with Bonferroni	Group PValue Corrected with Bonferroni	% Associated Genes	Nr. Genes	Associated Genes Found
acute inflammatory response	GO_BiologicalProcess-GOA_09.02.2016_16h18	67.0E-9	11.0E-6	740.0E-9	6.10	10.00	[C3, CRP, F3, GSTP1, LBP, PTGS2, SAA1, SERPINA3, SERPING1, TIMP1]
Fc receptor signaling pathway	GO_ImmuneSystemProcess-GOA_09.02.2016_16h18	650.0E-12	110.0E-9	3.4E-15	4.09	17.00	[CAMK2B, CAMK2D, CDKN1B, CFL1, CSF2, FYN, IL2, IL3, MAP2K1, MAP2K2, PEBP1, PSMA2, PTPN11, RPS27A, SHC1, UBE2N, YES1]
response to peptide hormone	GO_BiologicalProcess-GOA_09.02.2016_16h18	10.0E-15	1.8E-12	3.4E-15	4.07	26.00	[ANXA1, CAMK2B, CAMK2D, CDKN1B, CSF2, FYN, GOT1, GSTP1, IGFBP1, IGFBP2, IL2, IL3, MAP2K1, MAP2K2, NAMPT, PEBP1, POR, PSMA2, PTPN1, PTPN11, PTPN6, RETN, RPS27A, SHC1, TFF3, TIMP1]
cellular response to fibroblast growth factor stimulus	GO_BiologicalProcess-GOA_09.02.2016_16h18	4.1E-9	720.0E-9	3.4E-15	4.26	15.00	[CAMK2B, CAMK2D, CDKN1B, CSF2, CXCL8, FYN, IL2, IL3, MAP2K1, MAP2K2, PEBP1, PSMA2, PTPN11, RPS27A, SHC1]
response to insulin	GO_BiologicalProcess-GOA_09.02.2016_16h18	11.0E-12	1.9E-9	3.4E-15	4.18	20.00	[CAMK2B, CAMK2D, CSF2, FYN, GOT1, GSTP1, IGFBP1, IGFBP2, IL2, IL3, MAP2K1, MAP2K2, NAMPT, PEBP1, PSMA2, PTPN1, PTPN11, RETN, RPS27A, SHC1]
Fc-epsilon receptor signaling pathway	GO_ImmuneSystemProcess-GOA_09.02.2016_16h18	3.8E-9	660.0E-9	3.4E-15	4.29	15.00	[CAMK2B, CAMK2D, CDKN1B, CSF2, FYN, IL2, IL3, MAP2K1, MAP2K2, PEBP1, PSMA2, PTPN11, RPS27A, SHC1, UBE2N]
cellular response to insulin stimulus	GO_BiologicalProcess-GOA_09.02.2016_16h18	97.0E-12	17.0E-9	3.4E-15	4.27	18.00	[CAMK2B, CAMK2D, CSF2, FYN, GOT1, GSTP1, IGFBP1, IL2, IL3, MAP2K1, MAP2K2, NAMPT, PEBP1, PSMA2, PTPN1, PTPN11, RPS27A, SHC1]
insulin receptor signaling pathway	GO_BiologicalProcess-GOA_09.02.2016_16h18	380.0E-12	66.0E-9	3.4E-15	4.61	16.00	[CAMK2B, CAMK2D, CSF2, FYN, IGFBP1, IL2, IL3, MAP2K1, MAP2K2, NAMPT, PEBP1, PSMA2, PTPN1, PTPN11, RPS27A, SHC1]
epidermal growth factor receptor signaling pathway	GO_BiologicalProcess-GOA_09.02.2016_16h18	150.0E-12	27.0E-9	3.4E-15	4.49	17.00	[CAMK2B, CAMK2D, CDKN1B, CSF2, FYN, IL2, IL3, ITGA1, MAP2K1, MAP2K2, MMP9, PEBP1, PLAUR, PSMA2, PTPN11, RPS27A, SHC1]
phagocytosis	GO_BiologicalProcess-GOA_09.02.2016_16h18	160.0E-9	28.0E-6	1.8E-6	4.26	12.00	[AIF1, ANXA1, C3, CFL1, CRP, FCN1, FYN, HMGB1, LBP, PRTN3, SFTPD, YES1]
peptidase regulator activity	GO_MolecularFunction-GOA_09.02.2016_16h18	150.0E-9	26.0E-6	14.0E-12	4.85	11.00	[C3, C4B_2, CAST, CDKN1B, CST3, PEBP1, PI3, SERPINA3, SERPING1, TIMP1, TNFSF14]
endopeptidase inhibitor activity	GO_MolecularFunction-GOA_09.02.2016_16h18	160.0E-9	28.0E-6	14.0E-12	5.56	10.00	[C3, C4B_2, CAST, CST3, PEBP1, PI3, SERPINA3, SERPING1, TIMP1, TNFSF14]
regulation of endopeptidase activity	GO_BiologicalProcess-GOA_09.02.2016_16h18	490.0E-15	86.0E-12	14.0E-12	4.96	20.00	[C3, C4B_2, CAST, CDKN1B, CST3, F3, GPI, HMGB1, HTRA2, MMP9, PARK7, PEBP1, PI3, PLAUR, POR, S100A9, SERPINA3, SERPING1, TIMP1, TNFSF14]

GOTerm	Ontology Source	Term PValue	Term PValue Corrected with Bonferroni	Group PValue Corrected with Bonferroni	% Associated Genes	Nr. Genes	Associated Genes Found
negative regulation of endopeptidase activity	GO_BiologicalProcess-GOA_09.02.2016_16h18	69.0E-12	12.0E-9	14.0E-12	5.73	15.00	[C3, C4B_2, CAST, CST3, GPI, MMP9, PARK7, PEBP1, PI3, PLAUR, POR, SERPINA3, SERPING1, TIMP1, TNFSF14]
regulation of cysteine-type endopeptidase activity involved in apoptotic process	GO_BiologicalProcess-GOA_09.02.2016_16h18	110.0E-9	20.0E-6	14.0E-12	4.98	11.00	[CDKN1B, F3, GPI, HMGB1, HTRA2, MMP9, PARK7, PLAUR, POR, S100A9, TNFSF14]
positive regulation of leukocyte activation	GO_ImmuneSystemProcess-GOA_09.02.2016_16h18	21.0E-9	3.8E-6	25.0E-9	4.12	14.00	[AIF1, ANXA1, FYN, HAVCR2, HMGB1, IGFBP2, IL2, LBP, PTPN11, PTPN6, S100A9, TIMP1, TNFSF14, YES1]
leukocyte proliferation	GO_BiologicalProcess-GOA_09.02.2016_16h18	4.0E-9	710.0E-9	25.0E-9	4.70	14.00	[AIF1, ANXA1, CSF1, FYN, GSTP1, HMGB1, IGFBP2, IL2, IL3, PTPN6, S100A9, SFTPD, TIMP1, TNFSF14]
mononuclear cell proliferation	GO_BiologicalProcess-GOA_09.02.2016_16h18	180.0E-9	32.0E-6	25.0E-9	4.21	12.00	[AIF1, ANXA1, CSF1, FYN, HMGB1, IGFBP2, IL2, PTPN6, S100A9, SFTPD, TIMP1, TNFSF14]
T cell proliferation	GO_ImmuneSystemProcess-GOA_09.02.2016_16h18	280.0E-9	49.0E-6	25.0E-9	5.24	10.00	[AIF1, ANXA1, FYN, HMGB1, IGFBP2, IL2, PTPN6, SFTPD, TIMP1, TNFSF14]
regulation of T cell activation	GO_ImmuneSystemProcess-GOA_09.02.2016_16h18	83.0E-9	14.0E-6	25.0E-9	4.06	13.00	[AIF1, ANXA1, FYN, HAVCR2, HMGB1, IGFBP2, IL2, PTPN11, PTPN6, SFTPD, TIMP1, TNFSF14, YES1]
positive regulation of T cell activation	GO_ImmuneSystemProcess-GOA_09.02.2016_16h18	150.0E-9	26.0E-6	25.0E-9	4.85	11.00	[AIF1, ANXA1, FYN, HMGB1, IGFBP2, IL2, PTPN11, PTPN6, TIMP1, TNFSF14, YES1]
regulation of leukocyte proliferation	GO_BiologicalProcess-GOA_09.02.2016_16h18	11.0E-9	1.9E-6	25.0E-9	5.43	12.00	[AIF1, ANXA1, CSF1, GSTP1, HMGB1, IGFBP2, IL2, IL3, PTPN6, S100A9, SFTPD, TIMP1]
positive regulation of leukocyte proliferation	GO_BiologicalProcess-GOA_09.02.2016_16h18	250.0E-9	44.0E-6	25.0E-9	6.25	9.00	[AIF1, ANXA1, CSF1, HMGB1, IGFBP2, IL2, IL3, S100A9, TIMP1]
cellular response to interferon-gamma	GO_ImmuneSystemProcess-GOA_09.02.2016_16h18	280.0E-9	50.0E-6	150.0E-12	6.16	9.00	[AIF1, CAMK2B, CAMK2D, CCL14, CCL23, PTPN1, PTPN11, PTPN6, SYNCRIP]
ERK1 and ERK2 cascade	GO_BiologicalProcess-GOA_09.02.2016_16h18	91.0E-12	15.0E-9	150.0E-12	5.62	15.00	[C3, CAMK2D, CCL14, CCL23, EPHA2, GSTP1, HMGB1, MAP2K1, MAP2K2, PLA2G2A, PTPN1, PTPN11, PTPN6, S100A9, TIMP1]
regulation of ERK1 and ERK2 cascade	GO_BiologicalProcess-GOA_09.02.2016_16h18	4.3E-9	750.0E-9	150.0E-12	5.22	13.00	[C3, CAMK2D, CCL14, CCL23, EPHA2, GSTP1, HMGB1, PLA2G2A, PTPN1, PTPN11, PTPN6, S100A9, TIMP1]
granulocyte chemotaxis	GO_ImmuneSystemProcess-GOA_09.02.2016_16h18	2.0E-9	350.0E-9	22.0E-9	8.77	10.00	[ANXA1, CCL14, CCL23, CSF1, CXCL8, DAPK2, ITGA1, LBP, S100A9, SAA1]
neutrophil chemotaxis	GO_ImmuneSystemProcess-GOA_09.02.2016_16h18	120.0E-9	21.0E-6	22.0E-9	8.42	8.00	[CCL14, CCL23, CXCL8, DAPK2, ITGA1, LBP, S100A9, SAA1]
positive regulation of leukocyte chemotaxis	GO_ImmuneSystemProcess-GOA_09.02.2016_16h18	55.0E-9	9.7E-6	74.0E-24	9.30	8.00	[AIF1, CSF1, CXCL11, CXCL8, DAPK2, HMGB1, LBP, TNFSF14]

GOTerm	Ontology Source	Term PValue	Term PValue Corrected with Bonferroni	Group PValue Corrected with Bonferroni	% Associated Genes	Nr. Genes	Associated Genes Found
inflammatory response	GO_BiologicalProcess-GOA_09.02.2016_16h18	26.0E-18	4.6E-15	74.0E-24	4.20	30.00	[AIF1, ANXA1, C3, C4B_2, CCL14, CCL23, CMA1, CRP, CSF1, CTSS, CXCL11, CXCL8, F3, GSTP1, HAVCR2, HMGB1, IL2, LBP, MAPK13, OLR1, PARK7, PGLYRP1, PLA2G2A, PTGS2, S100A9, SAA1, SERPINA3, SERPING1, TIMP1, TNFRSF1B]
positive regulation of response to external stimulus	GO_BiologicalProcess-GOA_09.02.2016_16h18	20.0E-18	3.6E-15	74.0E-24	7.03	22.00	[AIF1, C3, CCL14, CCL23, CSF1, CTSS, CXCL11, CXCL8, DAPK2, F3, HAVCR2, HMGB1, IL16, IL2, LBP, MAPK13, PARK7, PLA2G2A, PTGS2, S100A9, SCARF1, TNFSF14]
positive regulation of defense response	GO_BiologicalProcess-GOA_09.02.2016_16h18	4.9E-12	860.0E-12	74.0E-24	4.38	20.00	[C3, CCL14, CCL23, CTSS, FCN1, FYN, HAVCR2, HMGB1, IL2, LBP, LTF, MAP2K1, MAPK13, PGLYRP1, PLA2G2A, PSMA2, PTGS2, RPS27A, S100A9, UBE2N]
positive regulation of chemotaxis	GO_BiologicalProcess-GOA_09.02.2016_16h18	9.7E-9	1.7E-6	74.0E-24	7.46	10.00	[AIF1, CSF1, CXCL11, CXCL8, DAPK2, F3, HMGB1, IL16, LBP, TNFSF14]
regulation of response to wounding	GO_BiologicalProcess-GOA_09.02.2016_16h18	23.0E-18	4.1E-15	74.0E-24	5.59	25.00	[ANXA1, ANXA2, C3, CCL14, CCL23, CMA1, CTSS, F3, GSTP1, HMGB1, IL2, LBP, MAP2K1, MAP2K2, MAPK13, PARK7, PGLYRP1, PLA2G2A, PLAUR, PTGS2, S100A9, SAA1, SCARF1, SERPING1, TNFRSF1B]
positive regulation of response to wounding	GO_BiologicalProcess-GOA_09.02.2016_16h18	3.0E-12	520.0E-12	74.0E-24	8.09	14.00	[ANXA1, C3, CCL14, CCL23, CTSS, F3, HMGB1, IL2, LBP, MAPK13, PLA2G2A, PTGS2, S100A9, SCARF1]
regulation of inflammatory response	GO_BiologicalProcess-GOA_09.02.2016_16h18	800.0E-15	140.0E-12	74.0E-24	5.70	18.00	[ANXA1, C3, CCL14, CCL23, CMA1, CTSS, GSTP1, IL2, LBP, MAPK13, PARK7, PGLYRP1, PLA2G2A, PTGS2, S100A9, SAA1, SERPING1, TNFRSF1B]
positive regulation of inflammatory response	GO_BiologicalProcess-GOA_09.02.2016_16h18	3.9E-9	680.0E-9	74.0E-24	8.20	10.00	[C3, CCL14, CCL23, CTSS, IL2, LBP, MAPK13, PLA2G2A, PTGS2, S100A9]
TNF signaling pathway	KEGG_10.02.2016	380.0E-9	66.0E-6	4.2E-6	7.27	8.00	[CASP10, CSF1, CSF2, MAP2K1, MAPK13, MMP9, PTGS2, TNFRSF1B]
Proteoglycans in cancer	KEGG_10.02.2016	48.0E-9	8.5E-6	530.0E-9	5.42	11.00	[CAMK2B, CAMK2D, ITGB1, MAP2K1, MAP2K2, MAPK13, MMP9, MSN, PLAUR, PTPN11, PTPN6]
positive regulation of tyrosine phosphorylation of Stat5 protein	GO_BiologicalProcess-GOA_09.02.2016_16h18	110.0E-9	20.0E-6	140.0E-15	25.00	5.00	[CSF2, FYN, IL2, IL3, TIMP1]
cytokine activity	GO_MolecularFunction-GOA_09.02.2016_16h18	1.1E-12	190.0E-12	140.0E-15	6.81	16.00	[CCL14, CCL23, CSF1, CSF2, CXCL11, CXCL8, GDF5, GPI, HMGB1, IL16, IL2, IL3, NAMPT, S100A9, TIMP1, TNFSF14]

cytokine receptor binding	GO_MolecularFunction-GOA_09.02.2016_16h18	31.0E-9	5.4E-6	140.0E-15	4.42	13.00	[CCL14, CCL23, CSF1, CSF2, CXCL11, CXCL8, GDF5, IL2, IL3, S100A9, SHC1, TIMP1, TNFSF14]

GOTerm	Ontology Source	Term PValue	Term PValue Corrected with Bonferroni	Group PValue Corrected with Bonferroni	% Associated Genes	Nr. Genes	Associated Genes Found
peptidyl-tyrosine phosphorylation	GO_BiologicalProcess-GOA_09.02.2016_16h18	5.5E-9	970.0E-9	140.0E-15	4.17	15.00	[CSF2, DDR2, EPHA2, FYN, IL2, IL3, ITGB1, MAP2K1, MAP2K2, PTPN1, PTPN6, S100A9, SHC1, TIMP1, YES1]
regulation of peptidyl-tyrosine phosphorylation	GO_BiologicalProcess-GOA_09.02.2016_16h18	210.0E-9	37.0E-6	140.0E-15	4.68	11.00	[CSF2, FYN, IL2, IL3, ITGB1, PTPN1, PTPN6, S100A9, SHC1, TIMP1, YES1]

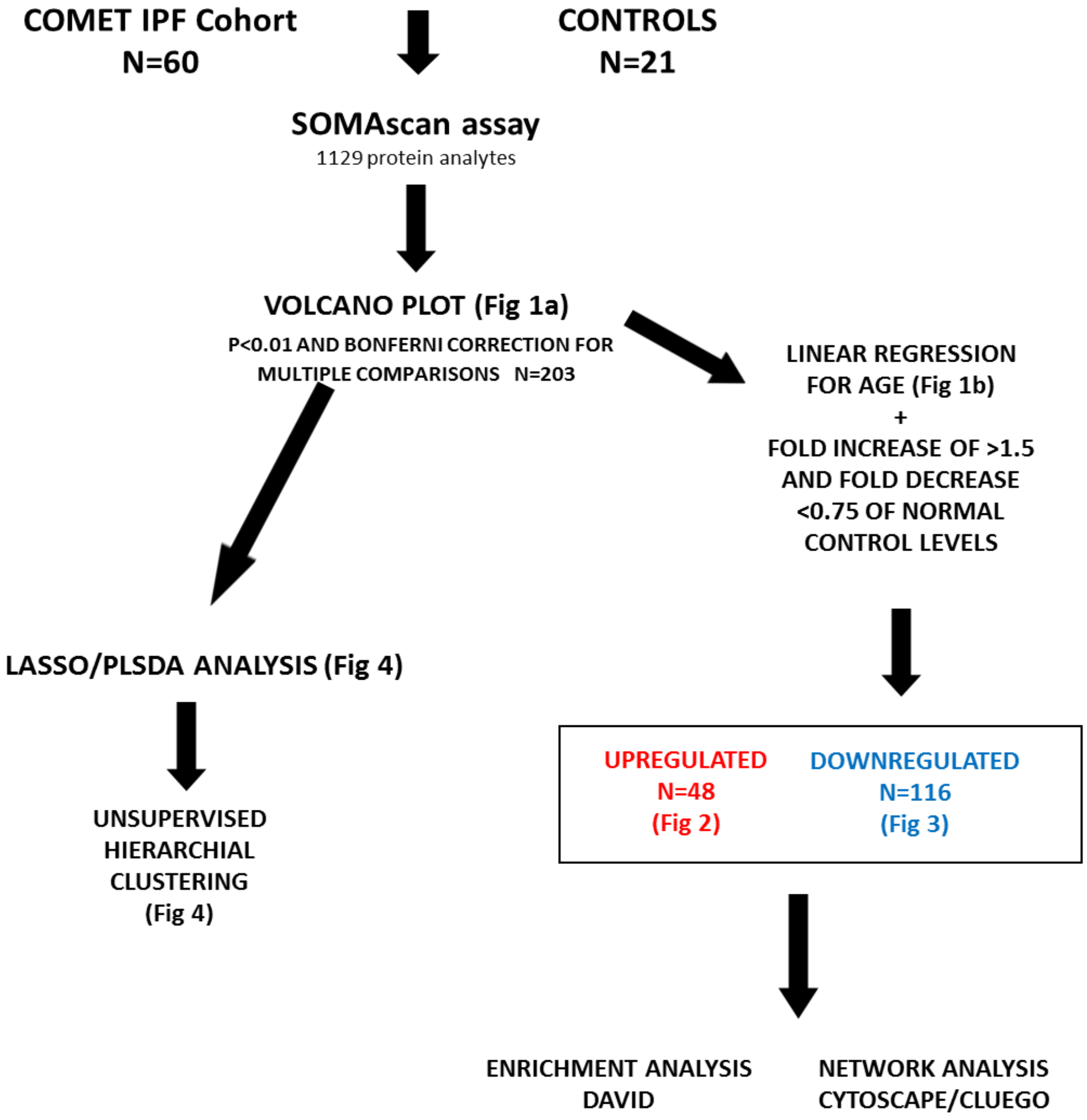


Figure A.S1. Schematic representation of approach adopted in this study.

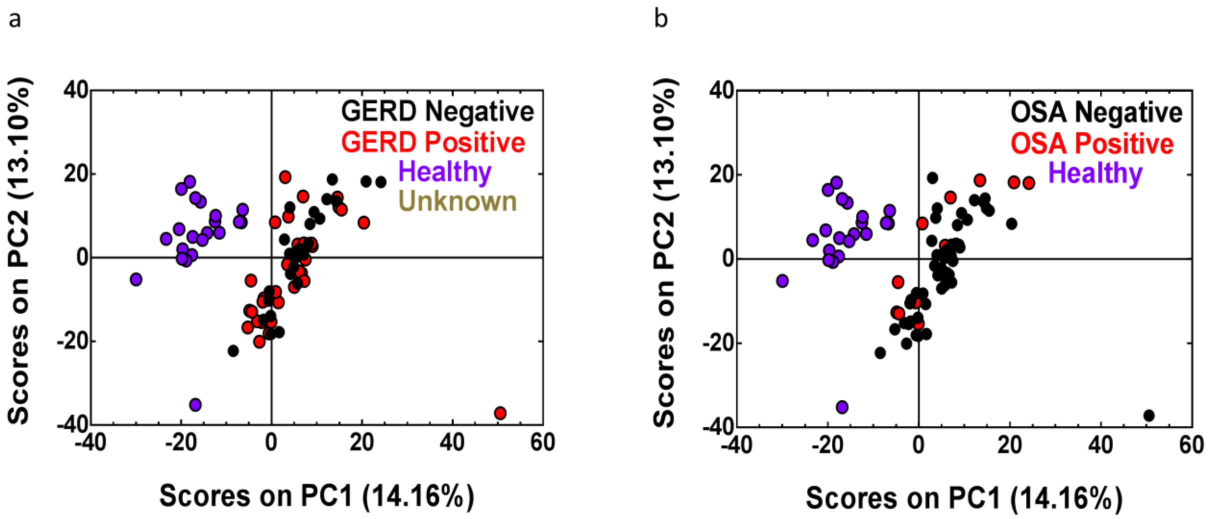
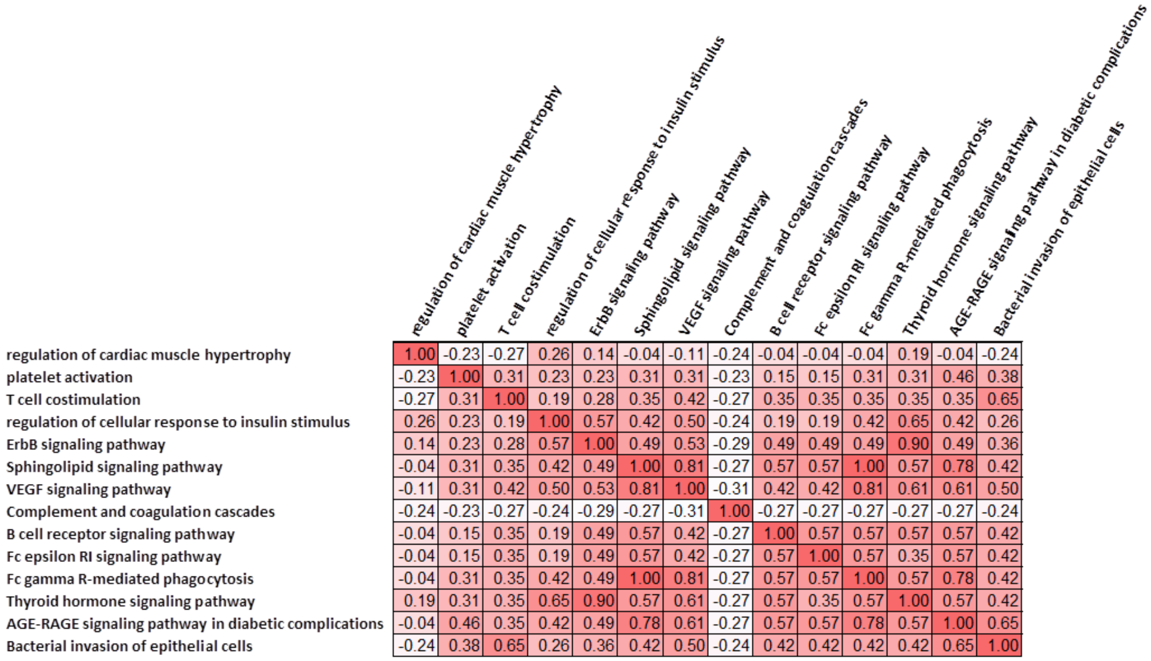


Figure A.S2. Principal Component Analysis (PCA) showed that the largest, unbiased difference in the IPF-Healthy dataset is between the IPF and healthy groups, with comorbidities in the IPF patients having little effect.
(a) A PCA model based on all 1129 measured blood proteins captured 27.26% of the total variance in the data, with PC1 explaining 14.16% of the variance and PC2, 13.10%. In this model, the healthy patients score in the negative region of PC1, and the IPF patients score mostly in the positive area of PC1. IPF patients with GERD do not cluster together within the IPF group; these patients are mixed evenly with the IPF patients who do not have GERD. **(b)** Similarly, when looking at IPF patients with obstructive sleep apnea (OSA), it can be seen that these patients are spread throughout the IPF grouping in the PCA and do not form their own cluster. The main difference in this PCA model is still between the healthy and IPF patients.



	regulation of cardiac muscle hypertrophy	platelet activation	T cell costimulation	regulation of cellular response to insulin stimulus	ErbB signaling pathway	Sphingolipid signaling pathway	VEGF signaling pathway	Complement and coagulation cascades	B cell receptor signaling pathway	Fc epsilon RI signaling pathway	Fc gamma R-mediated phagocytosis	Thyroid hormone signaling pathway	AGE-RAGE signaling pathway in diabetic complications	Bacterial invasion of epithelial cells
regulation of cardiac muscle hypertrophy	1.00	-0.23	-0.27	0.26	0.14	-0.04	-0.11	-0.24	-0.04	-0.04	-0.04	0.19	-0.04	-0.24
platelet activation	-0.23	1.00	0.31	0.23	0.23	0.31	0.31	-0.23	0.15	0.15	0.31	0.31	0.46	0.38
T cell costimulation	-0.27	0.31	1.00	0.19	0.28	0.35	0.42	-0.27	0.35	0.35	0.35	0.35	0.35	0.65
regulation of cellular response to insulin stimulus	0.26	0.23	0.19	1.00	0.57	0.42	0.50	-0.24	0.19	0.19	0.42	0.65	0.42	0.26
ErbB signaling pathway	0.14	0.23	0.28	0.57	1.00	0.49	0.53	-0.29	0.49	0.49	0.49	0.90	0.49	0.36
Sphingolipid signaling pathway	-0.04	0.31	0.35	0.42	0.49	1.00	0.81	-0.27	0.57	0.57	1.00	0.57	0.78	0.42
VEGF signaling pathway	-0.11	0.31	0.42	0.50	0.53	0.81	1.00	-0.31	0.42	0.42	0.81	0.61	0.61	0.50
Complement and coagulation cascades	-0.24	-0.23	-0.27	-0.24	-0.29	-0.27	-0.31	1.00	-0.27	-0.27	-0.27	-0.27	-0.27	-0.24
B cell receptor signaling pathway	-0.04	0.15	0.35	0.19	0.49	0.57	0.42	-0.27	1.00	0.57	0.57	0.57	0.57	0.42
Fc epsilon RI signaling pathway	-0.04	0.15	0.35	0.19	0.49	0.57	0.42	-0.27	0.57	1.00	0.57	0.35	0.57	0.42
Fc gamma R-mediated phagocytosis	-0.04	0.31	0.35	0.42	0.49	1.00	0.81	-0.27	0.57	0.57	1.00	0.57	0.78	0.42
Thyroid hormone signaling pathway	0.19	0.31	0.35	0.65	0.90	0.57	0.61	-0.27	0.57	0.35	0.57	1.00	0.57	0.42
AGE-RAGE signaling pathway in diabetic complications	-0.04	0.46	0.35	0.42	0.49	0.78	0.61	-0.27	0.57	0.57	0.78	0.57	1.00	0.65
Bacterial invasion of epithelial cells	-0.24	0.38	0.65	0.26	0.36	0.42	0.50	-0.24	0.42	0.42	0.42	0.42	0.65	1.00

Figure A.S3. Kappa statistics from the upregulated proteome.

Level of agreement between gene terms is measured by Kappa statistics (default <4). Red scale depicts level of agreement from very high (1) to very low (-1).

NAME	acute inflammatory response	positive regulation of leukocyte chemota	positive regulation of leukocyte activation	endopeptidase inhibitor activi	cytokine activity	cytokine receptor binding	phagocytosis	inflammatory response	epidermal growth factor receptor signaling pathway	insulin receptor signaling pathway	negative regulation of endopeptidase acti	peptidyl-tyrosine phosphorylation	neutrophil chemotaxis	positive regulation of defense response	positive regulation of response to external stimulus	response to insulin	cellular response to insulin stimulus	mononuclear cell proliferation	Fc receptor signaling pathway	Fc-epsilon receptor signaling pathwa	T cell proliferation	positive regulation of tyrosine phosphorylation of Stat5 protein	regulation of cysteine-type endopeptidase activity involved in apoptotic pro	response to peptide hormone	cellular response to fibroblast growth factor stimulus	regulation of inflammatory response	positive regulation of inflammatory respo	regulation of peptidyl-tyrosine phosphorylation	regulation of T cell activation	positive regulation of T cell activation	positive regulation of chemotaxis	regulation of endopeptidase activity	peptidase regulator activity	ERK1 and ERK2 cascade	regulation of ERK1 and ERK2 cascade	leukocyte proliferation	regulation of leukocyte proliferation	positive regulation of leukocyte prolifera	cellular response to interferon-gamma	granulocyte chemotaxis	regulation of response to wounding	positive regulation of response to wounding	TNF signaling pathway	Proteoglycans in canc
acute inflammatory response	1.00	0.00	0.02	0.31	-0.09	-0.07	0.16	0.38	-0.19	-0.18	0.20	-0.08	0.12	0.04	0.09	-0.12	-0.11	-0.05	-0.19	-0.18	-0.03	0.05	-0.04	-0.09	-0.18	0.32	0.20	-0.04	-0.07	-0.04	0.08	0.20	0.29	0.10	0.14	0.02	0.05	-0.02	-0.14	0.08	0.27	0.22	0.00	-0.15
positive regulation of leukocyte chemotaxis	0.00	1.00	0.27	0.00	0.33	0.29	0.20	0.19	-0.16	-0.16	-0.05	-0.15	0.30	0.00	0.45	-0.17	-0.16	0.32	-0.16	-0.15	0.25	-0.08	0.11	-0.18	-0.05	-0.07	0.00	-0.13	0.18	0.22	0.87	0.00	-0.01	-0.05	-0.03	0.27	0.20	0.28	0.01	0.37	-0.04	0.06	0.03	-0.13
positive regulation of leukocyte activation	0.02	0.27	1.00	0.02	0.18	0.15	0.36	0.16	0.00	0.01	-0.06	0.28	0.06	0.18	0.22	0.03	-0.01	0.72	0.08	0.03	0.71	0.25	0.10	0.16	0.03	0.06	0.12	0.38	0.87	0.86	0.22	0.03	0.00	0.20	0.24	0.65	0.54	0.55	0.14	0.12	0.04	0.22	-0.15	0.00
endopeptidase inhibitor activity	0.31	0.00	0.02	1.00	0.00	0.04	-0.05	0.07	-0.10	-0.09	0.76	-0.08	-0.13	-0.12	-0.06	-0.12	-0.11	0.05	-0.10	-0.08	0.08	0.05	-0.04	-0.09	-0.08	-0.02	-0.03	-0.04	0.04	0.07	-0.03	0.60	0.95	0.01	0.04	0.02	-0.05	-0.02	-0.14	-0.14	-0.08	-0.08	-0.13	-0.15
cytokine activity	-0.09	0.33	0.18	0.00	1.00	0.79	-0.12	0.17	-0.03	0.06	0.00	0.16	0.23	0.07	0.38	0.00	0.03	0.31	-0.03	0.00	0.18	0.31	0.16	-0.02	0.08	0.03	0.18	0.25	0.12	0.16	0.36	0.07	-0.02	0.16	0.20	0.34	0.31	0.39	0.02	0.27	0.00	0.18	0.04	-0.20
cytokine receptor binding	-0.07	0.29	0.15	0.04	0.79	1.00	-0.19	0.18	0.10	0.12	-0.04	0.31	0.29	0.05	0.32	0.05	0.08	0.29	0.10	0.13	0.14	0.39	0.02	0.05	0.22	0.08	0.24	0.41	0.08	0.12	0.24	-0.02	0.02	0.13	0.17	0.33	0.29	0.37	0.05	0.34	-0.01	0.15	0.07	-0.18
phagocytosis	0.16	0.20	0.36	-0.05	-0.12	-0.19	1.00	0.09	-0.13	-0.12	-0.11	-0.02	-0.02	0.15	0.05	-0.16	-0.14	0.31	0.04	-0.11	0.37	0.03	-0.07	-0.13	-0.11	0.02	0.05	0.03	0.38	0.34	0.16	-0.08	-0.07	-0.02	0.00	0.26	0.21	0.18	-0.04	0.05	0.01	0.17	-0.14	-0.17
inflammatory response	0.38	0.19	0.16	0.07	0.17	0.18	0.09	1.00	-0.32	-0.30	0.02	-0.16	0.19	0.31	0.55	-0.32	-0.28	0.15	-0.32	-0.28	0.07	0.01	-0.01	-0.32	-0.22	0.65	0.38	-0.07	0.06	0.05	0.20	0.08	0.05	0.14	0.18	0.16	0.21	0.22	-0.03	0.26	0.58	0.46	0.06	-0.19
epidermal growth factor receptor signaling pathway	-0.19	-0.16	0.00	-0.10	-0.03	0.10	-0.13	-0.32	1.00	0.73	-0.02	0.30	-0.07	0.05	-0.25	0.61	0.67	-0.05	0.78	0.84	-0.01	0.29	0.05	0.53	0.84	-0.21	-0.10	0.23	0.02	0.05	-0.19	-0.02	-0.03	0.06	-0.07	0.00	-0.05	0.01	0.10	-0.10	-0.09	-0.16	0.12	0.40
insulin receptor signaling pathway	-0.18	-0.16	0.01	-0.09	0.06	0.12	-0.12	-0.30	0.73	1.00	-0.16	0.40	-0.16	0.07	-0.24	0.86	0.93	-0.04	0.73	0.80	0.00	0.31	-0.20	0.68	0.80	-0.20	-0.09	0.33	0.03	0.07	-0.18	-0.22	-0.11	0.16	0.03	0.01	-0.04	0.02	0.20	-0.18	-0.13	-0.15	0.04	0.25
negative regulation of endopeptidase activity	0.20	-0.05	-0.06	0.76	0.00	-0.04	-0.11	0.02	-0.02	-0.16	1.00	-0.15	-0.15	-0.20	-0.08	-0.20	-0.18	-0.02	-0.17	-0.15	0.01	0.01	0.36	-0.12	-0.15	-0.03	-0.08	-0.10	-0.04	-0.01	-0.08	0.82	0.73	-0.07	-0.04	-0.06	-0.11	-0.07	-0.17	-0.18	-0.05	-0.14	-0.05	-0.01
peptidyl-tyrosine phosphorylation	-0.08	-0.15	0.28	-0.08	0.16	0.31	-0.02	-0.16	0.30	0.40	-0.15	1.00	-0.05	0.01	-0.15	0.31	0.35	0.24	0.37	0.34	0.20	0.45	-0.10	0.33	0.34	-0.11	0.01	0.82	0.22	0.27	-0.18	-0.13	-0.10	0.34	0.22	0.28	0.24	0.22	0.03	-0.08	-0.05	-0.06	0.05	0.18
neutrophil chemotaxis	0.12	0.30	0.06	-0.13	0.23	0.29	-0.02	0.19	-0.07	-0.16	-0.15	-0.05	1.00	0.16	0.30	-0.17	-0.16	-0.02	-0.16	-0.15	-0.13	-0.08	-0.01	-0.18	-0.05	0.28	0.37	-0.01	-0.14	-0.13	0.25	-0.09	-0.13	0.15	0.18	-0.04	-0.02	0.01	0.14	0.87	0.18	0.27	-0.11	-0.13
positive regulation of defense response	0.04	0.00	0.18	-0.12	0.07	0.05	0.15	0.31	0.05	0.07	-0.20	0.01	0.16	1.00	0.42	0.00	0.03	0.07	0.12	0.16	0.04	0.07	-0.06	-0.10	0.09	0.45	0.60	0.02	0.05	0.02	-0.04	-0.14	-0.14	0.23	0.21	0.03	0.00	0.06	-0.02	0.12	0.41	0.55	0.08	-0.06
positive regulation of response to external stimulus	0.09	0.45	0.22	-0.06	0.38	0.32	0.05	0.55	-0.25	-0.24	-0.08	-0.15	0.30	0.42	1.00	-0.30	-0.27	0.19	-0.25	-0.22	0.09	-0.03	0.14	-0.37	-0.15	0.40	0.55	-0.08	0.10	0.07	0.55	0.03	-0.08	0.13	0.17	0.15	0.12	0.19	0.04	0.32	0.43	0.65	0.06	-0.15
response to insulin	-0.12	-0.17	0.03	-0.12	0.00	0.05	-0.16	-0.32	0.61	0.86	-0.20	0.31	-0.17	0.00	-0.30	1.00	0.93	0.00	0.61	0.67	0.04	0.24	-0.22	0.82	0.67	-0.18	-0.12	0.25	0.05	0.10	-0.20	-0.27	-0.14	0.16	0.05	0.11	0.07	0.06	0.14	-0.20	-0.14	-0.19	0.00	0.17
cellular response to insulin stimulus	-0.11	-0.16	-0.01	-0.11	0.03	0.08	-0.14	-0.28	0.67	0.93	-0.18	0.35	-0.16	0.03	-0.27	0.93	1.00	-0.06	0.67	0.73	-0.02	0.28	-0.21	0.75	0.73	-0.15	-0.11	0.29	0.00	0.04	-0.19	-0.25	-0.13	0.20	0.08	0.06	0.02	0.00	0.17	-0.19	-0.11	-0.17	0.02	0.21
mononuclear cell proliferation	-0.05	0.32	0.72	0.05	0.31	0.29	0.31	0.15	-0.05	-0.04	-0.02	0.24	-0.02	0.07	0.19	0.00	-0.06	1.00	-0.05	-0.02	0.89	0.29	0.14	0.14	-0.02	0.02	0.05	0.34	0.76	0.75	0.26	0.07	0.03	0.15	0.19	0.91	0.80	0.73	0.07	0.16	0.01	0.17	-0.02	-0.07
Fc receptor signaling pathway	-0.19	-0.16	0.08	-0.10	-0.03	0.10	0.04	-0.32	0.78	0.73	-0.17	0.37	-0.16	0.12	-0.25	0.61	0.67	-0.05	1.00	0.92	-0.01	0.29	-0.12	0.53	0.84	-0.21	-0.10	0.31	0.10	0.14	-0.19	-0.16	-0.03	0.06	-0.07	0.00	-0.05	0.01	0.10	-0.19	-0.15	-0.16	0.03	0.23
Fc-epsilon receptor signaling pathway	-0.18	-0.15	0.03	-0.08	0.00	0.13	-0.11	-0.28	0.84	0.80	-0.15	0.34	-0.15	0.16	-0.22	0.67	0.73	-0.02	0.92	1.00	0.01	0.34	-0.10	0.58	0.92	-0.18	-0.08	0.27	0.05	0.08	-0.18	-0.13	-0.01	0.09	-0.04	0.03	-0.02	0.03	0.13	-0.18	-0.11	-0.14	0.05	0.27
T cell proliferation	-0.03	0.25	0.71	0.08	0.18	0.14	0.37	0.07	-0.01	0.00	0.01	0.20	-0.13	0.04	0.09	0.04	-0.02	0.89	-0.01	0.01	1.00	0.34	0.07	0.18	0.01	-0.02	-0.03	0.29	0.85	0.84	0.20	0.04	0.07	0.10	0.14	0.80	0.68	0.58	0.10	-0.03	-0.01	0.12	-0.13	-0.04
positive regulation of tyrosine phosphorylation of Stat5 protein	0.05	-0.08	0.25	0.05	0.31	0.39	0.03	0.01	0.29	0.31	0.01	0.45	-0.08	0.07	-0.03	0.24	0.28	0.29	0.29	0.34	0.34	1.00	-0.10	0.24	0.34	-0.01	0.05	0.59	0.27	0.32	-0.09	-0.02	0.04	0.01	0.02	0.36	0.29	0.38	-0.09	-0.09	-0.04	0.01	0.08	-0.10
regulation of cysteine-type endopeptidase activity involved in ap	-0.04	0.11	0.10	-0.04	0.16	0.02	-0.07	-0.01	0.05	-0.20	0.36	-0.10	-0.01	-0.06	0.14	-0.22	-0.21	0.14	-0.12	-0.10	0.07	-0.10	1.00	-0.11	-0.10	-0.04	-0.04	-0.06	0.02	0.05	0.18	0.65	0.05	-0.01	0.02	0.10	0.03	0.09	-0.14	-0.04	0.10	0.10	-0.01	0.05
response to peptide hormone	-0.09	-0.18	0.16	-0.09	-0.02	0.05	-0.13	-0.32	0.53	0.68	-0.12	0.33	-0.18	-0.10	-0.37	0.82	0.75	0.14	0.53	0.58	0.18	0.24	-0.11	1.00	0.58	-0.18	-0.16	0.29	0.18	0.23	-0.22	-0.16	-0.04	0.20	0.11	0.22	0.20	0.14	0.14	-0.16	-0.19	-0.17	-0.04	0.16
cellular response to fibroblast growth factor stimulus	-0.18	-0.05	0.03	-0.08	0.08	0.22	-0.11	-0.22	0.84	0.80	-0.15	0.34	-0.05	0.09	-0.15	0.67	0.73	-0.02	0.84	0.92	0.01	0.34	-0.10	0.58	1.00	-0.18	-0.08	0.27	0.05	0.08	-0.08	-0.13	-0.01	0.09	-0.04	0.03	-0.02	0.03	0.13	-0.08	-0.11	-0.14	0.05	0.27
regulation of inflammatory response	0.32	-0.07	0.06	-0.02	0.03	0.08	0.02	0.65	-0.21	-0.20	-0.03	-0.11	0.28	0.45	0.40	-0.18	-0.15	0.02	-0.21	-0.18	-0.02	-0.01	-0.04	-0.18	-0.18	1.00	0.66	-0.04	-0.08	-0.04	-0.11	-0.04	-0.04	0.20	0.24	0.06	0.10	0.08	0.00	0.32	0.78	0.61	0.11	-0.13
positive regulation of inflammatory response	0.20	0.00	0.12	-0.03	0.18	0.24	0.05	0.38	-0.10	-0.09	-0.08	0.01	0.37	0.60	0.55	-0.12	-0.11	0.05	-0.10	-0.08	-0.03	0.05	-0.04	-0.16	-0.08	0.66	1.00	0.07	-0.07	-0.04	-0.03	-0.04	-0.04	0.29	0.34	0.02	0.05	0.10	0.10	0.31	0.48	0.80	0.12	-0.04
regulation of peptidyl-tyrosine phosphorylation	-0.04	-0.13	0.38	-0.04	0.25	0.41	0.03	-0.07	0.23	0.33	-0.10	0.82	-0.01	0.02	-0.08	0.25	0.29	0.34	0.31	0.27	0.29	0.59	-0.06	0.29	0.27	-0.04	0.07	1.00	0.31	0.37	-0.15	-0.06	-0.06	0.18	0.21	0.38	0.34	0.31	0.09	-0.04	-0.10	0.00	-0.01	0.05
regulation of T cell activation	-0.07	0.18	0.87	0.04	0.12	0.08	0.38	0.06	0.02	0.03	-0.04	0.22	-0.14	0.05	0.10	0.05	0.00	0.76	0.10	0.05	0.85	0.27	0.02	0.18	0.05	-0.08	-0.07	0.31	1.00	0.90	0.14	-0.02	0.02	0.13	0.17	0.69	0.57	0.47	0.16	-0.07	-0.07	0.06	-0.14	0.02
positive regulation of T cell activation	-0.04	0.22	0.86	0.07	0.16	0.12	0.34	0.05	0.05	0.07	-0.01	0.27	-0.13	0.02	0.07	0.10	0.04	0.75	0.14	0.08	0.84	0.32	0.05	0.23	0.08	-0.04	-0.04	0.37	0.90	1.00	0.18	0.02	0.05	0.18	0.21	0.67	0.54	0.54	0.20	-0.04	-0.03	0.10	-0.13	0.05
positive regulation of chemotaxis	0.08	0.87	0.22	-0.03	0.36	0.24	0.16	0.20	-0.19	-0.18	-0.08	-0.18	0.25	-0.04	0.55	-0.20	-0.19	0.26	-0.19	-0.18	0.20	-0.09	0.18	-0.22	-0.08	-0.11	-0.03	-0.15	0.14	0.18	1.00	0.04	-0.04	-0.08	-0.07	0.22	0.16	0.22	-0.02	0.31	-0.01	0.12	0.00	-0.15
regulation of endopeptidase activity	0.20	0.00	0.03	0.60	0.07	-0.02	-0.08	0.08	-0.02	-0.22	0.82	-0.13	-0.09	-0.14	0.03	-0.27	-0.25	0.07	-0.16	-0.13	0.04	-0.02	0.65	-0.16	-0.13	-0.04	-0.04	-0.06	-0.02	0.02	0.04	1.00	0.65	0.01	0.05	0.03	0.00	0.06	-0.19	-0.12	0.04	0.03	-0.09	-0.06
peptidase regulator activity	0.29	-0.01	0.00	0.95	-0.02	0.02	-0.07	0.05	-0.03	-0.11	0.73	-0.10	-0.13	-0.14	-0.08	-0.14	-0.13	0.03	-0.03	-0.01	0.07	0.04	0.05	-0.04	-0.01	-0.04	-0.04	-0.06	0.02	0.05	-0.04	0.65	1.00	-0.01	0.02	0.00	-0.07	-0.03	-0.14	-0.15	-0.10	-0.09	-0.13	-0.16
ERK1 and ERK2 cascade	0.10	-0.05	0.20	0.01	0.16	0.13	-0.02	0.14	0.06	0.16	-0.07	0.34	0.15	0.23	0.13	0.16	0.20	0.15	0.06	0.09	0.10	0.01	-0.01	0.20	0.09	0.20	0.29	0.18	0.13	0.18	-0.08	0.01	-0.01	1.00	0.91	0.20	0.24	0.13	0.42	0.10	0.28	0.28	-0.05	0.27
regulation of ERK1 and ERK2 cascade	0.14	-0.03	0.24	0.04	0.20	0.17	0.00	0.18	-0.07	0.03	-0.04	0.22	0.18	0.21	0.17	0.05	0.08	0.19	-0.07	-0.04	0.14	0.02	0.02	0.11	-0.04	0.24	0.34	0.21	0.17	0.21	-0.07	0.05	0.02	0.91	1.00	0.24	0.29	0.16	0.47	0.14	0.19	0.33	-0.14	0.12
leukocyte proliferation	0.02	0.27	0.65	0.02	0.34	0.33	0.26	0.16	0.00	0.01	-0.06	0.28	-0.04	0.03	0.15	0.11	0.06	0.91	0.00	0.03	0.80	0.36	0.10	0.22	0.03	0.06	0.02	0.38	0.69	0.67	0.22	0.03	0.00	0.20	0.24	1.00	0.91	0.75	0.04	0.12	0.04	0.13	-0.04	-0.09
regulation of leukocyte proliferation	0.05	0.20	0.54	-0.05	0.31	0.29	0.21	0.21	-0.05	-0.04	-0.11	0.24	-0.02	0.00	0.12	0.07	0.02	0.80	-0.05	-0.02	0.68	0.29	0.03	0.20	-0.02	0.10	0.05	0.34	0.57	0.54	0.16	0.00	-0.07	0.24	0.29	0.91	1.00	0.84	0.07	0.16	0.08	0.17	-0.02	-0.07
positive regulation of leukocyte proliferation	-0.02	0.28	0.55	-0.02	0.39	0.37	0.18	0.22	0.01	0.02	-0.07	0.22	0.01	0.06	0.19	0.06	0.00	0.73	0.01	0.03	0.58	0.38	0.09	0.14	0.03	0.08	0.10	0.31	0.47	0.54	0.22	0.06	-0.03	0.13	0.16	0.75	0.84	1.00	0.00	0.22	0.08	0.24	0.01	-0.14
cellular response to interferon-gamma	-0.14	0.01	0.14	-0.14	0.02	0.05	-0.04	-0.03	0.10	0.20	-0.17	0.03	0.14	-0.02	0.04	0.14	0.17	0.07	0.10	0.13	0.10	-0.09	-0.14	0.14	0.13	0.00	0.10	0.09	0.16	0.20	-0.02	-0.19	-0.14	0.42	0.47	0.04	0.07	0.00	1.00	0.10	-0.06	0.04	-0.12	0.31
granulocyte chemotaxis	0.08	0.37	0.12	-0.14	0.27	0.34	0.05	0.26	-0.10	-0.18	-0.18	-0.08	0.87	0.12	0.32	-0.20	-0.19	0.16	-0.19	-0.18	-0.03	-0.09	-0.04	-0.16	-0.08	0.32	0.31	-0.04	-0.07	-0.04	0.31	-0.12	-0.15	0.10	0.14	0.12	0.16	0.22	0.10	1.00	0.20	0.32	0.00	-0.15
regulation of response to wounding	0.27	-0.04	0.04	-0.08	0.00	-0.01	0.01	0.58	-0.09	-0.13	-0.05	-0.05	0.18	0.41	0.43	-0.14	-0.11	0.01	-0.15	-0.11	-0.01	-0.04	0.10	-0.19	-0.11	0.78	0.48	-0.10	-0.07	-0.03	-0.01	0.04	-0.10	0.28	0.19	0.04	0.08	0.08	-0.06	0.20	1.00	0.64	0.11	0.04
positive regulation of response to wounding	0.22	0.06	0.22	-0.08	0.18	0.15	0.17	0.46	-0.16	-0.15	-0.14	-0.06	0.27	0.55	0.65	-0.19	-0.17	0.17	-0.16	-0.14	0.12	0.01	0.10	-0.17	-0.14	0.61	0.80	0.00	0.06	0.10	0.12	0.03	-0.09	0.28	0.33	0.13	0.17	0.24	0.04	0.32	0.64	1.00	0.06	-0.09
TNF signaling pathway	0.00	0.03	-0.15	-0.13	0.04	0.07	-0.14	0.06	0.12	0.04	-0.05	0.05	-0.11	0.08	0.06	0.00	0.02	-0.02	0.03	0.05	-0.13	0.08	-0.01	-0.04	0.05	0.11	0.12	-0.01	-0.14	-0.13	0.00	-0.09	-0.13	-0.05	-0.14	-0.04	-0.02	0.01	-0.12	0.00	0.11	0.06	1.00	0.22
Proteoglycans in cancer	-0.15	-0.13	0.00	-0.15	-0.20	-0.18	-0.17	-0.19	0.40	0.25	-0.01	0.18	-0.13	-0.06	-0.15	0.17	0.21	-0.07	0.23	0.27	-0.04	-0.10	0.05	0.16	0.27	-0.13	-0.04	0.05	0.02	0.05	-0.15	-0.06	-0.16	0.27	0.12	-0.09	-0.07	-0.14	0.31	-0.15	0.04	-0.09	0.22	1.00

Figure A.S4. Kappa statistics from the downregulated proteome.

Level of agreement between gene terms is measured by Kappa statistics (default < 4). Blue scale depicts level of agreement from very high (1) to very low (-1)

P13	C3	SERPINA5	TNFSF14	GSK3B	GSK3A		P Value
■	■	■	■	■	■	negative regulation of catalytic activity	0.0246
■	■	■	■	□	□	peptidase inhibitor activity	0.0095
■	■	■	■	□	□	endopeptidase regulator activity	0.009
■	■	■	■	□	□	peptidase regulator activity	0.0164
■	■	■	■	□	□	endopeptidase inhibitor activity	0.0082

Figure A.S5. DAVID analysis by GO Biological process of the 8 protein identified LASSO signature.

APPENDIX B. Supplement to: Identification of a Unique Temporal Signature in Blood and BAL Associated With IPF Progression

Katy C. Norman[1], David N. O'Dwyer[2], Margaret L. Salisbury[3], Katarina M. DiLillo[1], Vibha N. Lama[2], Meng Xia[4], Stephen J. Gurczynski[2], Eric S. White[2], Kevin R. Flaherty[2], Fernando J. Martinez[5], Susan Murray[4], Bethany B. Moore[2,6], and Kelly B. Arnold[1]

[1] Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI 48109

[2] Department of Internal Medicine, Division of Pulmonary and Critical Care Medicine, University of Michigan Medical School, Ann Arbor, MI, USA

[3] Department of Medicine, Division of Allergy, Pulmonary and Critical Care Medicine Vanderbilt University Medical Center, Nashville, TN, USA

[4] Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, USA

[5] Department of Internal Medicine, Weill Cornell School of Medicine, New York, NY, USA

[6] Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, MI, USA

Scientific Reports. 2020: In Press.

Supplemental methods: Sample acquisitions and measurements

Peripheral blood samples were collected from 60 COMET patients at three time points (week 0/baseline, week 48 and week 80) in EDTA-containing vacutainers and were shipped overnight from individual sites to the University of Michigan. Blood samples were centrifuged and plasma was stored at -80°C until transported to SomaLogic (Boulder, CO). Slow off-rate modified aptamers (SOMAmer©) technology was used to measure 1129 proteins present in blood samples at each collection time point.

Bronchoscopy was performed at enrollment in patients who were clinically stable and without evidence of active infection. BAL samples were collected and pooled from 4 installations of 50 mL sterile isotonic saline aliquots. Cell-free fluid was stored at -80°C. Luminex FlexMAP 3D (Luminex Corporation, Austin, TX) technology was used to measure 29 cytokines/chemokines in the BAL samples. Samples below the lower limit of detection were set to be ½ the lowest minimum detectable concentration across the standard curves of all analytes. Before inclusion in any analyses, all BAL protein concentrations were normalized to total protein concentration as quantified by a Pierce BCA Protein Assay Kit (Pierce Protein Biology, Rockford, IL).

Table B.S1. Demographic and lung function test descriptions from progressors and non-progressors whose baseline blood protein measurements were used in creating models based on blood proteins alone.

	Non-progressor (N=25)	**Progressor (N=34)**	**P-value**
Age	63.72	64.86	0.5855
Sex (Male)	76%	61.76%	0.2551
Number Never Smokers	7	12	0.5614
Number Former Smokers	17	22	0.796
Number Current Smokers	1	0	0.2469
FVC % Predicted	68.19	70.78	0.5511
DLCO % Predicted	44.75	47.61	0.441

Table B.S2. Demographic and lung function test descriptions from progressors and non-progressors whose baseline BAL protein measurements were used in creating models based on BAL proteins alone.

	Non-progressor (N=20)	**Progressor (N=31)**	**P-value**
Age	62.43	65.43	0.1924
Sex (Male)	16 (80%)	20 (64.5%)	0.2446
Number Never Smokers	6 (30%)	11 (35.48%)	0.6922
Number Former Smokers	13 (65%)	20 (65.42%)	0.9725
Number Current Smokers	1 (5%)	0	0.2165
FVC % Predicted	66.88%	71.84%	0.3248
DLCO % Predicted	45.78%	47.42%	0.6803

Table B.S3. Pearson’s correlation between proteins measured by SOMAmer aptamers and by ELISA in a subset of the COMET samples.

	Pearson's correlation coefficient	P-value
CCL22	0.672	0.006
CCL18	0.706	0.003
CCL2	0.566	0.028
IL-10	-0.208	0.456
CXCL12	-0.081	0.775

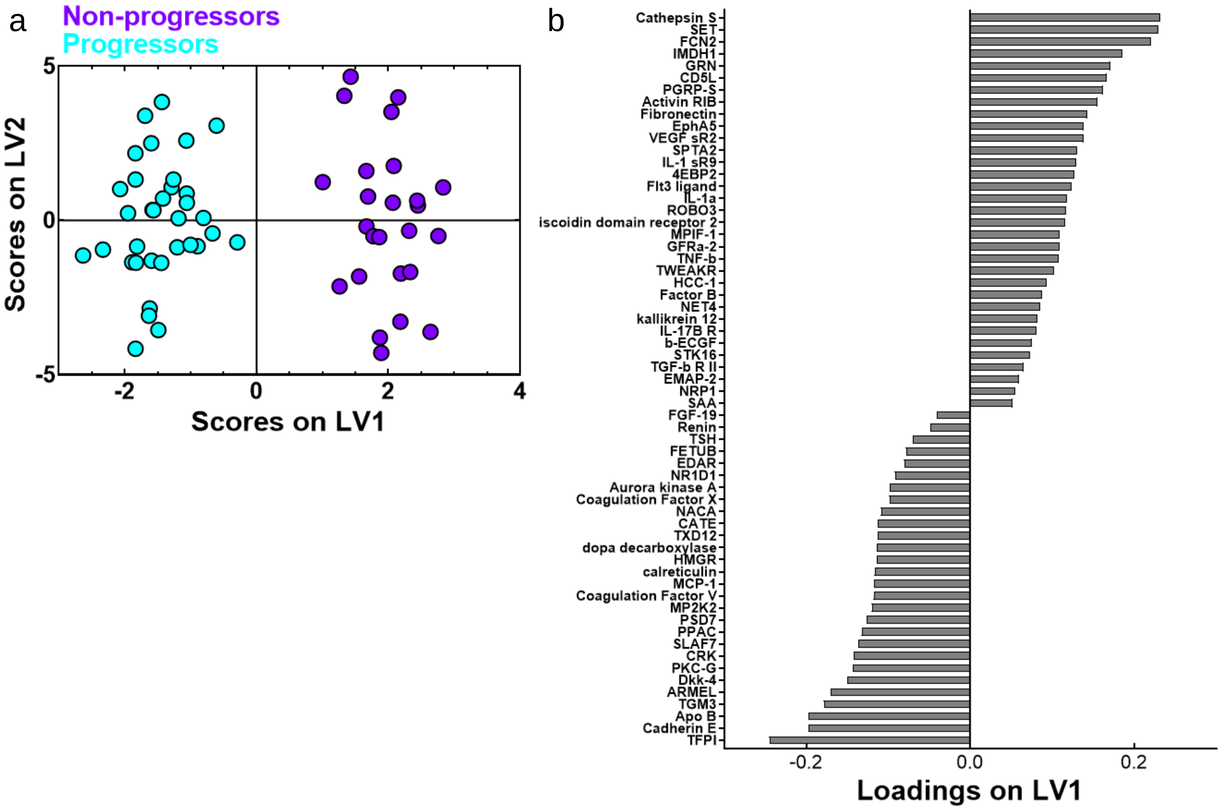


Figure B.S1. The PLSDA model based on LASSO-identified signature of blood proteins is accurately able to differentiate IPF progressors and non-progressors.

(**a**) LASSO identified a signature of 61 blood proteins that differentiated progressors and non-progressors with 100% calibration and 96.53% cross-validation accuracy. (**b**) The associated loadings on latent variable 1 (LV1) captured 6.28% of the total variance in the data. Proteins that are loaded negatively on LV1 are comparatively upregulated in IPF progressors, and positively loaded proteins have a comparative reduction in IPF progressors.

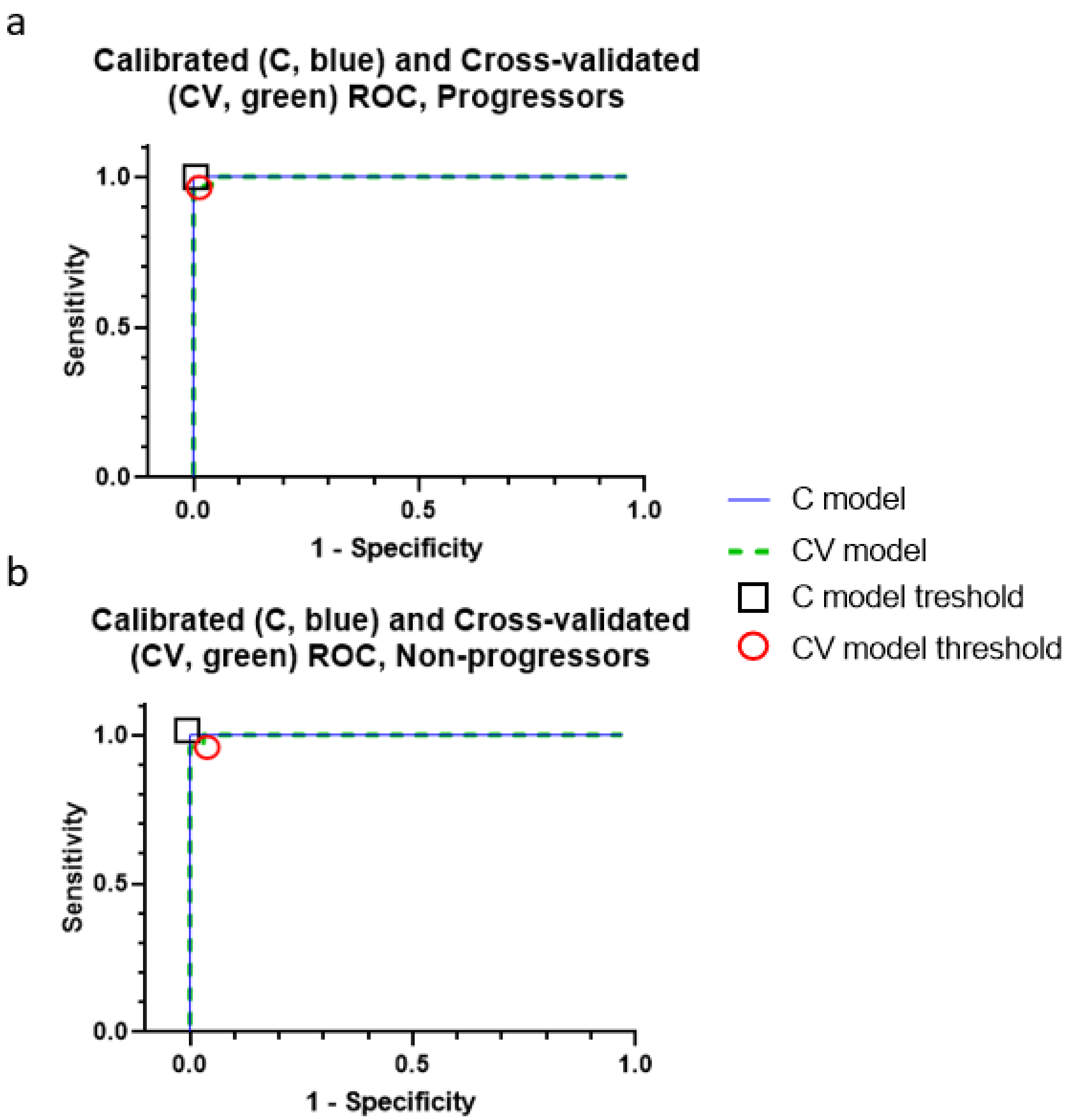


Figure B.S2. The receiver operator characteristic (ROC) curves associated with the PLSDA model based on the LASSO-identified signature of 61 blood proteins.

(**a**) The cross-validated PLSDA model reported a sensitivity of 97.06% and a specificity of 99.56% for the progressors.
(**b**) The cross-validated PLSDA model reported a sensitivity of 96% and specificity of 97.38% for the non-progressors. C: Calibrated; CV = Cross-validated; AUC = area under curve.

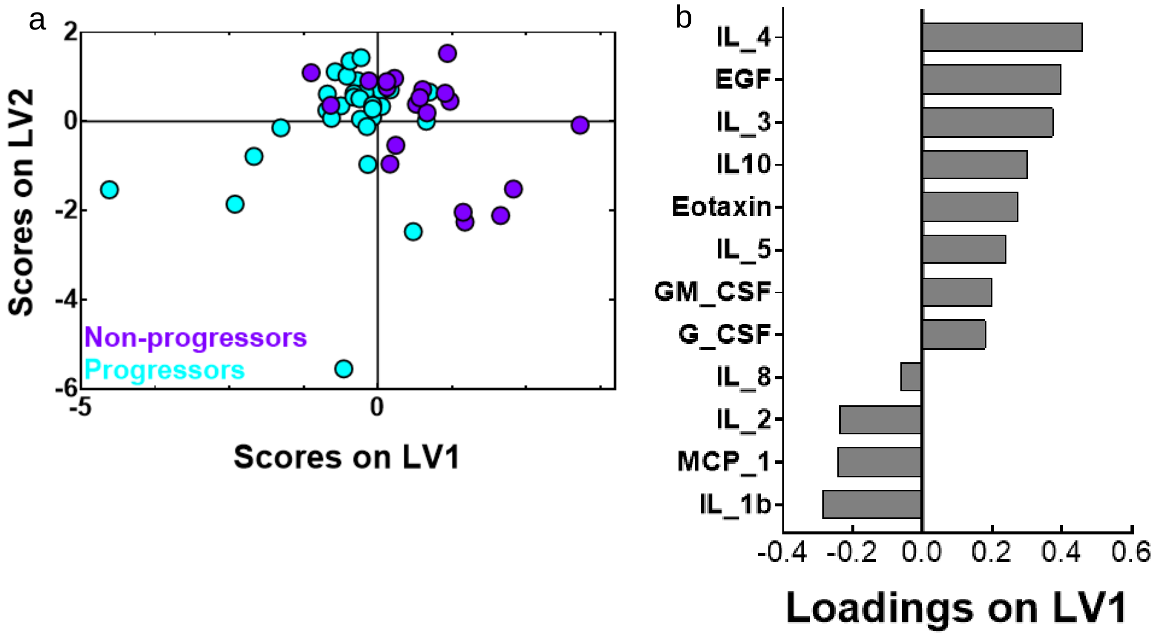


Figure B.S3. PLSDA model based on VIP-selected signature of BAL proteins is moderately able to differentiate IPF progressors and non-progressors, with 78.55% calibration and 67.82% cross-validation accuracy.

(**a**) The PLSDA scores plot of the 12 feature BAL protein signature highlights moderate separation between baseline progressors and non-progressors, with progressors generally having negative scores on LV1 and non-progressors having positive scores. (**b**) The associated loadings on LV1 captured 16.49% of the total variance in the data. Proteins that are loaded negatively on LV1 are comparatively upregulated in IPF progressors, and positively loaded proteins have a comparative reduction in IPF progressors.

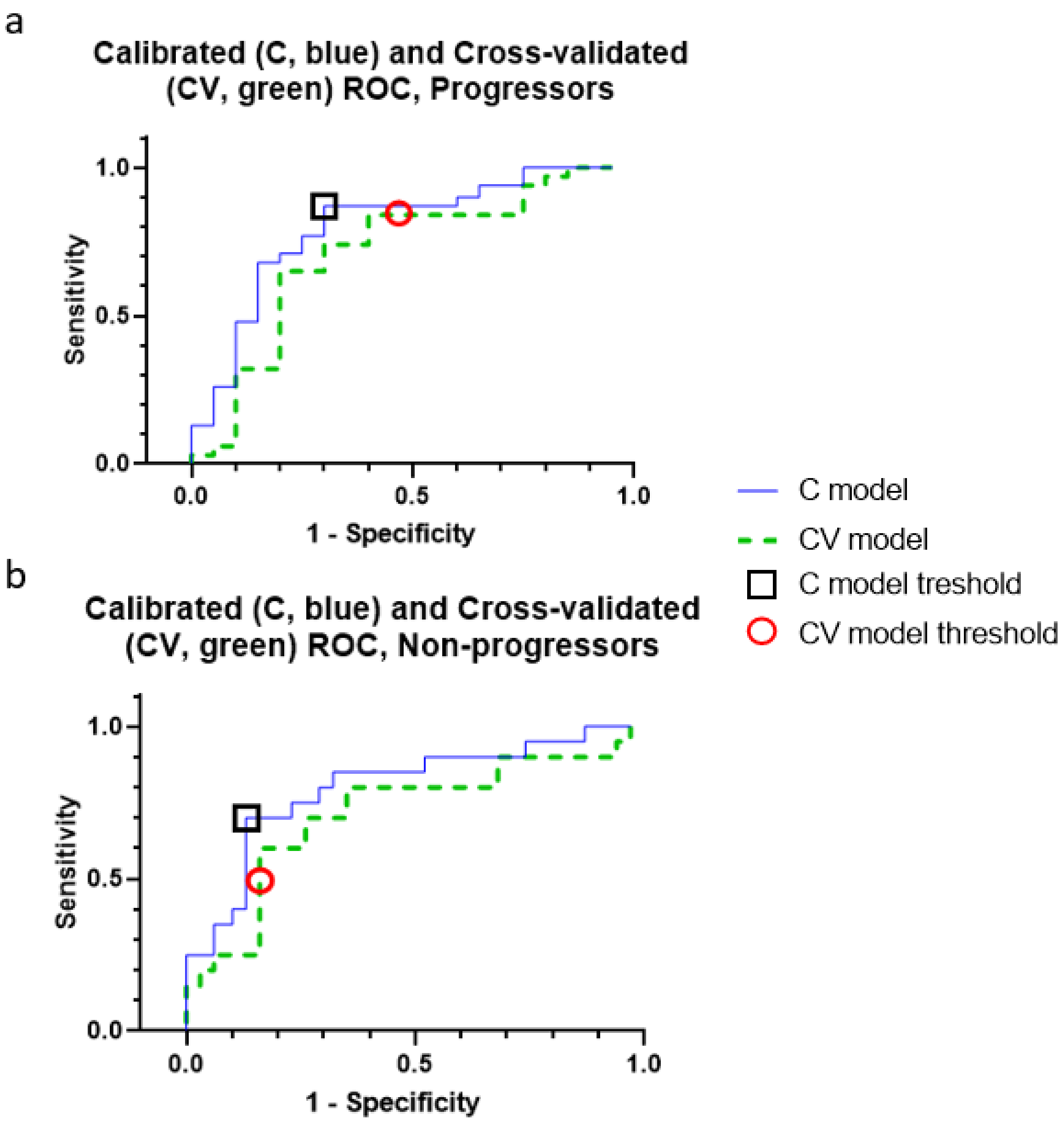


Figure B.S4. The receiver operator characteristic (ROC) curves associated with the PLSDA model based on the VIP-selected signature of BAL proteins.

(**a**) The cross-validated PLSDA model reported a sensitivity of 83.87% and a specificity of 54.94% for the progressors. (**b**) The cross-validated PLSDA model reported a sensitivity of 49.94% and specificity of 83.87% for the non-progressors. C: Calibrated; CV = Cross-validated; AUC = area under curve.

Figure B.S5

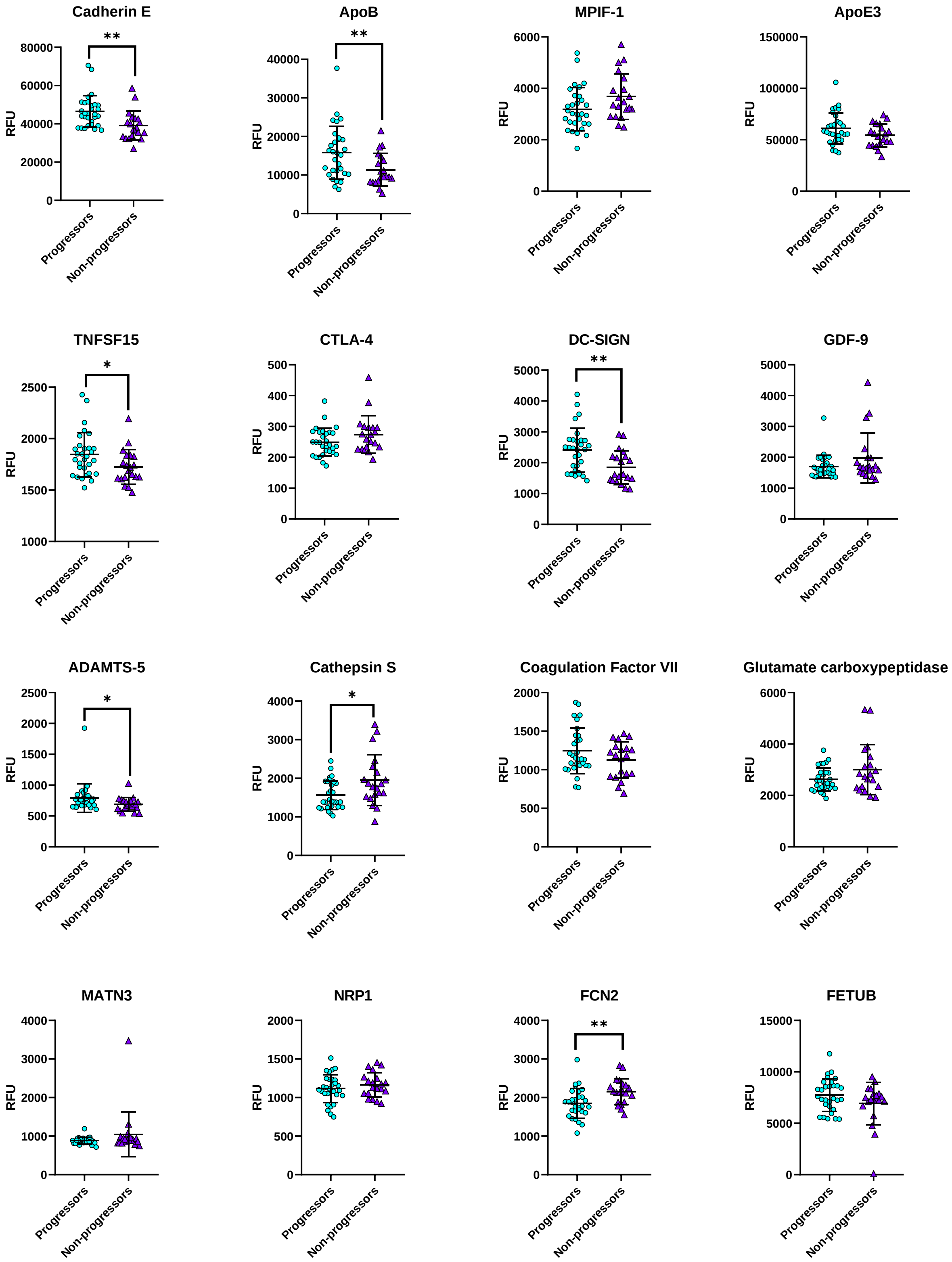

Cadherin E
**
80000
60000
40000
20000
0
RFU
Progressors
Non-progressors
ApoB
**
40000
30000
20000
10000
0
RFU
Progressors
Non-progressors
MPIF-1
6000
4000
2000
0
RFU
Progressors
Non-progressors
ApoE3
150000
100000
50000
0
RFU
Progressors
Non-progressors
TNFSF15
*
2500
2000
1500
1000
RFU
Progressors
Non-progressors
CTLA-4
500
400
300
200
100
0
RFU
Progressors
Non-progressors
DC-SIGN
**
5000
4000
3000
2000
1000
0
RFU
Progressors
Non-progressors
GDF-9
5000
4000
3000
2000
1000
0
RFU
Progressors
Non-progressors
ADAMTS-5
*
2500
2000
1500
1000
500
0
RFU
Progressors
Non-progressors
Cathepsin S
*
4000
3000
2000
1000
0
RFU
Progressors
Non-progressors
Coagulation Factor VII
2000
1500
1000
500
0
RFU
Progressors
Non-progressors
Glutamate carboxypeptidase
6000
4000
2000
0
RFU
Progressors
Non-progressors
MATN3
4000
3000
2000
1000
0
RFU
Progressors
Non-progressors
NRP1
2000
1500
1000
500
0
RFU
Progressors
Non-progressors
FCN2
**
4000
3000
2000
1000
0
RFU
Progressors
Non-progressors
FETUB
15000
10000
5000
0
RFU
Progressors
Non-progressors

Figure B.S5 continued

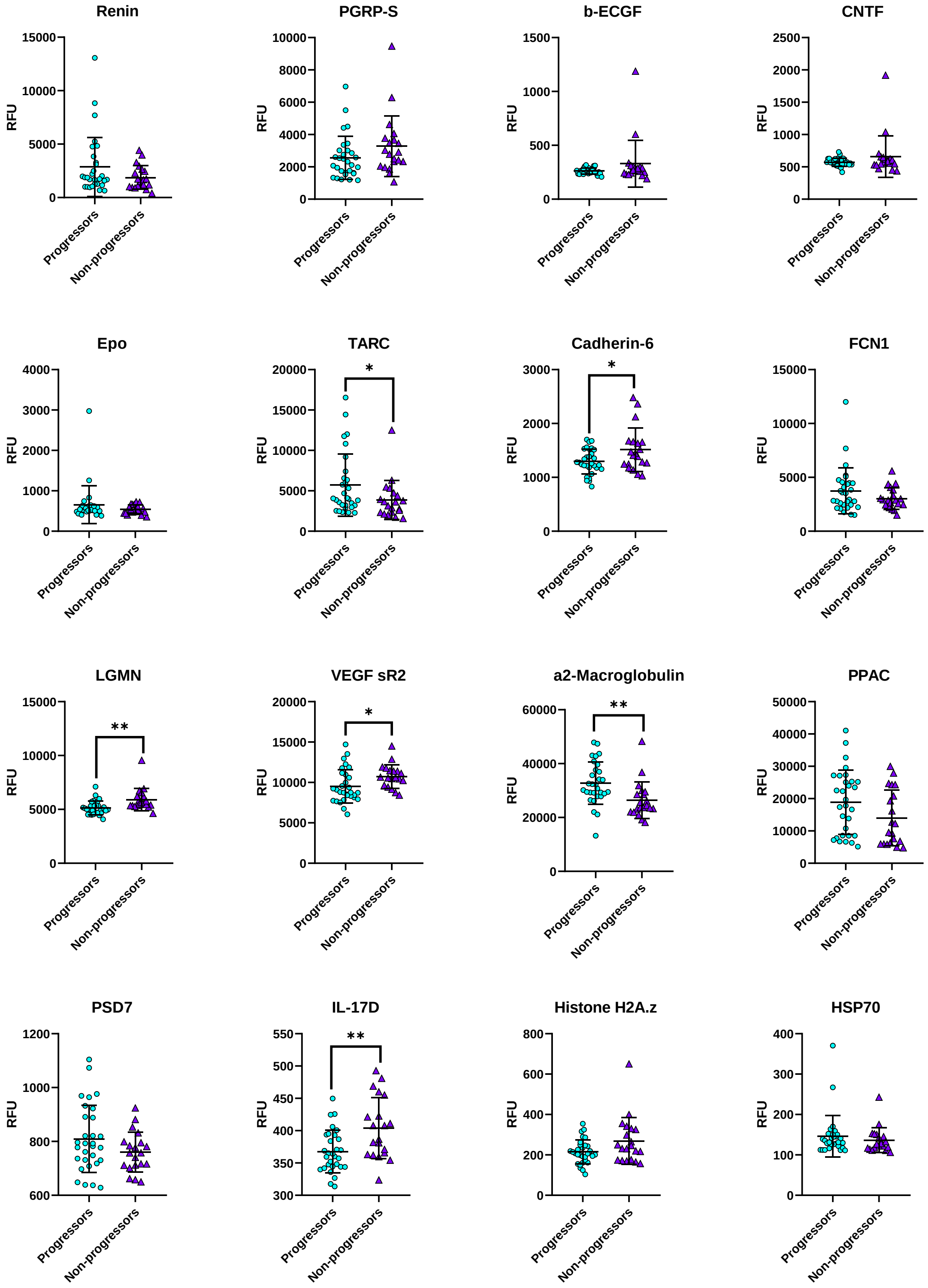
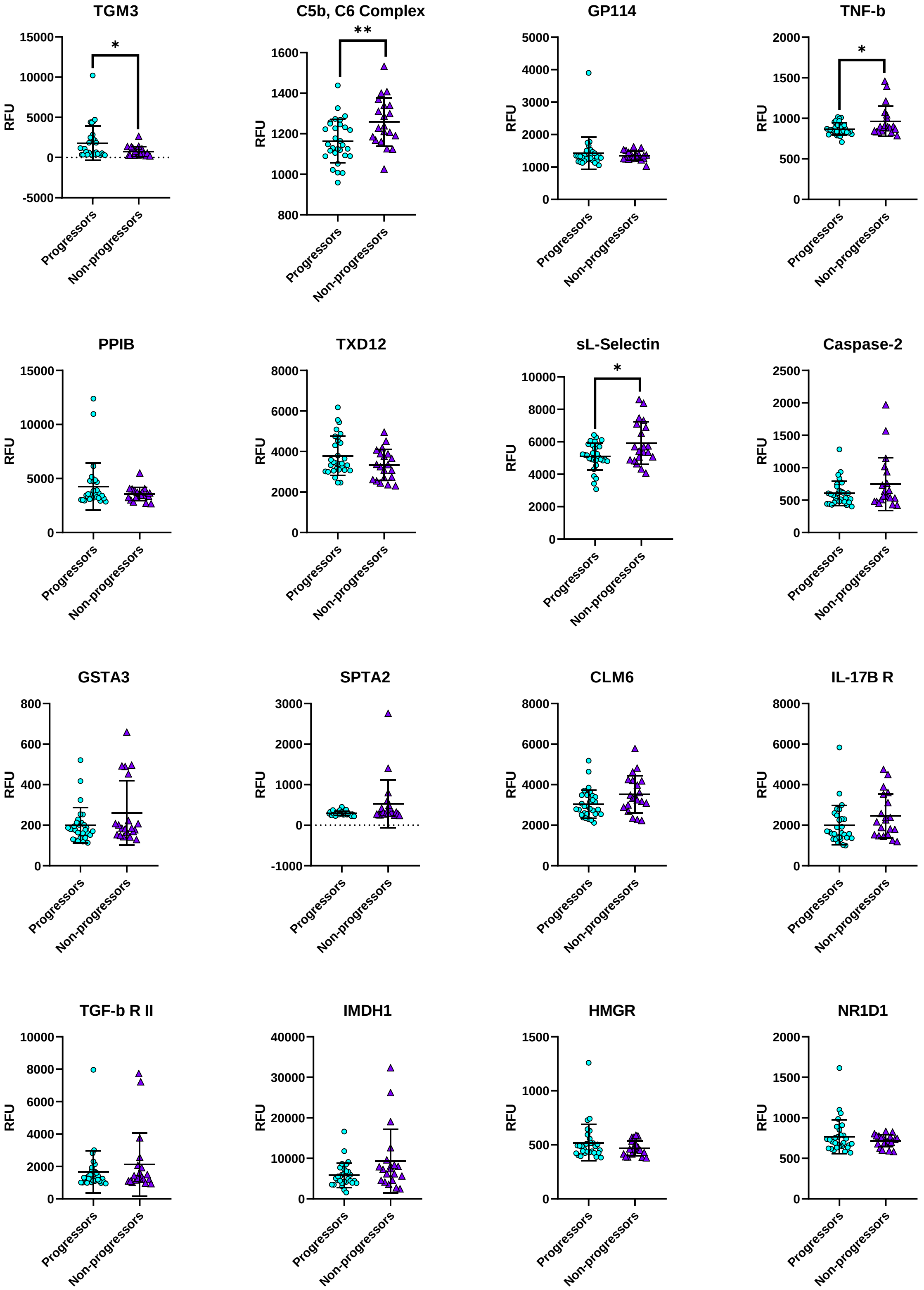


Figure B.S5 continued



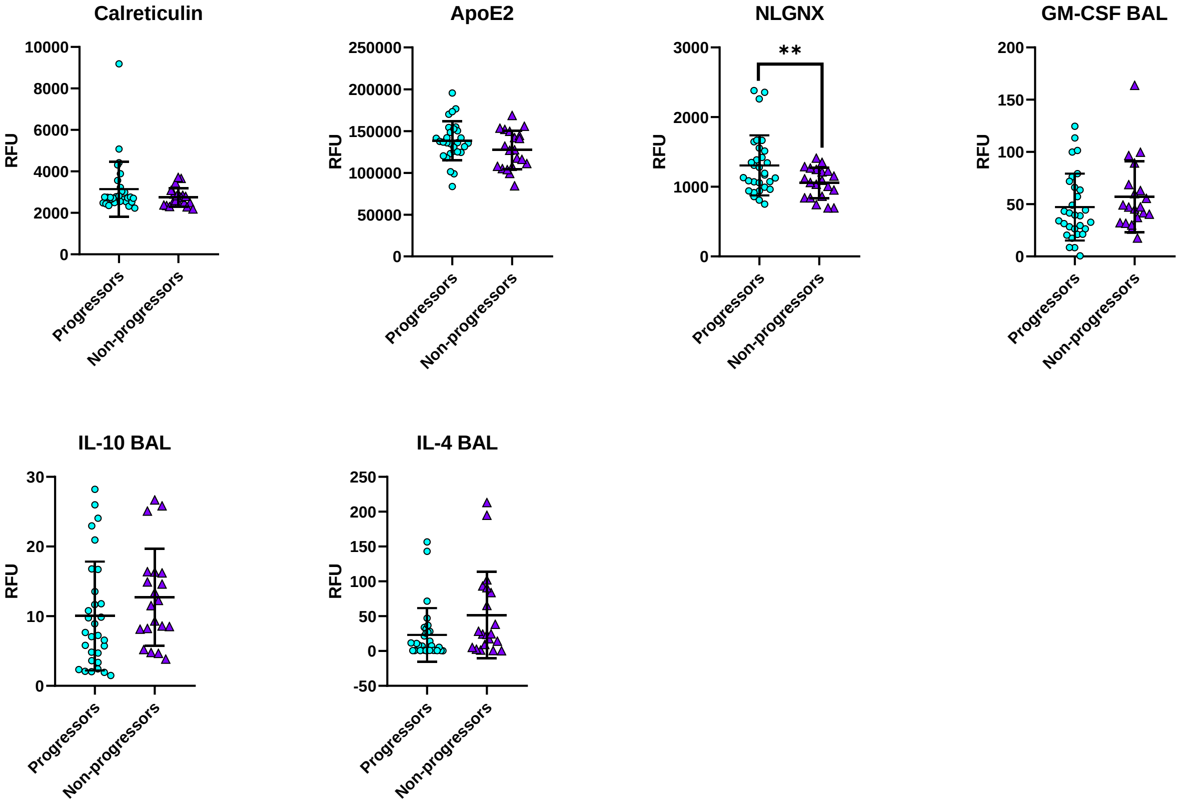


Figure B.S5. Direct comparison of expression of the blood and BAL proteins in the LASSO-identified signature in both progressors and non-progressors.

Significance according to a two-sample t-test is marked on each graph, with ** indicating p < 0.01 and * indicating p < 0.05.

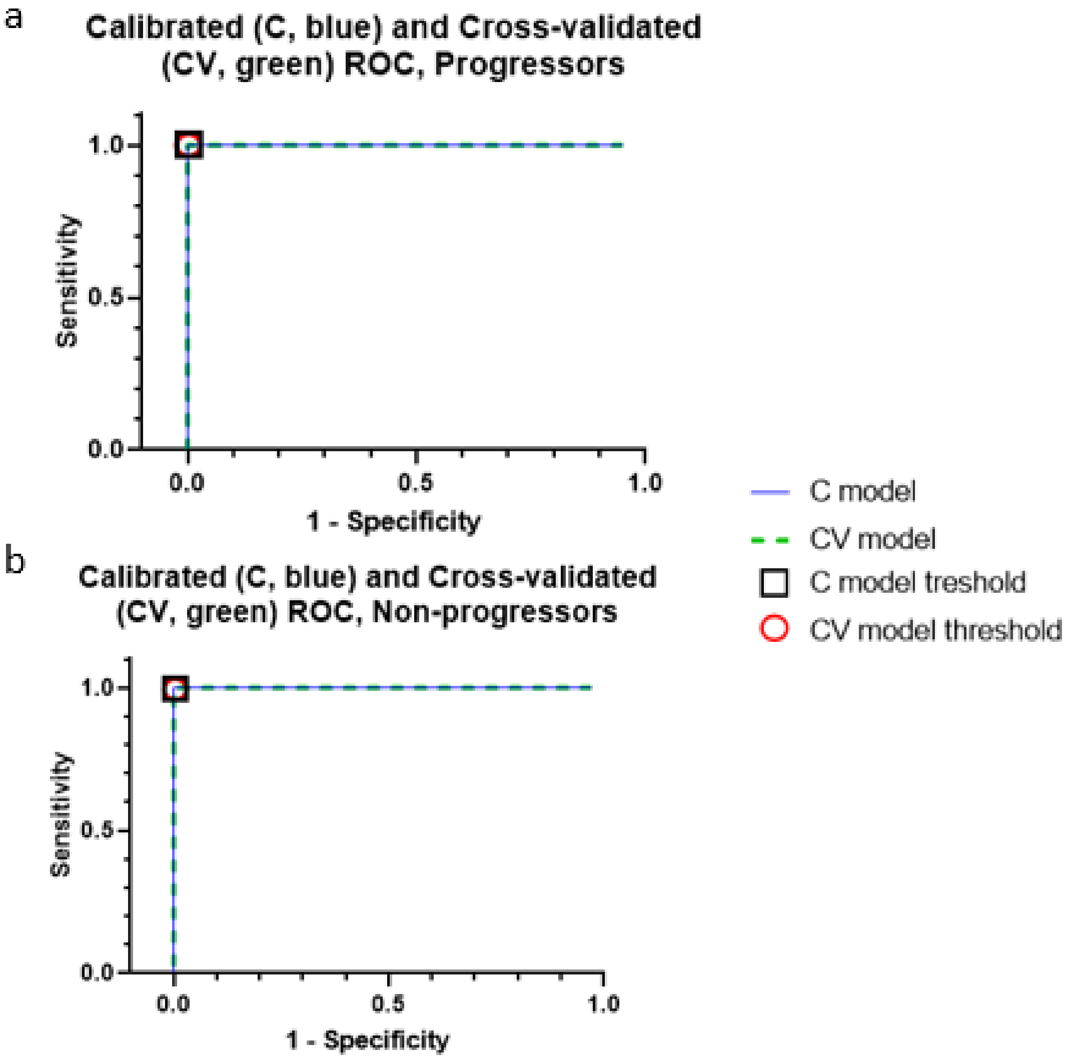


Figure B.S6. The receiver operator characteristic (ROC) curves associated with the PLSDA model based on the LASSO-identified signature of blood BAL proteins.

(a) The cross-validated PLSDA model reported a sensitivity of 100% and a specificity of 100% for the progressors. **(b)** The cross-validated PLSDA model reported a sensitivity of 100% and specificity of 100% for the non-progressors. C: Calibrated; CV = Cross-validated; AUC = area under curve.

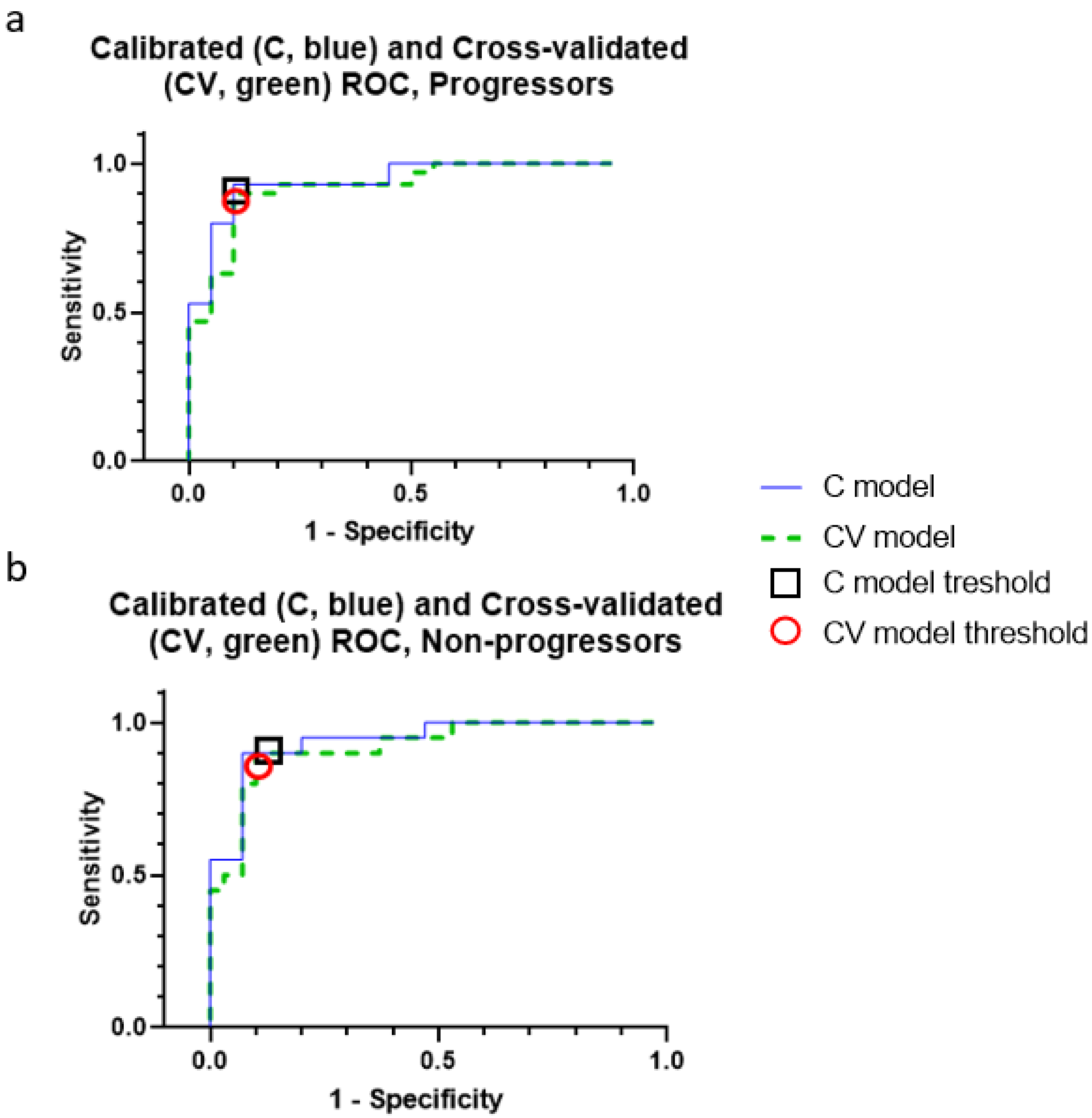


Figure B.S7. The receiver operator characteristic (ROC) curves associated with the PLSDA model based on the 28 proteins that were identified as being significantly differentially expressed across progressors and non-progressors in the volcano plot.

(**a**) The cross-validated PLSDA model reported a sensitivity of 88.29% and a specificity of 87.56% for the progressors. (**b**) The cross-validated PLSDA model reported a sensitivity of 90% and specificity of 90% for the non-progressors. C: Calibrated; CV = Cross-validated; AUC = area under curve.

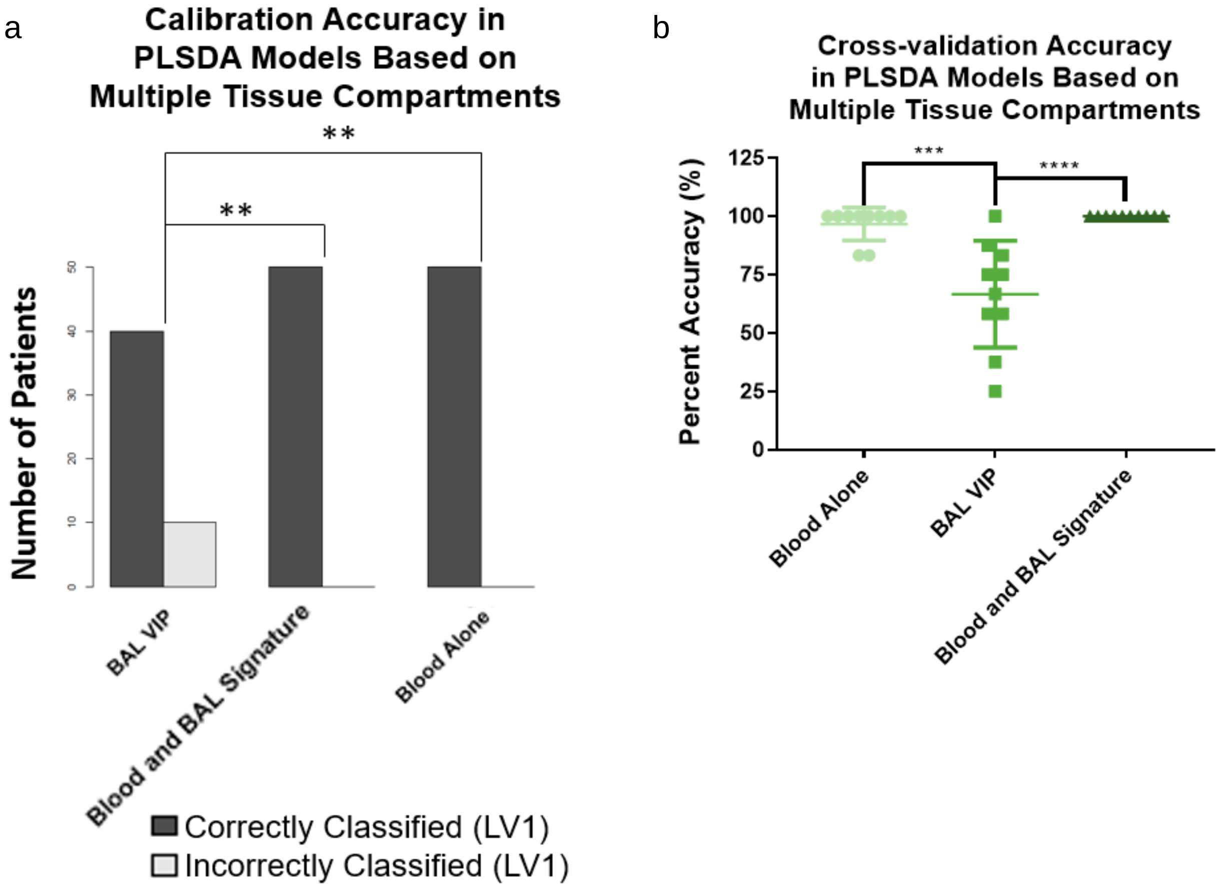


Figure B.S8. Comparison of calibration and cross-validation accuracy in the PLSDA models based on blood proteins alone, BAL proteins alone, and blood and BAL proteins combined shows that the blood only and the combination model are both significantly better than the model based on BAL proteins alone.

(**a**) Comparison of the calibration accuracies in the three PLSDA models. ** indicate $p < 0.01$ after administration of Cochran's Q test with McNemar's post hoc test. (**b**) Comparison of the cross-validation accuracies of the same three PLSDA models shown in panel **a**. *** indicates $p = 0.0001$, and **** indicates $p < 0.0001$ after administration of a one-way ANOVA with Tukey's multiple comparison test.

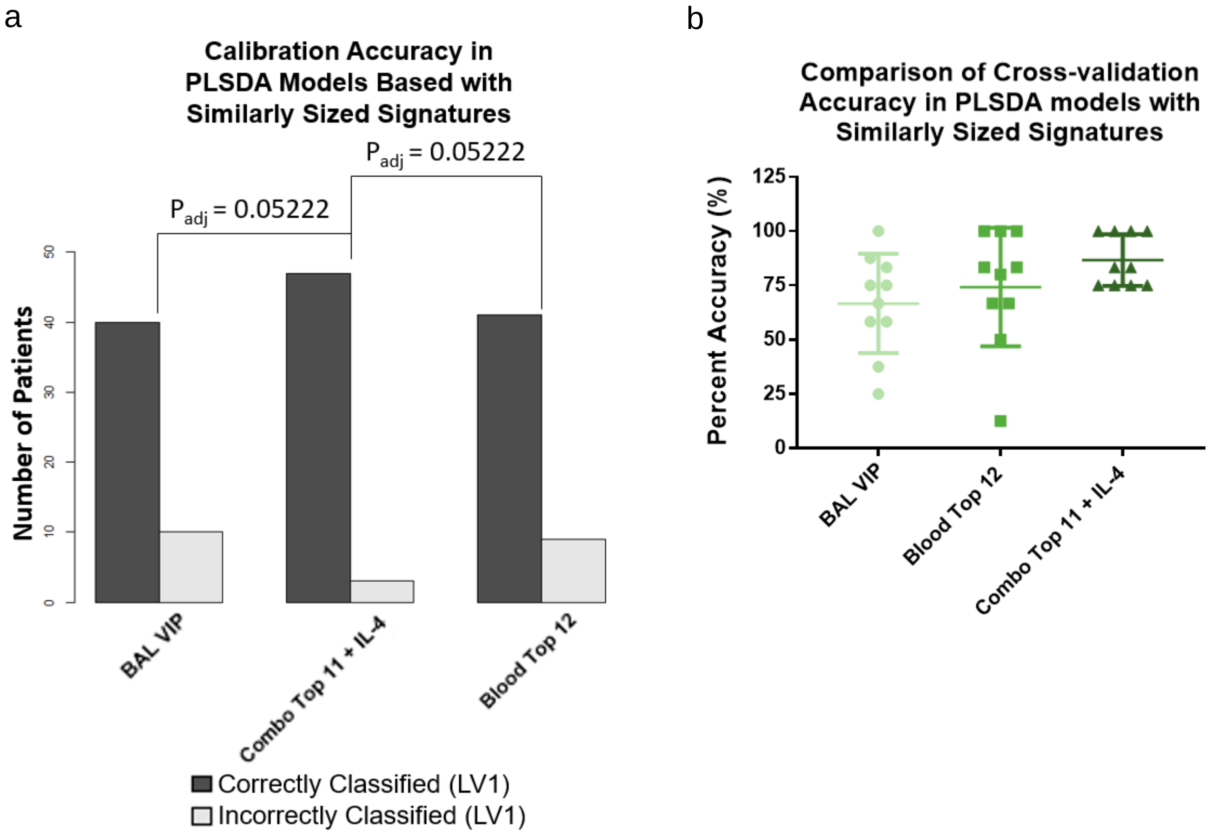


Figure B.S9. Statistical comparison of calibration and cross-validation accuracies of PLSDA models with similar number of features included in each signature showed only trends towards being significantly different from each other.

(**a**) Statistical analysis of the calibration accuracies via Cochran's Q test showed that the shortened signature based on blood and BAL proteins combined approached being significantly better than the BAL VIP and the shortened blood signature (p = 0.052, McNemar's post hoc test). (**b**) When comparing cross-validation accuracies of the three models, none were significantly different from each other.

a

neuropilin 1(NRP1)	kinase insert domain receptor(KDR)	fibroblast growth factor 1(FGF1)	TGFBR2	IL4	complement C5	colony stimulating factor 2(CSF2)	ciliary neurotrophic factor(CNTF)	CCL23	IL10	growth differentiation factor 9(GDF9)	caspase 2(CASP2)	lymphotoxin alpha(LTA)	complement C6	spectrin alpha, non-erythrocytic 1(SPTAN1)	legumain(LGMN)	IL17RB	CTLA4	peptidoglycan recognition protein 1(PGLYRP1)	cathepsin S(CTSS)	selectin L	inosine monophosphate dehydrogenase 1(IMPDH1)	ficolin 2(FCN2)	IL17D	cadherin 6(CDH6)	Pathway	Bonferroni Corrected P-value
■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	□	□	■	□	□	positive regulation of response to stimulus	4.68E-08
■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	□	■	□	■	□	□	□	■	□	□	response to external stimulus	8.52E-05
■	■	■	■	■	■	■	■	■	□	■	■	■	□	□	■	■	■	□	□	□	□	□	■	■	cell surface receptor signaling pathway	0.002974
■	■	■	■	■	■	■	■	■	□	■	■	□	■	□	□	□	□	□	□	□	□	□	□	□	positive regulation of cellular protein metabolic process	0.02651
■	■	■	■	■	■	■	■	■	□	■	■	□	■	□	□	□	□	□	□	□	□	□	□	□	positive regulation of protein metabolic process	0.046396
■	■	■	■	■	□	■	■	■	■	■	□	■	□	□	□	□	■	□	□	□	■	□	□	□	cell proliferation	0.058653
■	■	■	■	■	■	■	■	□	■	□	□	■	■	□	□	■	□	□	□	□	□	□	□	□	positive regulation of multicellular organismal process	0.028502
■	■	■	■	■	■	□	■	■	■	□	□	■	■	□	□	■	□	■	□	□	□	□	□	□	regulation of response to external stimulus	2.06E-06
■	■	■	■	■	■	■	■	■	□	■	□	□	□	□	□	□	□	□	□	□	□	□	□	□	positive regulation of protein phosphorylation	0.025269
■	■	■	■	■	■	■	■	■	□	■	□	□	□	□	□	□	□	□	□	□	□	□	□	□	positive regulation of phosphorylation	0.03636
■	■	■	■	■	■	□	■	□	■	■	□	■	■	□	□	□	■	■	□	□	□	□	□	□	regulation of multicellular organismal development	0.026881
□	■	■	■	■	■	■	■	■	■	■	□	■	□	□	□	□	□	□	□	□	□	□	■	□	receptor binding	0.003002
■	■	■	■	■	□	■	■	□	■	■	□	■	□	□	□	□	□	□	□	□	□	□	□	□	positive regulation of cell proliferation	0.017344
□	□	□	■	■	■	■	■	■	■	■	□	■	□	□	□	□	□	□	□	□	□	□	■	□	cytokine receptor binding	9.62E-08
■	■	■	□	■	■	□	□	■	■	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	cell chemotaxis	0.007705
□	■	□	■	■	□	■	■	□	■	□	□	■	□	□	□	■	□	□	□	□	□	□	□	□	tokine-cytokine receptor interaction	0.000223
□	□	■	□	■	□	■	■	□	■	■	□	□	□	□	□	□	□	□	□	□	□	□	□	□	growth factor activity	0.001478
■	□	■	■	□	□	□	□	■	□	□	□	□	□	□	□	□	□	■	□	■	□	□	□	□	glycosaminoglycan binding	0.004944

■ Indicates involvement in process

□ Indicates no involvement in process

b

IL10	lymphotoxin alpha(LTA)	IL4	complement C5	peptidoglycan recognition protein 1(PGLYRP1)	complement C6	colony stimulating factor 2(CSF2)	TGFBR2	ficolin 2(FCN2)	IL17RB	CTLA4	C-C motif chemokine ligand 23(CCL23)	cathepsin S(CTSS)	legumain(LGMN)	IL17D	Pathway	Bonferroni Corrected P-value
■	■	■	■	■	■	□	■	■	□	■	□	■	■	□	positive regulation of immune system process	0.005574
■	■	■	■	■	■	□	□	□	■	□	■	■	■	□	regulation of defense response	0.002697
■	■	■	■	■	■	□	□	□	■	□	■	□	□	■	inflammatory response	0.018036
■	■	■	■	■	■	□	□	□	■	□	■	□	□	□	regulation of inflammatory response	0.001707
■	□	■	□	□	□	■	■	□	□	□	□	□	□	□	myeloid dendritic cell activation	0.029336

■ Indicates involvement in process

□ Indicates no involvement in process

Figure B.S10. Additional DAVID enrichment analyses of the proteins in the LASSO-signature that were found to be comparatively upregulated in the non-progressors. (**a**) This cluster was mostly enriched for processes involving cell signaling and regulation of basic cell processes, with an enrichment score of 2.57. (**b**) This cluster was also enriched for processes involving the function and regulation of the immune, defense and inflammatory responses, with an enrichment score of 2.50. Black squares indicate protein involvement in a particular pathway, while white squares indicate non-involvement.

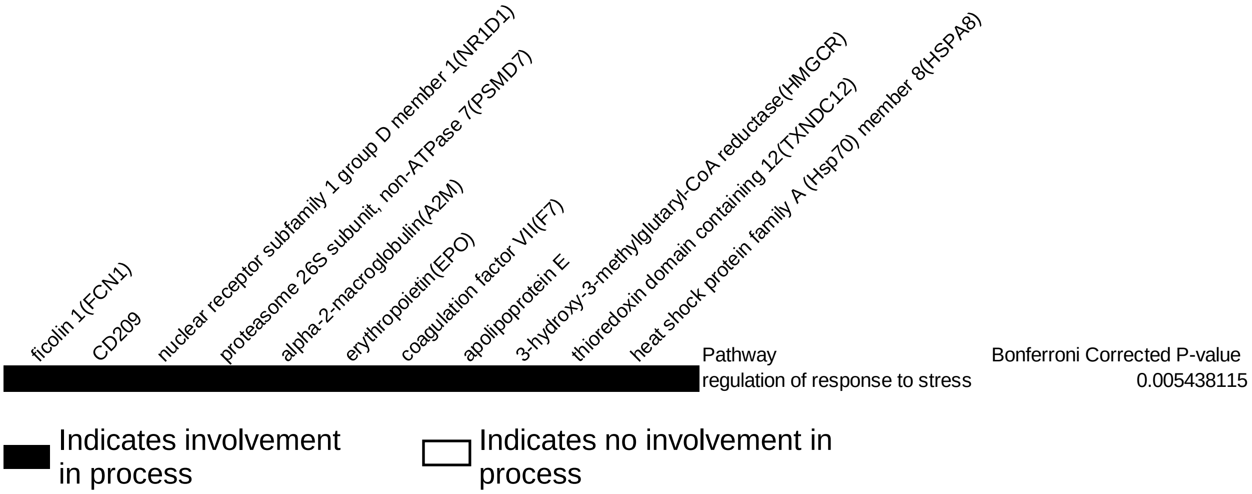
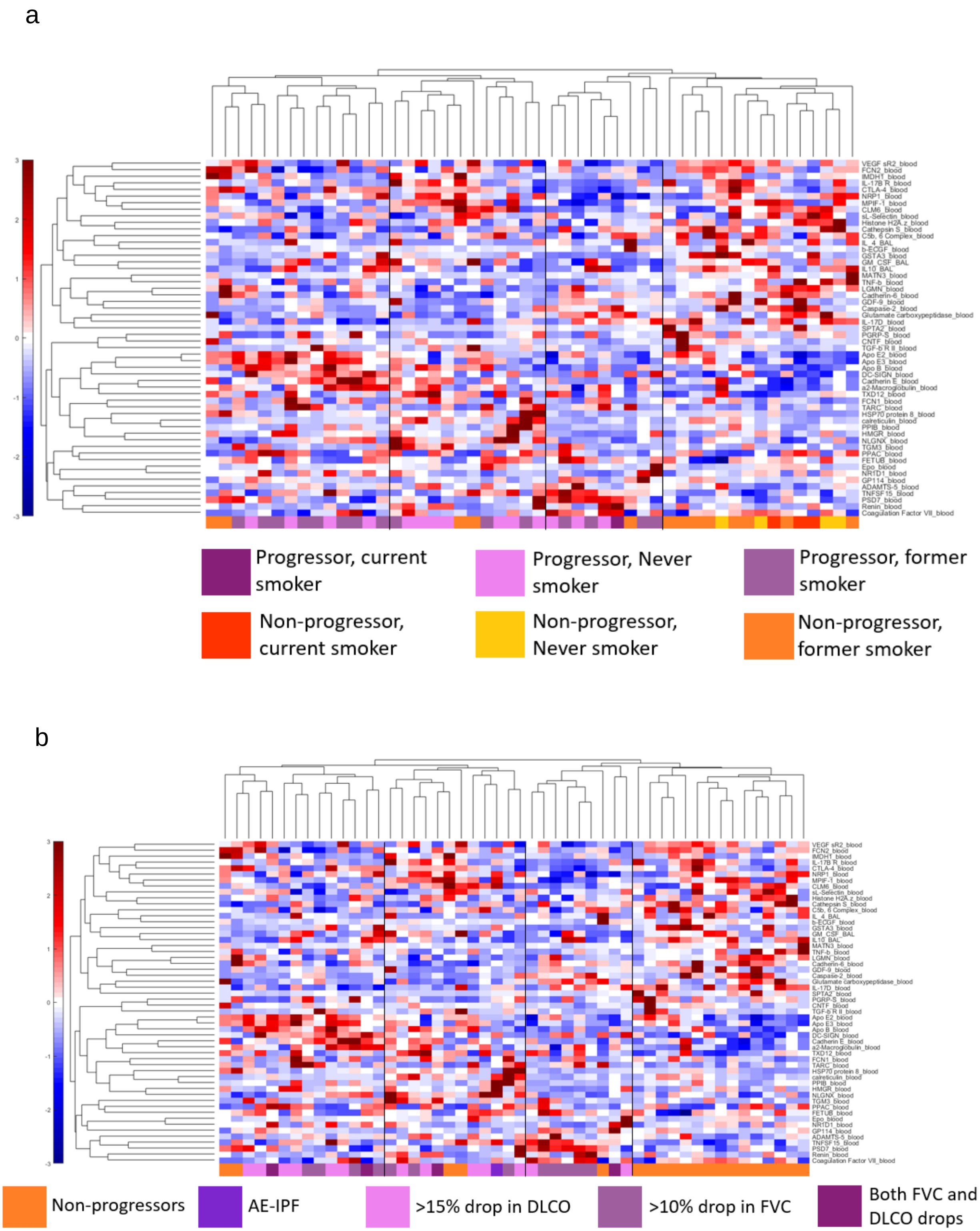
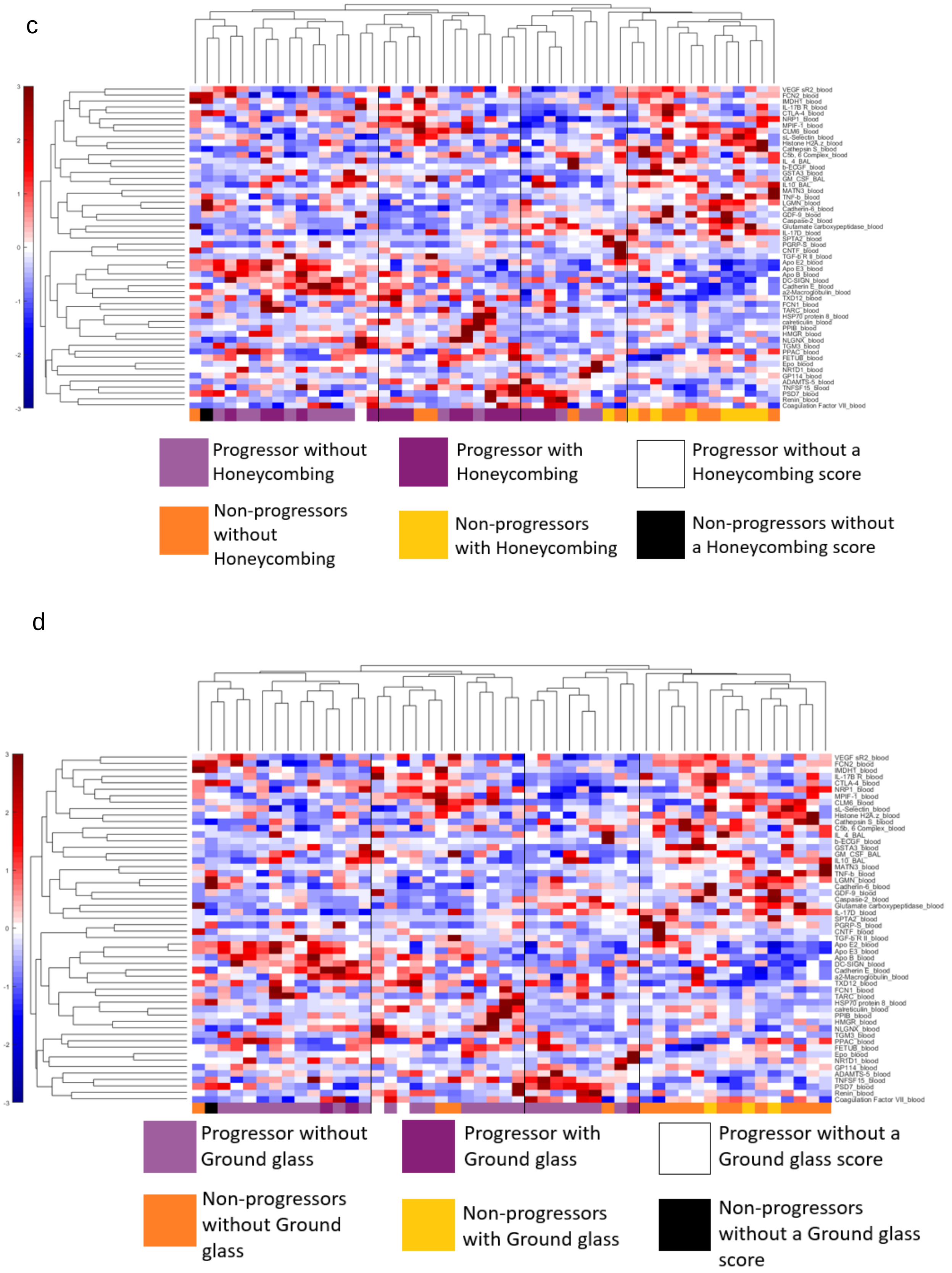
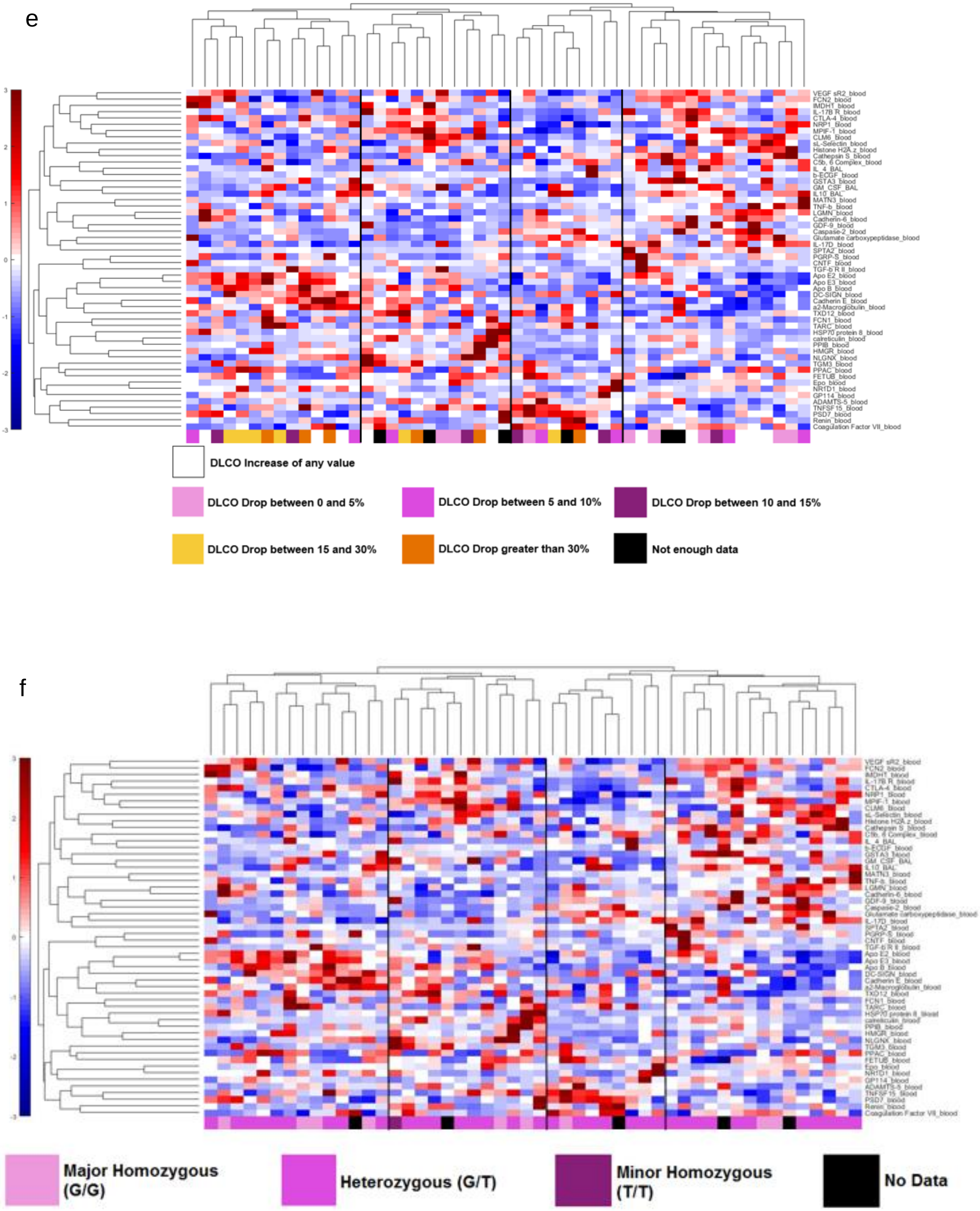


Figure B.S11. DAVID enrichment analysis of the proteins that were comparatively upregulated in the progressors in the LASSO-identified signature based on blood and BAL proteins measured in COMET IPF patients.

The enrichment score of this cluster is 2.05. Black squares indicate protein involvement in a particular pathway, while white squares indicate non-involvement.


a
3
2
1
0
-1
-2
-3
VEGF sR2_blood
FCN2_blood
IMDHT_blood
IL-17B R_blood
CTLA-4_blood
NRP1_blood
MPIF-1_blood
CLM6_blood
sL-Selectin_blood
Histone H2A.z_blood
Cathepsin S_blood
C5b, 6 Complex_blood
IL_4_BAL
b-ECGF_blood
GSTA3_blood
GM_CSF_BAL
IL10_BAL
MATN3_blood
TNF-b_blood
LGMN_blood
Cadherin-6_blood
GDF-9_blood
Caspase-2_blood
Glutamate carboxypeptidase_blood
IL-17D_blood
SPTA2_blood
PGRP-S_blood
CNTF_blood
TGF-b R II_blood
Apo E2_blood
Apo E3_blood
Apo B_blood
DC-SIGN_blood
Cadherin E_blood
a2-Macroglobulin_blood
TXD12_blood
FCN1_blood
TARC_blood
HSP70 protein 8_blood
calreticulin_blood
PPIB_blood
HMGR_blood
NLGNX_blood
TGM3_blood
PPAC_blood
FETUB_blood
Epo_blood
NR1D1_blood
GP114_blood
ADAMTS-5_blood
TNFSF15_blood
PSD7_blood
Renin_blood
Coagulation Factor VII_blood
Progressor, current smoker
Progressor, Never smoker
Progressor, former smoker
Non-progressor, current smoker
Non-progressor, Never smoker
Non-progressor, former smoker
b
3
2
1
0
-1
-2
-3
VEGF sR2_blood
FCN2_blood
IMDHT_blood
IL-17B R_blood
CTLA-4_blood
NRP1_blood
MPIF-1_blood
CLM6_blood
sL-Selectin_blood
Histone H2A.z_blood
Cathepsin S_blood
C5b, 6 Complex_blood
IL_4_BAL
b-ECGF_blood
GSTA3_blood
GM_CSF_BAL
IL10_BAL
MATN3_blood
TNF-b_blood
LGMN_blood
Cadherin-6_blood
GDF-9_blood
Caspase-2_blood
Glutamate carboxypeptidase_blood
IL-17D_blood
SPTA2_blood
PGRP-S_blood
CNTF_blood
TGF-b R II_blood
Apo E2_blood
Apo E3_blood
Apo B_blood
DC-SIGN_blood
Cadherin E_blood
a2-Macroglobulin_blood
TXD12_blood
FCN1_blood
TARC_blood
HSP70 protein 8_blood
calreticulin_blood
PPIB_blood
HMGR_blood
NLGNX_blood
TGM3_blood
PPAC_blood
FETUB_blood
Epo_blood
NR1D1_blood
GP114_blood
ADAMTS-5_blood
TNFSF15_blood
PSD7_blood
Renin_blood
Coagulation Factor VII_blood
Non-progressors
AE-IPF
>15% drop in DLCO
>10% drop in FVC
Both FVC and DLCO drops


c
3
2
1
0
-1
-2
-3
VEGF sR2_blood
FCN2_blood
IMDH1_blood
IL-17B R_blood
CTLA-4_blood
NRP1_blood
MPIF-1_blood
CLM6_blood
sL-Selectin_blood
Histone H2A.z_blood
Cathepsin S_blood
C5b, 6 Complex_blood
IL_4_BAL
b-ECGF_blood
GSTA3_blood
GM_CSF_BAL
IL10_BAL
MATN3_blood
TNF-b_blood
LGMN_blood
Cadherin-6_blood
GDF-9_blood
Caspase-2_blood
Glutamate carboxypeptidase_blood
IL-17D_blood
SPTA2_blood
PGRP-S_blood
CNTF_blood
TGF-b R II_blood
Apo E2_blood
Apo E3_blood
Apo B_blood
DC-SIGN_blood
Cadherin E_blood
a2-Macroglobulin_blood
TXD12_blood
FCN1_blood
TARC_blood
HSP70 protein 8_blood
calreticulin_blood
PPIB_blood
HMGR_blood
NLGNX_blood
TGM3_blood
PPAC_blood
FETUB_blood
Epo_blood
NR1D1_blood
GP114_blood
ADAMTS-5_blood
TNFSF15_blood
PSD7_blood
Renin_blood
Coagulation Factor VII_blood
Progressor without Honeycombing
Progressor with Honeycombing
Progressor without a Honeycombing score
Non-progressors without Honeycombing
Non-progressors with Honeycombing
Non-progressors without a Honeycombing score
d
3
2
1
0
-1
-2
-3
VEGF sR2_blood
FCN2_blood
IMDH1_blood
IL-17B R_blood
CTLA-4_blood
NRP1_blood
MPIF-1_blood
CLM6_blood
sL-Selectin_blood
Histone H2A.z_blood
Cathepsin S_blood
C5b, 6 Complex_blood
IL_4_BAL
b-ECGF_blood
GSTA3_blood
GM_CSF_BAL
IL10_BAL
MATN3_blood
TNF-b_blood
LGMN_blood
Cadherin-6_blood
GDF-9_blood
Caspase-2_blood
Glutamate carboxypeptidase_blood
IL-17D_blood
SPTA2_blood
PGRP-S_blood
CNTF_blood
TGF-b R II_blood
Apo E2_blood
Apo E3_blood
Apo B_blood
DC-SIGN_blood
Cadherin E_blood
a2-Macroglobulin_blood
TXD12_blood
FCN1_blood
TARC_blood
HSP70 protein 8_blood
calreticulin_blood
PPIB_blood
HMGR_blood
NLGNX_blood
TGM3_blood
PPAC_blood
FETUB_blood
Epo_blood
NR1D1_blood
GP114_blood
ADAMTS-5_blood
TNFSF15_blood
PSD7_blood
Renin_blood
Coagulation Factor VII_blood
Progressor without Ground glass
Progressor with Ground glass
Progressor without a Ground glass score
Non-progressors without Ground glass
Non-progressors with Ground glass
Non-progressors without a Ground glass score


e
3
2
1
0
-1
-2
-3
VEGF sR2_blood
FCN2_blood
IMDHT_blood
IL-17B R_blood
CTLA-4_blood
NRP1_blood
MPIF-1_blood
CLM6_blood
sL-Selectin_blood
Histone H2A.z_blood
Cathepsin S_blood
C5b, 6 Complex_blood
IL_4_BAL
b-ECGF_blood
GSTA3_blood
GM_CSF_BAL
IL10_BAL
MATN3_blood
TNF-b_blood
LGMN_blood
Cadherin-6_blood
GDF-9_blood
Caspase-2_blood
Glutamate carboxypeptidase_blood
IL-17D_blood
SPTA2_blood
PGRP-S_blood
CNTF_blood
TGF-b R II_blood
Apo E2_blood
Apo E3_blood
Apo B_blood
DC-SIGN_blood
Cadherin E_blood
a2-Macroglobulin_blood
TXD12_blood
FCN1_blood
TARC_blood
HSP70 protein 8_blood
calreticulin_blood
PPIB_blood
HMGR_blood
NLGNX_blood
TGM3_blood
PPAC_blood
FETUB_blood
Epo_blood
NR1D1_blood
GP114_blood
ADAMTS-5_blood
TNFSF15_blood
PSD7_blood
Renin_blood
Coagulation Factor VII_blood
DLCO Increase of any value
DLCO Drop between 0 and 5%
DLCO Drop between 5 and 10%
DLCO Drop between 10 and 15%
DLCO Drop between 15 and 30%
DLCO Drop greater than 30%
Not enough data
f
3
2
1
0
-1
-2
-3
VEGF sR2_blood
FCN2_blood
IMDHT_blood
IL-17B R_blood
CTLA-4_blood
NRP1_blood
MPIF-1_blood
CLM6_blood
sL-Selectin_blood
Histone H2A.z_blood
Cathepsin S_blood
C5b, 6 Complex_blood
IL_4_BAL
b-ECGF_blood
GSTA3_blood
GM_CSF_BAL
IL10_BAL
MATN3_blood
TNF-b_blood
LGMN_blood
Cadherin-6_blood
GDF-9_blood
Caspase-2_blood
Glutamate carboxypeptidase_blood
IL-17D_blood
SPTA2_blood
PGRP-S_blood
CNTF_blood
TGF-b R II_blood
Apo E2_blood
Apo E3_blood
Apo B_blood
DC-SIGN_blood
Cadherin E_blood
a2-Macroglobulin_blood
TXD12_blood
FCN1_blood
TARC_blood
HSP70 protein 8_blood
calreticulin_blood
PPIB_blood
HMGR_blood
NLGNX_blood
TGM3_blood
PPAC_blood
FETUB_blood
Epo_blood
NR1D1_blood
GP114_blood
ADAMTS-5_blood
TNFSF15_blood
PSD7_blood
Renin_blood
Coagulation Factor VII_blood
Major Homozygous (G/G)
Heterozygous (G/T)
Minor Homozygous (T/T)
No Data

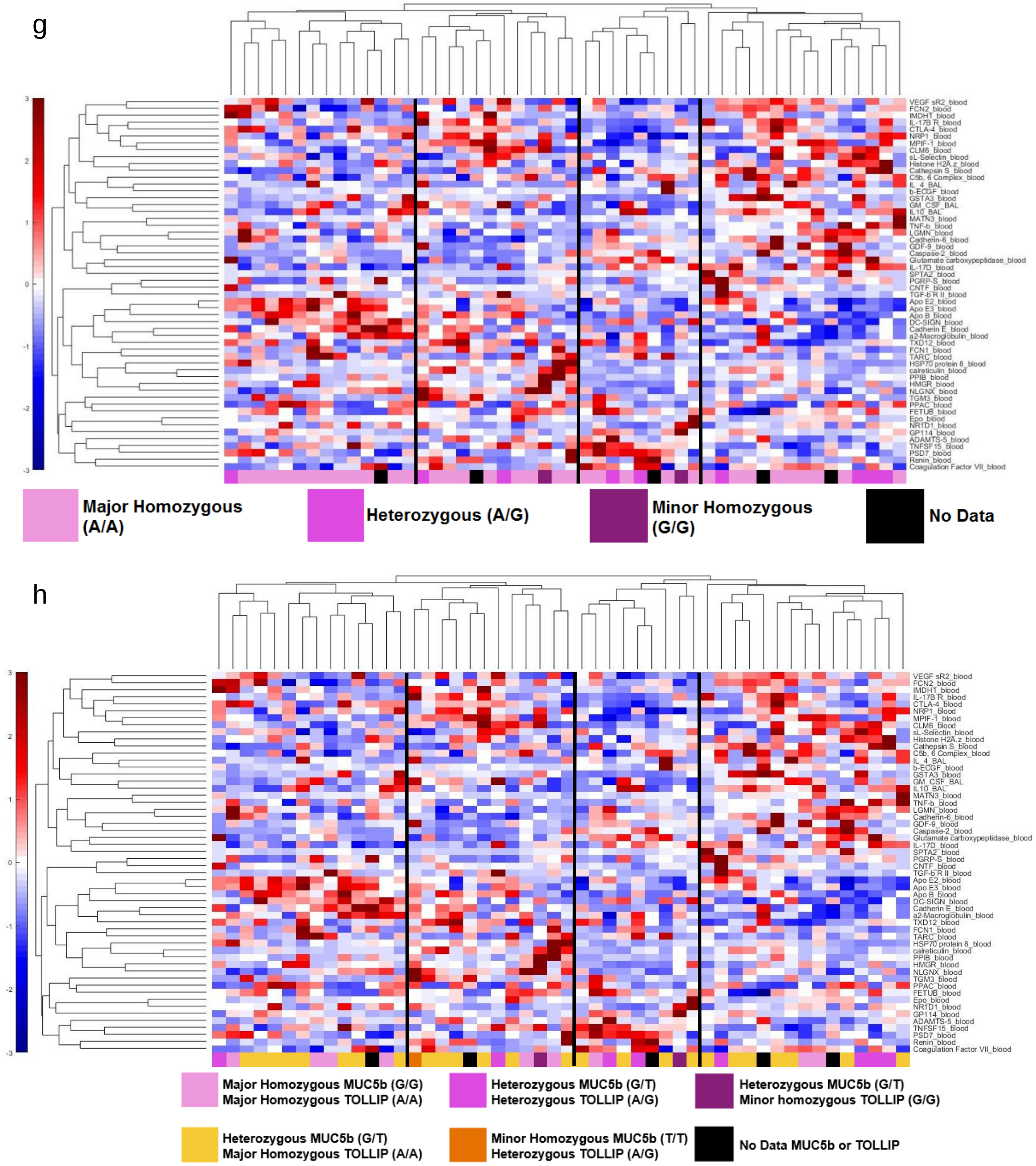


Figure B.S12. Hierarchical cluster based on the combination signature did not cluster according to the following clinical and pulmonary variables: A. smoking status, B. how progression occurred in that specific patient, C. presence of honeycombing in the CT scan, D. presence of ground glass in the CT scan, E. DLCO increase or decrease over the 80-week time period of the COMET study, F. MUC5b genotyping results, G. TOLLIP genotyping results, and H. MUC5b and TOLLIP genotyping results together.

Color bars are shown to the left of each figure, with red indicating higher protein expression level from the mean, white unchanged, and blue a lower expression. AE-IPF: acute exacerbations of IPF, DLCO: diffusing capacity of the lungs for carbon monoxide, FVC: forced vital capacity.

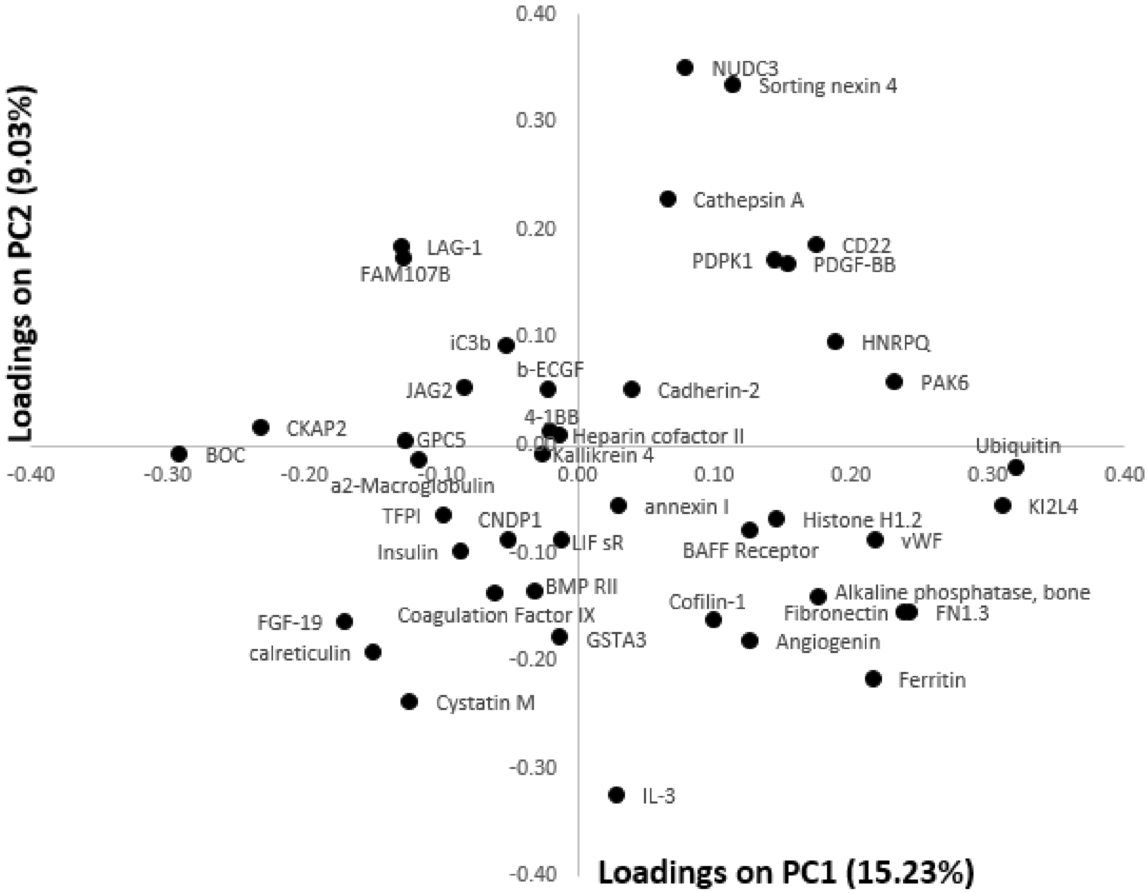


Figure B.S13. The LASSO-identified trajectory PCA signature chosen to separate the non-progressors across the three time points captured 24.26% of the natural variance in the data across the first two principal components.

APPENDIX C. Supplement to: Inference of Cellular Immune Environments in Sputum and Peripheral Blood Associated With Acute Exacerbations of COPD

Katy C. Norman[1 *], Christine M. Freeman[2, 3, 4 *], Neha S. Bidthanapally[1], MeiLan K. Han[2], Fernando J. Martinez[5], Jeffrey L. Curtis[2, 3, 6 #], and Kelly B. Arnold[1 #]

*co-first authors, #co-corresponding authors

1 Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI 48109

2 Department of Internal Medicine, Division of Pulmonary & Critical Care, University of Michigan, Ann Arbor, MI 48109

3 Research Service, VA Ann Arbor Healthcare System, Ann Arbor, MI 48105

4 Graduate Program in Immunology, Rackham Graduate School, University of Michigan, Ann Arbor, MI 48109

5 Joan & Sanford I. Weill Department of Medicine, Division of Pulmonary & Critical Care Medicine, Weill Cornell Medical College, New York, NY 10065

6 Medicine Service, Pulmonary & Critical Care Section, VA Ann Arbor Healthcare System, Ann Arbor, MI 48105

Cellular and Molecular Bioengineering. 2019; 12(2): 165-177.

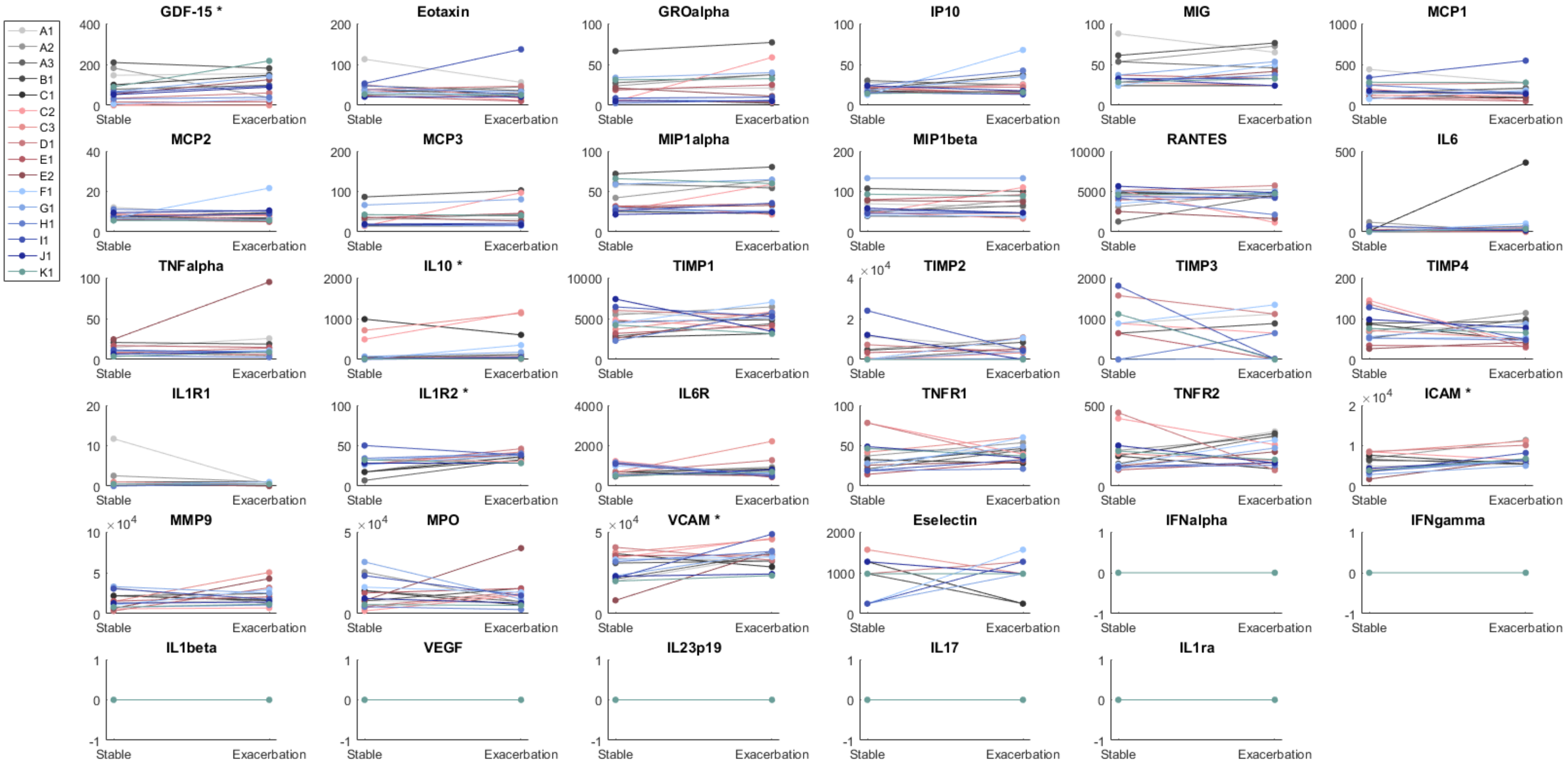


Figure C.S1. Changes in serum protein expression level in all patients used in the volcano plot analysis across the stable and exacerbated state.
* indicates significance ($p < 0.05$) after application of a paired Wilcoxon signed rank test. Patient number is indicated in the legend, with number references for multiple exacerbation visits. Unconnected dots represent measurements that were only made in one state (and were excluded from statistical analysis but included in the fold change calculation). Proteins that were not present in any individuals in any state overlap and appear as a single line. These measurements were not included in the PLSDA analysis.

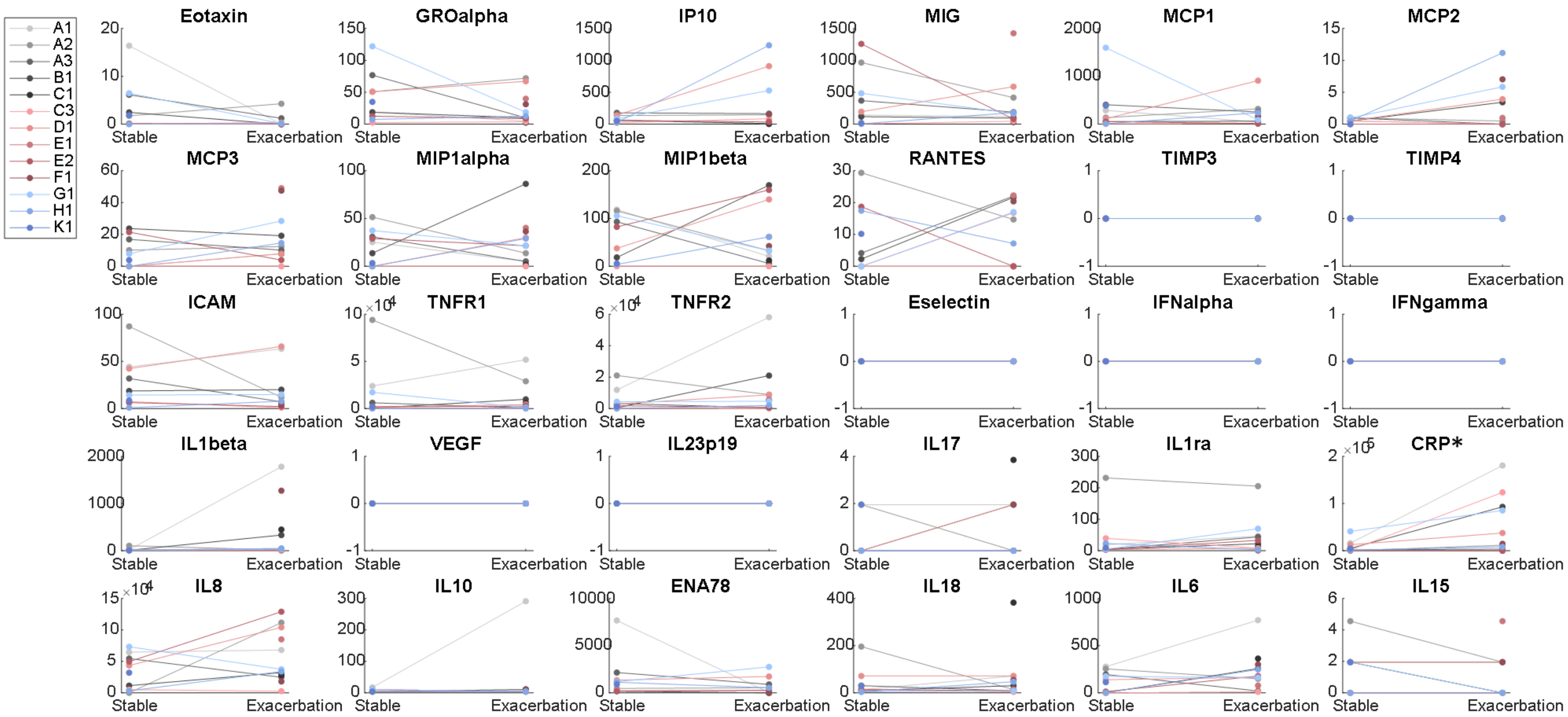


Figure C.S2. Changes in sputum protein expression level in all patients used in the volcano plot analysis across the stable and exacerbation state.
* indicates significance ($p < 0.05$) after application of a paired Wilcoxon signed rank test. Patient number is indicated in the legend, with number references for multiple exacerbation visits. Unconnected dots represent measurements that were only made in one state (and were excluded from statistical analysis but included in the fold change calculation). Proteins that were not present in any individuals in any state overlap and appear as a single line. These measurements were not included in the PLSDA analysis. the PLSDA analysis.

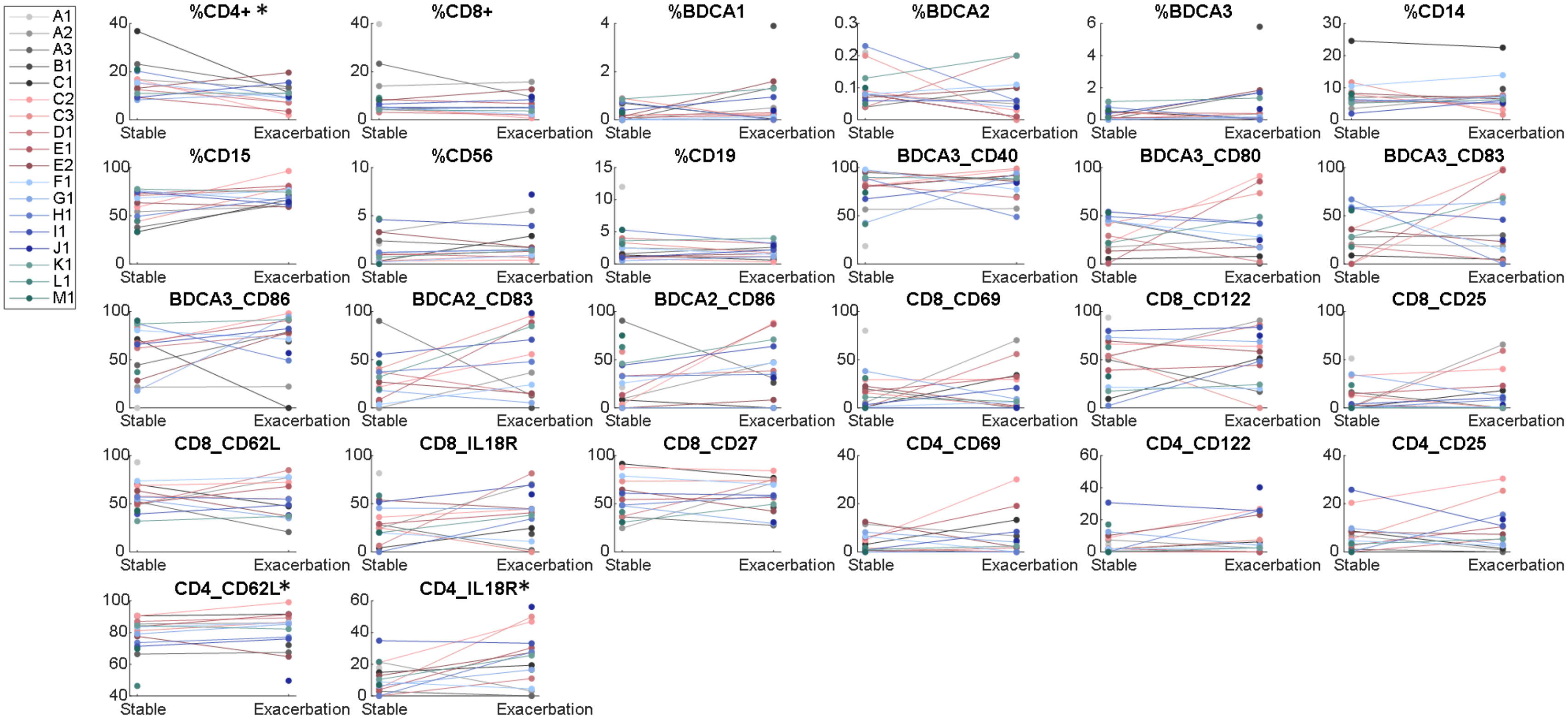


Figure C.S3. Changes in blood cell marker protein expression level in all patients used in the volcano plot analysis across the stable and exacerbation state. * indicates significance ($p < 0.05$) after application of a paired Wilcoxon signed rank test. Patient number is indicated in the legend, with number references for multiple exacerbation visits. Unconnected dots represent measurements that were only made in one state (and were excluded from statistical analysis but included in the fold change calculation). Proteins that were not present in any individuals in any state overlap and appear as a single line. These measurements were not included in the PLSDA analysis.

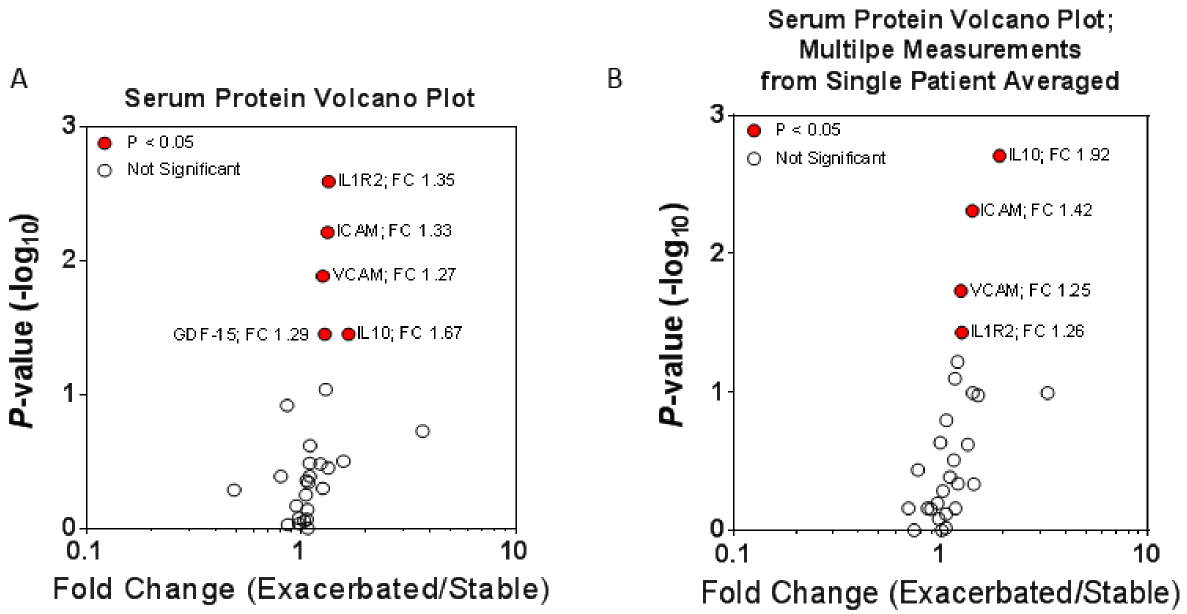


Figure C.S4. Volcano plot based on all stable and exacerbation serum protein measurements (A) compared to volcano plot created after first averaging the serum protein concentrations from all of the stable points and all of the exacerbation measurements collected across multiple visits from patient A, from patient C, and from patient E separately (B) before calculating fold change and performing the Wilcoxon signed rank test.

The volcano plots illustrate serum proteins that are both differentially expressed (x axis) and significantly different (y axis) between the stable and exacerbated state. Points in red indicate significantly different expression between the stable and exacerbated state via paired Wilcoxon signed rank test, with significance being defined as p < 0.05.

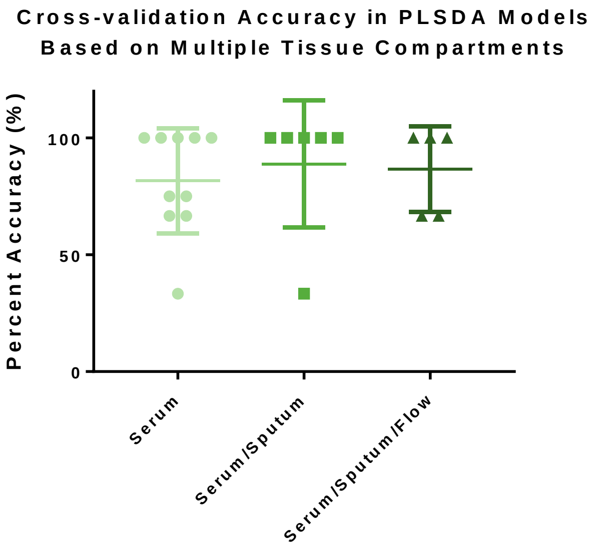


Figure C.S5. Comparison of the cross-validation accuracies associated with each training and test set created during cross-validation of the PLSDA models based on VIP-identified serum proteins, serum and sputum proteins, and serum and sputum proteins and flow markers.

No model had significantly higher cross-validation accuracy according to Tukey's multiple comparison test (one-way ANOVA).

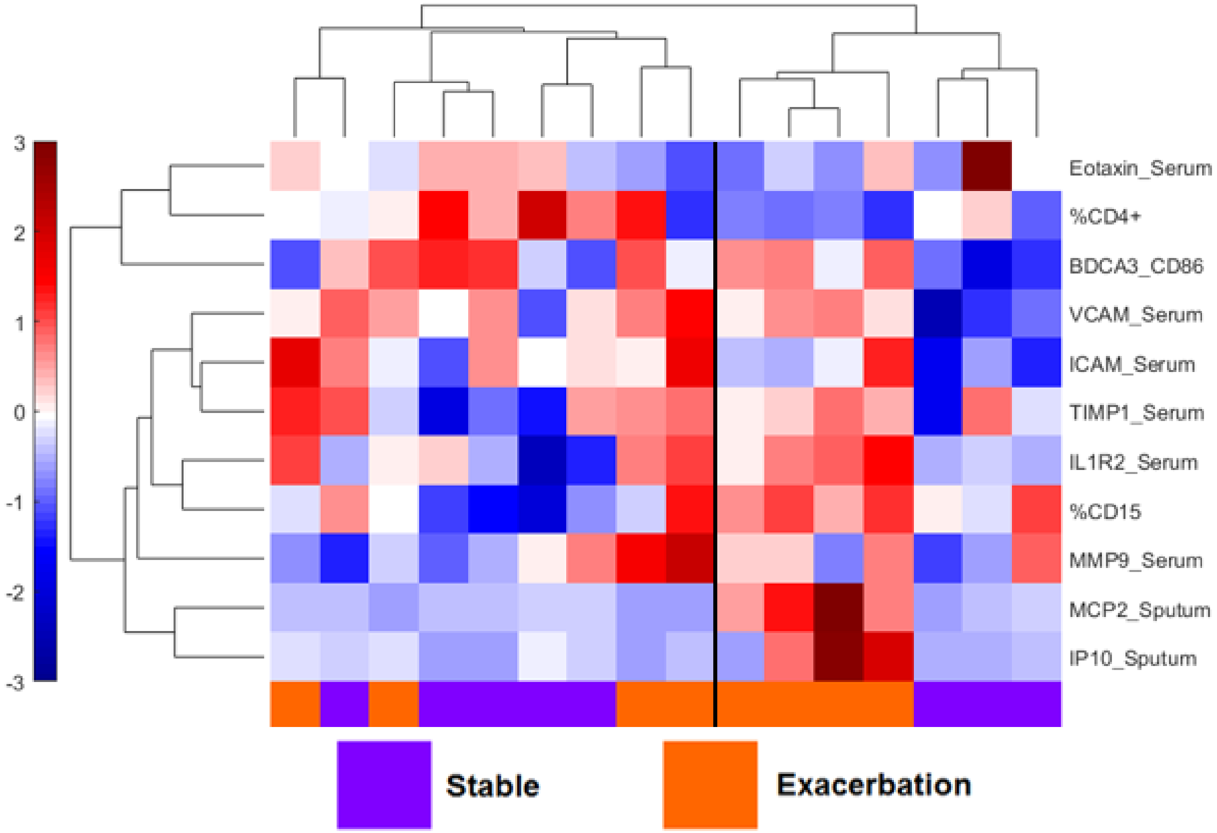


Figure C.S6. Hierarchical clustering of the patient samples included in the PLSDA model is ultimately unable to accurately classify stable from exacerbation, with seven patients out of 16 included being misclassified.
Protein abundance is shown on a colorimetric scale, with red indicating overabundant, white unchanged, and blue under abundant protein level compared to the mean. Color bar scale is to the left of the figure.

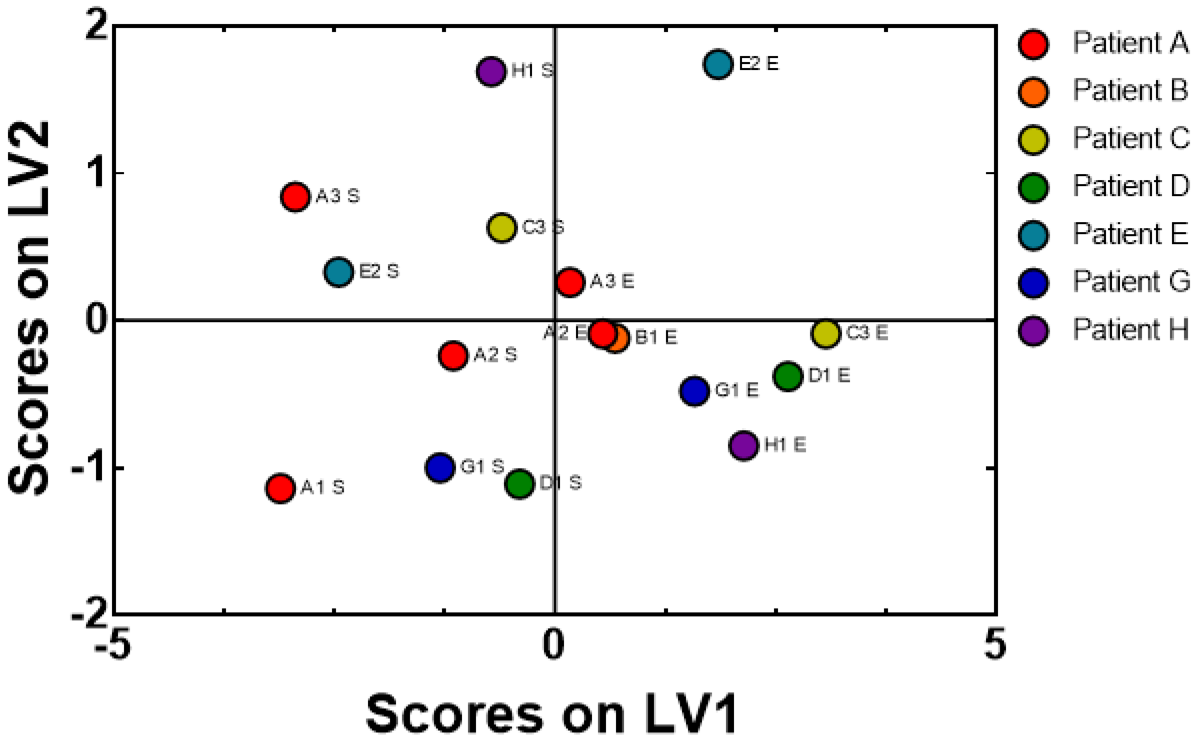


Figure C.S7. An investigation of the effect of including multiple paired stable and exacerbation measurements from the same patient in a PLSDA model based on serum and sputum proteins and blood flow marker data showed that there is no clustering by patient, only by stable and exacerbated states.

This PLSDA scores plot is the same as the one as shown in **Figure 4A**, except the samples are now colored by which patient they came from. Each point is additionally labeled to convey information about the state of the patient for that sample (i.e. stable or exacerbated), as well as with information about which visit the point is referring to, in the cases where multiple exacerbations were captured for one patient.

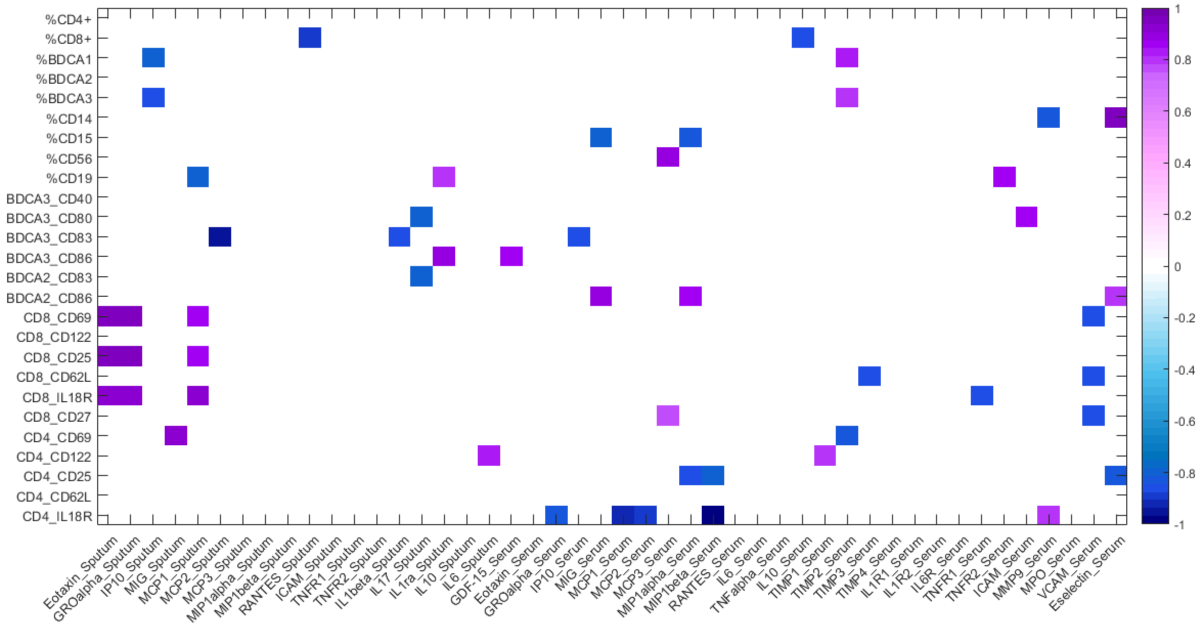


Figure C.S8. A correlation coefficient heat map based on the change in concentration between the stable and exacerbated state of all measured blood cell markers and serum and sputum proteins highlights how cellular concentrations could potentially affect protein concentration during exacerbation.

Correlation coefficients were calculated using Spearman’s rank correlation. Color bar scale is shown to the right of the figure.

APPENDIX D. Appendix to SPIROMICS Results

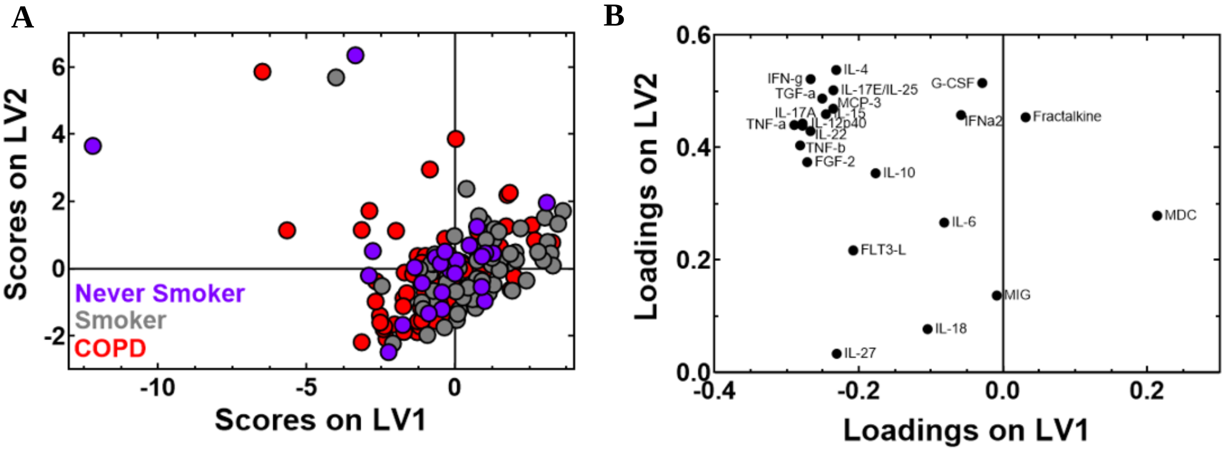


Figure D.1. Signature of VIP-selected, Luminex-measured plasma cytokines is unable to differentiate COPD disease state in a PLSDA model.

(**A**) There is no differentiation between smokers (grey), never smokers (purple), and COPD subjects (red) in the PLSDA scores plot. This model had a 65.69% calibration and 57.88% cross-validation accuracy. (**B**) Latent variable 1 (LV1) captured 28.97% of the variance in the data.

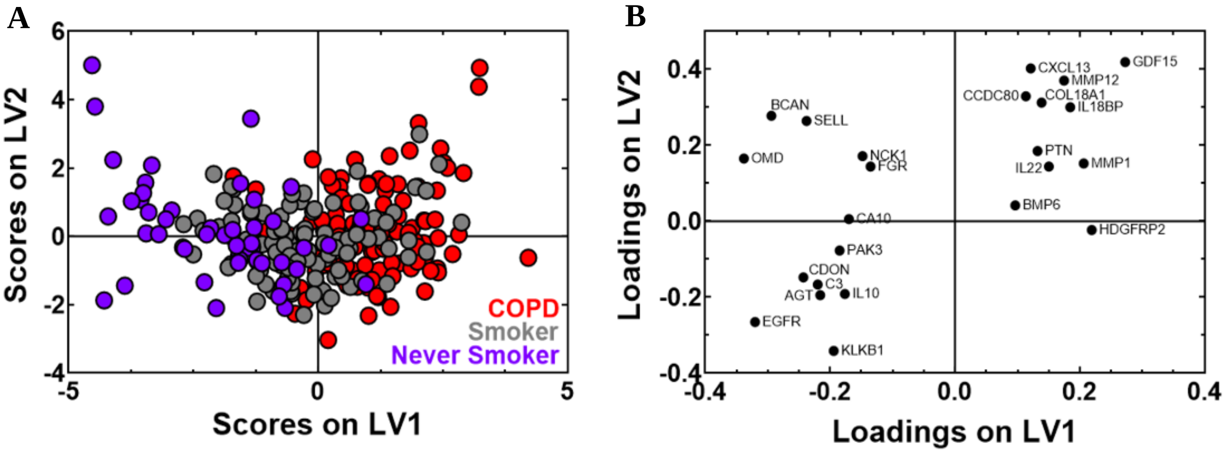


Figure D.2. Signature of SOMAmer-measured blood proteins is able to differentiate COPD disease state moderately well, but has slightly lower accuracy than model based on BAL proteins.

(**A**) PLSDA scores plot based on SOMAmer-measured blood proteins moderately separates the three groups with 74.56% calibration and 67.59% cross-validation accuracy. (**B**) PLSDA loadings plot captured 12.62% of the variance on LV1. Proteins that are positively loaded on LV1 are comparatively increased in most of the COPD subjects and around half of the smokers.

A

ret proto-oncogene(RET)	FGF10	p21 (RAC1) activated kinase 3(PAK3)	cell adhesion associated, oncogene regulated(CDON)	serpin family F member 2(SERPINF2)	LDL receptor related protein 8(LRP8)	IL15	macrophage migration inhibitory factor (MIF)	contactin 1(CNTN1)	BH3 interacting domain death agonist(BID)	lymphotoxin alpha(LTA)	IL19	methionyl aminopeptidase 1(METAP1)	caspase 10	kallikrein B1(KLKB1)	complement C2	Pathway	Bonferroni corrected P-value
■	■	■	■	■	■	■	■	■	■	□	□	■	□	■	■	regulation of protein metabolic process	0.003462898
■	■	■	■	■	■	■	■	■	■	□	□	■	□	■	□	regulation of cellular protein metabolic process	0.015297948
■	■	■	■	■	■	■	■	■	■	□	□	□	□	■	□	positive regulation of cellular protein metabolic process	0.001275865
■	■	■	■	■	■	■	■	■	■	□	□	□	□	■	□	positive regulation of protein metabolic process	0.002224591
■	■	■	■	■	■	■	■	■	□	■	□	□	□	□	□	positive regulation of multicellular organismal process	0.014824832
■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	positive regulation of protein phosphorylation	0.004124438
■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	positive regulation of phosphate metabolic process	0.013929143
■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	positive regulation of phosphorylation	0.005816171
■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	positive regulation of phosphorus metabolic process	0.013929143
■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	positive regulation of protein modification process	0.02600925
■	■	■	■	■	□	■	■	□	■	■	■	□	■	■	■	positive regulation of response to stimulus	0.000330764
■	■	■	■	■	□	■	■	□	■	□	■	□	■	□	□	positive regulation of cell communication	0.02898437
■	■	■	■	■	□	■	■	□	■	□	■	□	■	□	□	positive regulation of signaling	0.030197501
■	■	■	■	■	□	■	■	□	■	□	■	□	■	□	□	positive regulation of signal transduction	0.014573905
■	□	■	■	■	■	■	□	■	□	■	□	□	□	□	□	positive regulation of cell differentiation	0.034061891
□	■	■	■	■	□	■	■	□	■	□	■	□	■	□	□	positive regulation of intracellular signal transduction	0.005071393
□	■	□	□	□	■	■	■	■	□	□	□	□	□	□	□	positive regulation of peptidyl-tyrosine phosphorylation	0.04397408

B

macrophage migration inhibitory factor (MIF)	FGF10)	IL15	lymphotoxin alpha	LDL receptor related protein 8(LRP8)	serpin family F member 2(SERPINF2)	complement C2	kallikrein B1	contactin 1	methionyl aminopeptidase 1(METAP1)	Pathway	Bonferroni corrected P-value
■	■	■	■	□	■	■	■	□	■	Regulation of response to external stimulus	0.012623595
■	■	■	■	□	■	□	■	□	□	Positive regulation of response to external stimulus	0.014078805
■	■	■	□	■	□	□	□	■	□	Positive regulation of peptidyl-tyrosine phosphorylation	0.04397408

■ Indicates involvement in process □ Indicates no involvement in process

Figure D.3. DAVID identified a cluster of significant pathways (Bonferroni corrected p < 0.05) involving (A) metabolic process regulation (enrichment score 2.27) and (B) regulation of stimulus response (ES 1.92) that were enriched in the proteins that were comparatively increased in the never smokers and smokers in the cross-tissue compartment LASSO-identified signature.

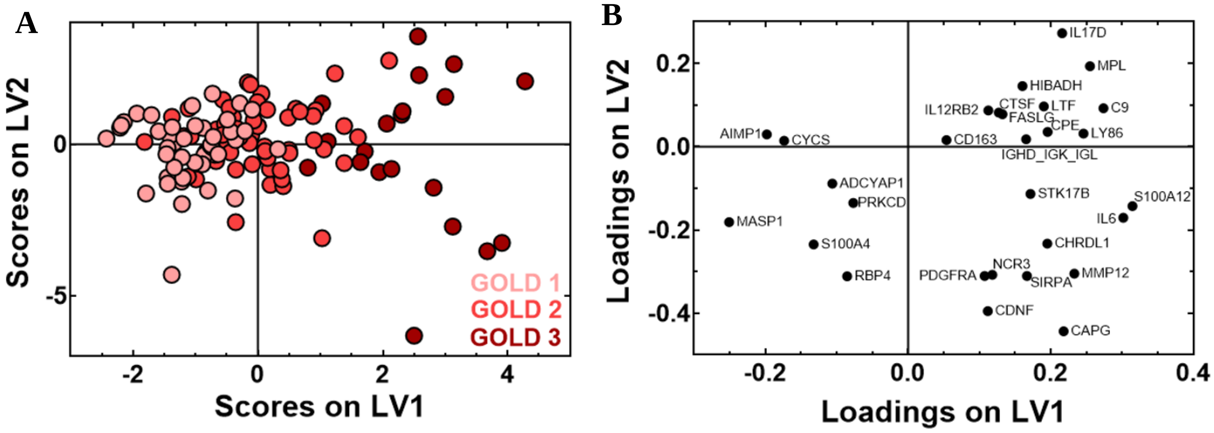


Figure D.4. LASSO-identified signature of 29 SOMAmer-measured blood proteins differentiates COPD severity moderately well.
(**A**) A signature of 29 SOMAmer-measured blood proteins differentiated three GOLD stages of COPD severity with 84.74% calibration and 77.77% cross-validation accuracy. (**B**) LV1 captured 7.938% of the variance in the data. COPD subjects with a higher GOLD classification had more positive scores on LV1.

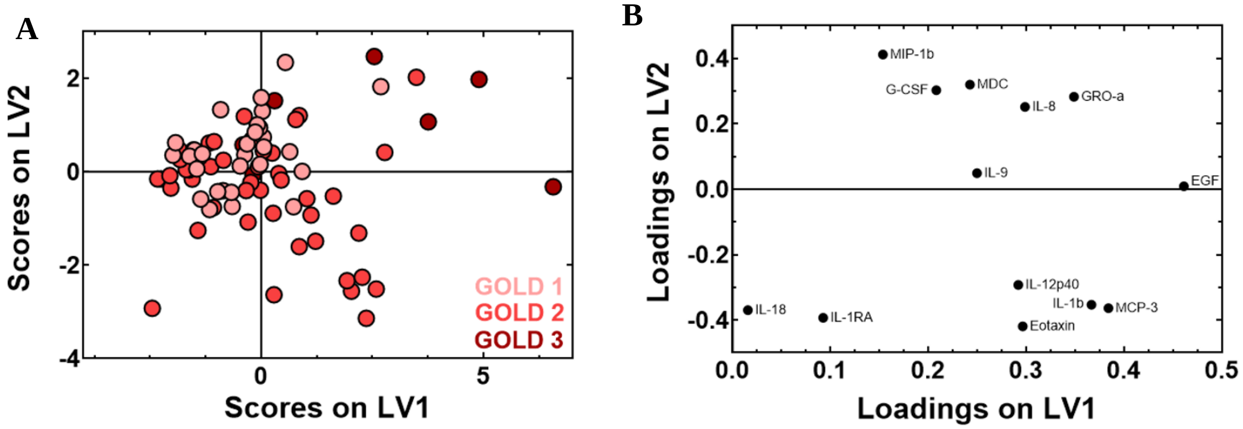


Figure D.5. Feature selected signature of 13 Luminex-measured BAL cytokines is not a strong differentiator of COPD disease severity.
(**A**) A VIP-selected signature of 13 BAL cytokines differentiated three GOLD stages of COPD severity with 69.53% calibration and 62.09% cross-validation accuracy. (**B**) LV1 captured 25.3% and LV2 captured 11.61% of the variance in the data. GOLD 3 subjects tended to have positive scores on LV1 and LV2. LV2 also separated GOLD 1 and GOLD 2 subjects.

APPENDIX E. Brief Results of Other Non-Published Experiments

Introduction

This appendix briefly reviews modeling approaches and situations that were explored during the course of this thesis, but were ultimately not followed up on, be it for having low model accuracy or for not contributing much biological insight to the field.

Biological differences in IPF patients requiring a diagnostic biopsy

In the COMET IPF cohort, 35 of the 60 recruited IPF patients received a biopsy to confirm diagnosis of IPF due to lack of clear usual interstitial pneumonia (UIP) patterns visible on chest computed tomography (CT) scans. We wanted to explore if there were distinct proteomic signatures in individuals who had to undergo a biopsy vs. those who were diagnosed with non-invasive mechanisms. Identification of a proteomic signature common to both IPF groups could suggest a basis for a new, less invasive diagnostic method, and potentially eliminate the need for diagnostic lung biopsies. We first set out to determine whether unsupervised approaches could identify differences between the two IPF groups and healthy patients. When we performed hierarchical clustering to investigate the differences between the healthy patients and the IPF patients who did or did not get biopsies (**Figure E.1A**), most of the healthy patients were grouped together, whereas there were no obvious differences in proteomes of the biopsy and no biopsy proteomes. We then used principal component analysis (PCA) (another unsupervised approach) to look at inherent differences between these three groups of

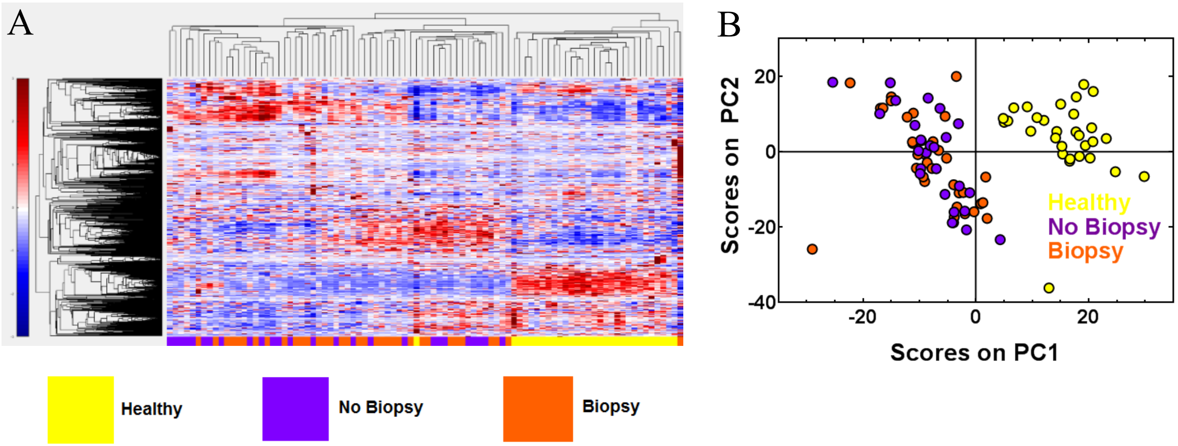


Figure E.1. Unsupervised methods of classification highlighted differences between healthy and IPF patients, but did not capture proteomic differences between IPF patients who had received a biopsy to confirm diagnosis versus those who did not.
A. Unsupervised hierarchical clustering of the three groups of patients based on 1129 proteins measured in the SOMAlogics assay did not show major differences between IPF patients who had received a biopsy vs. those who had not, but did separate healthy patients out from the IPF patients well. This indicates that the IPF patients had similar blood proteomes overall. Abundance of each protein is shown in color, with red indicating overabundant proteins, white unchanged, and blue underabundant proteins when compared to the mean expression (color bar to left of scale). **B**. PCA, another unsupervised method used to visualize inherent differences between data points, was also used to explore the differences in the blood proteome of healthy and IPF patients who did or did not receive a diagnostic biopsy. Again the healthy patients are seen to be visually distinct from the IPF patients, and there seemed to be little difference between the IPF patients who had received a biopsy and those who had not.

patients (**Figure E.1B**). Again, healthy patients clustered differently than all IPF patients, with no biopsy and biopsy groups occupying the same area in the multivariate space, suggesting little inherent differences between blood proteomes in these two groups of patients. Both PCA and hierarchical clustering suggested that the blood proteome of healthy patients was quite different from that of patients diagnosed with IPF, and additionally that the blood proteomes of IPF patients were similar, regardless of diagnostic method. This has implications in how IPF is diagnosed: if a person is exhibiting symptoms of IPF according to CT scans, then a diagnosis of IPF by biopsy might not be necessary. We next turned to the supervised LASSO technique to identify a signature that differentiated the patients who did and did not receive a biopsy; when visualized using PLSDA, this signature had a calibration and cross-validation accuracy of 77.43%. This may suggest there could be differences in the plasma proteome between IPF patients who do need a biopsy for diagnosis vs. those who don’t, but this result would need to be investigated and validated in another cohort before a more definitive conclusion could be reached.

IPF endotype investigation

Due to the heterogeneity present in the disease course of IPF, it is possible that patients may present similar losses in lung function that are actually caused by different pathophysiological mechanisms[320]. If disease subgroups or endotypes could be identified that are associated with specific biomarkers, this could help in the discovery of new diagnostic or treatment options for IPF and could lead to more personalized treatment options for patients. We attempted to identify potential endotypes within the 60 IPF patients recruited in the COMET cohort using unsupervised hierarchical clustering. Based on the dendrogram separating the IPF patients and the visual expression patterns of all 1129 proteins measured in the SOMAscan assay, we identified three clusters of IPF patients (**Figure E.2**). Although we explored proteomic

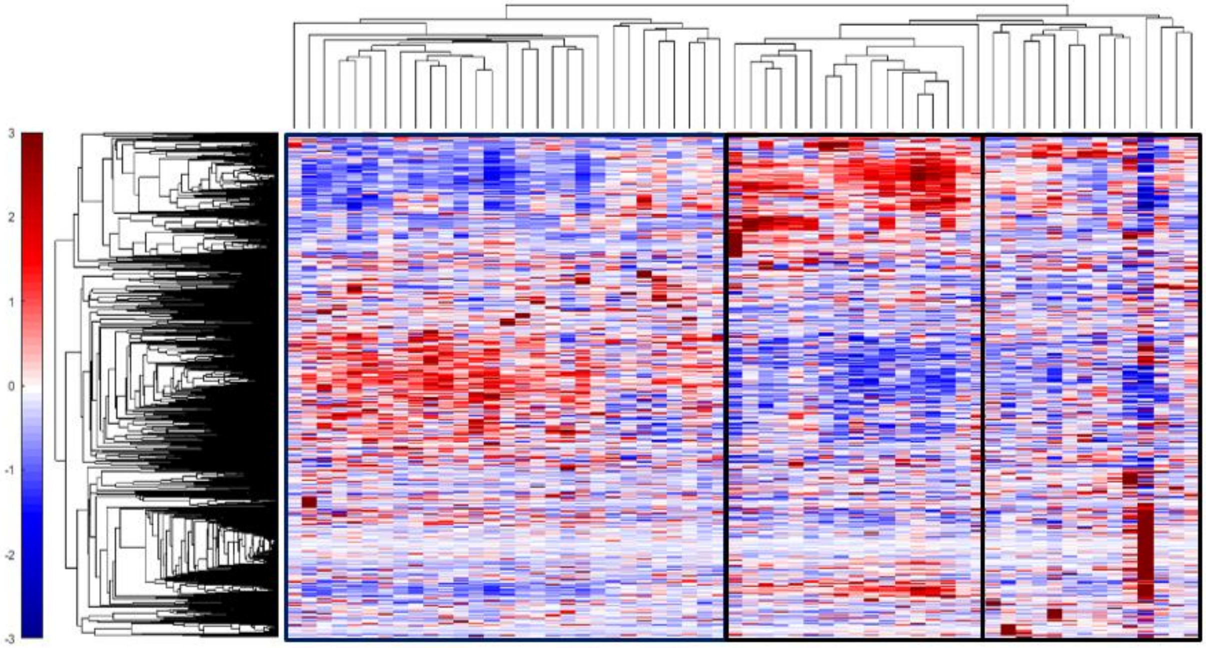


Figure E.2. Hierarchical clustering of all 60 COMET IPF patients based on expression of all 1129 proteins measured in the SOMAlogics assay identified three groups of IPF patients with distinct proteomic expression, as outlined in the black boxes.
Abundance of each protein is shown in color, with red indicating overabundant proteins, white unchanged, and blue underabundant proteins when compared to the mean expression (color bar to left of scale).

differences in these three groups using LASSO and PLSDA and clinical differences (such as differences in radiology, PFTs, and comorbidities), we did not discover novel differences, and did not follow up on these identified groups.

Validation of proteomic signature that differentiated healthy and COMET IPF patients

One difficulty with working with human data is that cohorts that employ the same assays across a similar patient population can be difficult to fund or obtain, which makes it difficult to validate a proteomic signature for a diagnostic or prognostic purpose. However, our contacts at MedImmune were able to share SOMAmer data from healthy, IPF and COPD patients not enrolled in the COMET study shared with us for validating our signatures that differentiated

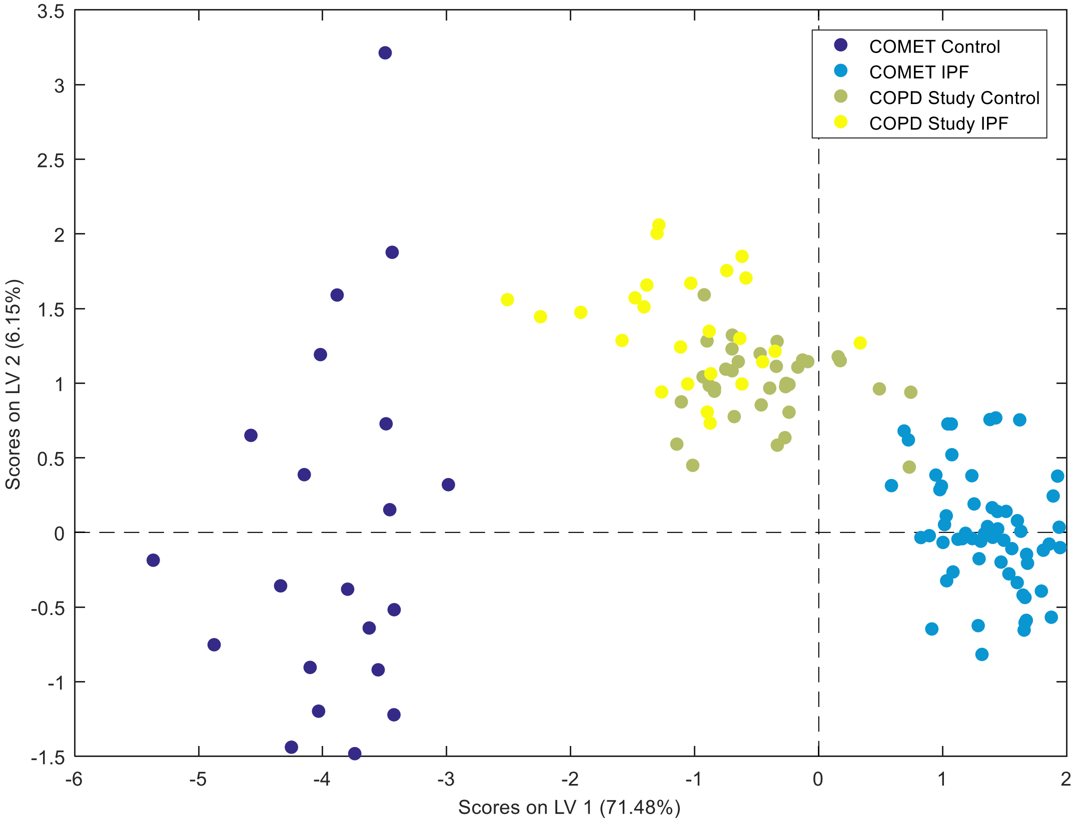


Figure E.3. Validation of the 8 protein signature that differentiated COMET control/healthy (dark blue) and IPF (lighter blue) patients[174] was not successful in another cohort (COPD Study cohort) of control (green) and IPF (yellow) patients shared with us by MedImmune, although this could have been due to unknown demographic and diagnostic guidelines.

Latent variable 1 (LV1) accounted for 71.48% of the variance in the data, and latent variable 2 (LV2) accounted for 6.15% of the variance in the data.

healthy and COMET IPF patients. However, when we tested our model using this unseen healthy and IPF data (called “COPD study” in **Figure 5.3**), we did not see clustering patterns that we expected: the new healthy and IPF patients were located in between the COMET IPF and healthy patients, with the new control patients scoring slightly more positive on LV1 and thus clustering nearer to the COMET IPF patients (**Figure E.3**). We did not receive any demographic data associated with these new patients, which made us unable to account for clinical or demographic parameters that may differ across the two cohorts. In addition, we did not know the exact guidelines that were used when diagnosing the IPF patients, which could have led to greater differences in the proteome of the IPF patients from the two cohorts. We did investigate how the model classified the COPD patients, and saw that they tended to cluster on the scores plot relatively closely to the COMET control patients (data not shown), but we did not explore this relationship any further.

Separation of IPF patients based on radiological variables

A subset of the COMET IPF patients received high resolution computed tomography (HRCT) scans as part of the IPF diagnosis process. Using the measured ALV score relating to the ground glass opacity seen in the HRCT scan and the INT score, which quantifies the fibrosis level in the HRCT scan[321], we used partial least squares regression (PLSR) to identify a signature of baseline blood proteins measured by the SOMAscan assay that could differentiate across the continuous range of these two scores (model based on ALV scores seen in **Figure E.4**; INT model not shown). We identified signatures using LASSO that accurately differentiated patients based on ALV and INT scores with high accuracy (94.2% R^2 and 83.3% Q^2 values on the ALV model). However, when we further investigated these signatures, we found that these identified proteins were novel in their association with IPF. Based on this lack of supporting evidence in the literature, we decided not to follow up on these models.

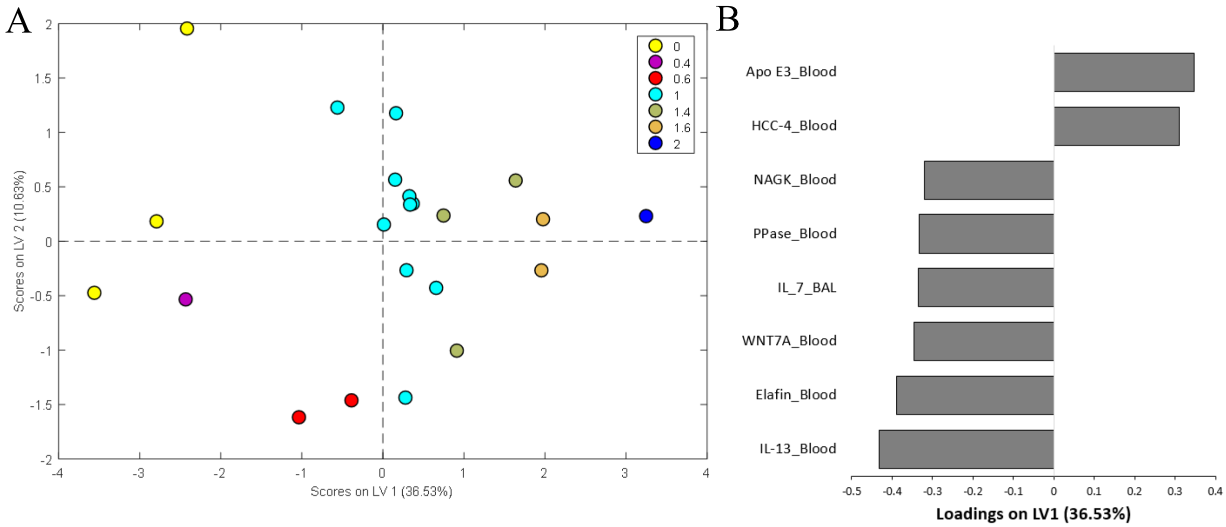


Figure E.4. LASSO/PLSR identified a signature of proteins that best differentiated patients by HRCT ALV score. **A**. LASSO identified a signature that differentiated patients across a range of continuous ALV scores, with a calibration R2 of 94.2% and a cross-validation R2 of 83.3%. **B**. The loadings plot indicates protein contributions of the LASSO-identified signature, with positive loadings positively associated with patients with higher ALV scores, and negative loadings comparatively reduced in patients with higher ALV scores.

Temporal models of IPF progression

We created a variety of models while investigating temporal differences in the blood proteome of IPF progressors and non-progressors. While we ended up publishing the trajectory PCA models of the progressors and non-progressors, we also explored PLSDA models of the blood proteome at week 48 and week 80 separately (data not shown), though we did not develop these models further because they were not as clinically useful as the signature based on week 0 protein expression. We also explored a PLSDA model where we used LASSO to identify a signature of blood proteins from both the week 0 and the week 48 time points that separated the two groups with 88.48% calibration and cross-validation accuracy. The scores plot for this model can be seen in **Figure E.5A**, with latent variable one capturing 27.86% of the variance between progressors and non-progressors, which is nearly the same amount of variance captured by the

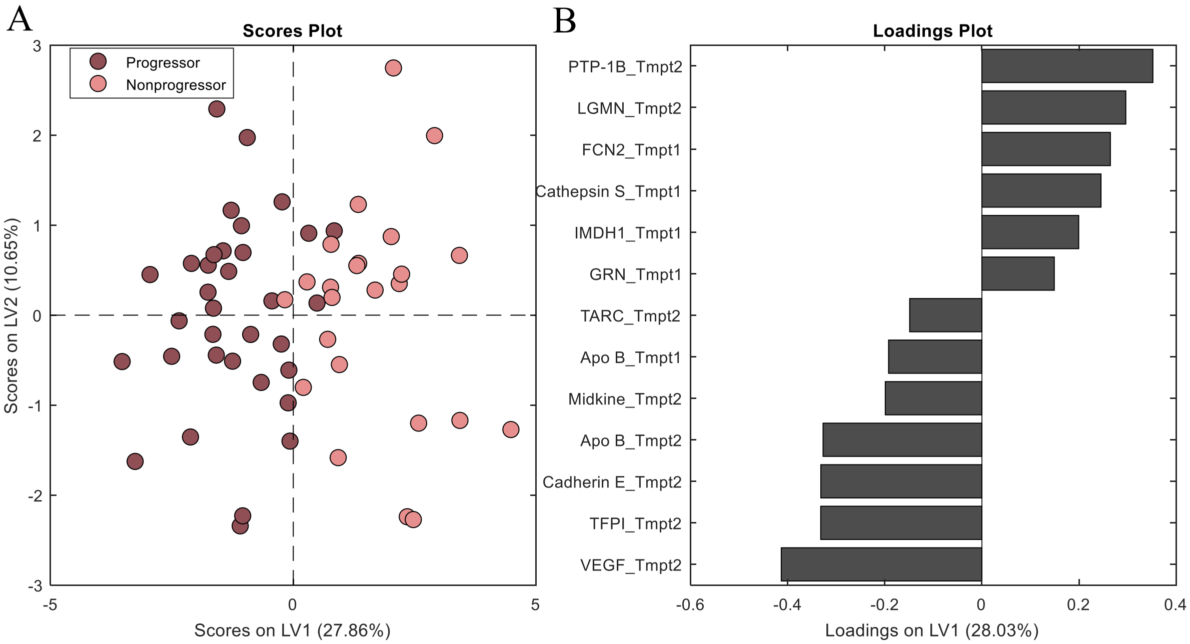


Figure E.5. A PLSDA model based on blood proteins from baseline (Tmpt 1) and 48 weeks (Tmpt 2) separates progressors and non-progressors well and contains mostly proteins from week 48 in the LASSO-identified signature. **A.** The LASSO-identified signature differentiated IPF progressors from non-progressors with 88.48% calibration and cross-validation accuracy in a PLSDA model. Latent variable 1 captured 27.86% of the variance in the data. **B.** The protein loadings associated with this model; proteins loaded negatively are comparatively upregulated in progressors, while proteins loaded positively are comparatively downregulated in progressors.

model based solely on week 48 proteins. It was interesting to note that out of the 13 proteins LASSO identified in this signature, eight of them are from week 48, and this includes the top two positively and negatively loaded proteins. This suggested to us that the proteins at later time points are more important in classifying the two groups. Proteins that were found to be comparatively upregulated in the progressors were mostly from the week 48 time point (6 out of 7). Some of the top loaded proteins in the progressors include apolipoprotein B, E-cadherin and TFPI (**Figure E.5B**), which was intriguing to us because these proteins were also chosen in the baseline (week 0) only (**Chapter 3**) and the week 48 only models as well. While we were able to hypothesize potential mechanisms involving these proteins that may be associated with progression, we did not follow up on these results because using data from the week 48 time point was not as clinically useful because progression should be attempted to be slowed or halted as soon as it could be detected.

Proteomic signatures associated with method of IPF progression

In the COMET study, patients who were classified as progressors experienced at least one of four events throughout the course of the 80-week study: (1) An acute exacerbation of IPF (AE-IPF), (2) A lung transplant, (3) A drop of 10% or more in FVC, or 4. A drop if 15% or more in DLCO measurements. Some patients even experienced a drop in both FVC and DLCO values that would classify them as progressors, but no patients that went through a lung transplant had blood samples measured by SomaLogic. We were interested in exploring if the way in which patients experienced IPF progression was related to their peripheral blood protein expression. We used LASSO and PLSDA to identify and visualize a signature of 20 blood proteins that differentiated patients who experienced an AE-IPF, a drop in just DLCO or FVC only, or a drop in both DLCO and FVC in the 80 weeks of the COMET study. The PLSDA model performed with high calibration and moderate cross-validation accuracy (95.26% calibration and 82.91% cross-validation accuracy), especially considering that the AE-IPF patients and the both FVC and DLCO patients had low numbers (n = 1 and n = 4, respectively) (**Figure E.6A**, **Figure E.6B**). In addition, when we looked at the individual expression of the proteins in the signature using hierarchical clustering, we did not see strong evidence for unsupervised clustering that corresponded well with our clinical groups (**Figure E.6C**), so we did not pursue these models further.

Validation of the blood protein IPF progression signature with later time points of blood protein data from the COMET cohort

After we identified a signature of blood proteins from week 0 (Tmpt1) of the COMET study that differentiated IPF progressors and non-progressors (**Chapter 3**), we wanted to validate

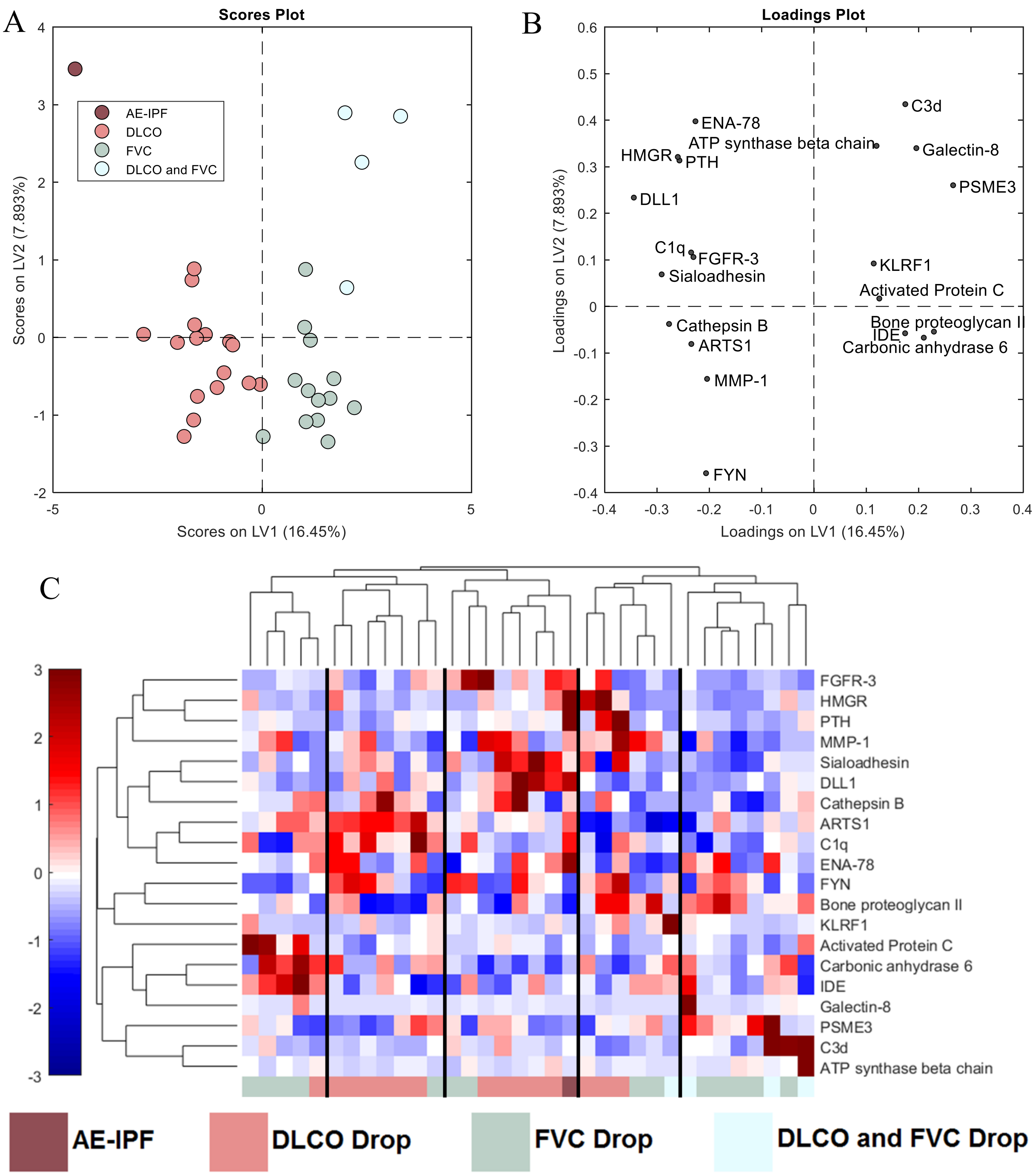


Figure E.6. A PLSDA model based on a LASSO-identified blood protein signature from baseline separates progressors according to how they progressed in the COMET study moderately well.
A. The LASSO-identified signature differentiated IPF progressors by how they progressed (AE-IPF, drop in only DLCO or only FVC, or drops in both DLCO and FVC throughout the 80 weeks of the COMET study) 11.52% calibration and cross-validation accuracy in a PLSDA model. Latent variable 1 captured 16.45% of the variance in the data, and latent variable 2 captured 7.89% of the variance. **B.** The protein loadings associated with this model; proteins loaded in each quadrant are comparatively increased in the group that is scored in the same quadrant. **C.** Hierarchical clustering of the proteins in the LASSO signature did not result in groups that corresponded strongly with the clinical progression groups.

this signature using the protein expression data from these patients at the two other time points in

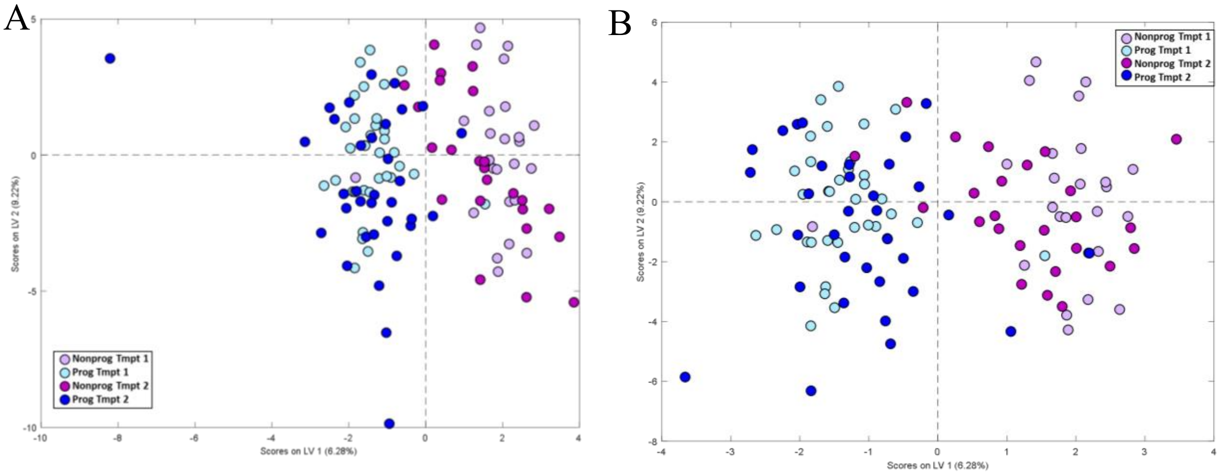


Figure E.7. Validation of our week 0 blood protein signature that differentiated COMET IPF progressors and non-progressors using protein expression data from the COMET patients at A. week 48, and B. week 80.
Nonprog: non-progressor. Prog: progressor. Tmpt 1 = week 0. Tmpt 2 = week 48. Tmpt 3 = week 80.

the study, week 48 (Tmpt2) and week 80 (Tmpt3). We saw on the scores plots after applying the model to the data from week 48 and week 80 that overall, this signature based on week 0 protein expression was still able to separate the same patients at later points throughout the COMET study. Using positive and negative scores on LV1 as the dividing mark between IPF progressors and non-progressors, we saw that there were only two progressors and two non-progressors misclassified based on the week 48 expression of these proteins (**Figure E.7A**), while at the later, week 80-time point, there were three progressors and three non-progressors misclassified (**Figure E.7B**). While this was promising to us and suggested that the signature we identified may be still useful in differentiating IPF progressors and non-progressors even throughout the course of disease progression, we acknowledge that in this case we were validating our original signature using data from the same patients at later time points. This is not the same as using a completely separate and unrelated validation cohort, which could result in biases towards a positive validation of our signature. Thus we decided not to continue further with these results.

Signature of blood and BAL proteins and blood cell markers that differentiated IPF progressors and non-progressors

A subset of blood samples collected at baseline/week 0 of the COMET study were used to measure common cell marker phenotypes in the IPF progressors and non-progressors using flow cytometry. We then investigated a signature of blood and BAL proteins and blood cell markers that could differentiate IPF progression. However, when we applied LASSO to this combined dataset, the best signature that we found that differentiated the subset of COMET-IPF progressors and non-progressors with all three measurements (n = 11 progressors and 8 non-progressors) was only based on blood and BAL proteins (data not shown). Because there were no cell markers selected by LASSO, we could not move forward with any cellular-based hypotheses, and due to the low sample size of patients who also had flow cytometry data, we did not pursue this model further.

Data-driven models from mouse models of pulmonary fibrosis

We had also been interested in applying our systems-focused analysis to mouse models of pulmonary fibrosis so we could infer the most important proteomic relationships associated with bleomycin-caused pulmonary fibrosis. Working with animal models would allow us to formulate hypotheses based on the identified proteomic relationships and test our hypotheses in the same system to validate our models. Working with our collaborator Dr. Beth Moore, we collected blood plasma, BAL, and lung homogenate samples from C57Bl/6 mice who were 21 days post injection with either the fibrosis-causing agent bleomycin (n=11), or with saline (n=5), and measured the concentrations of 32 cytokines in these samples using the Luminex platform. A hydroxyproline assay was used to quantify collagen levels in the lung homogenate. The

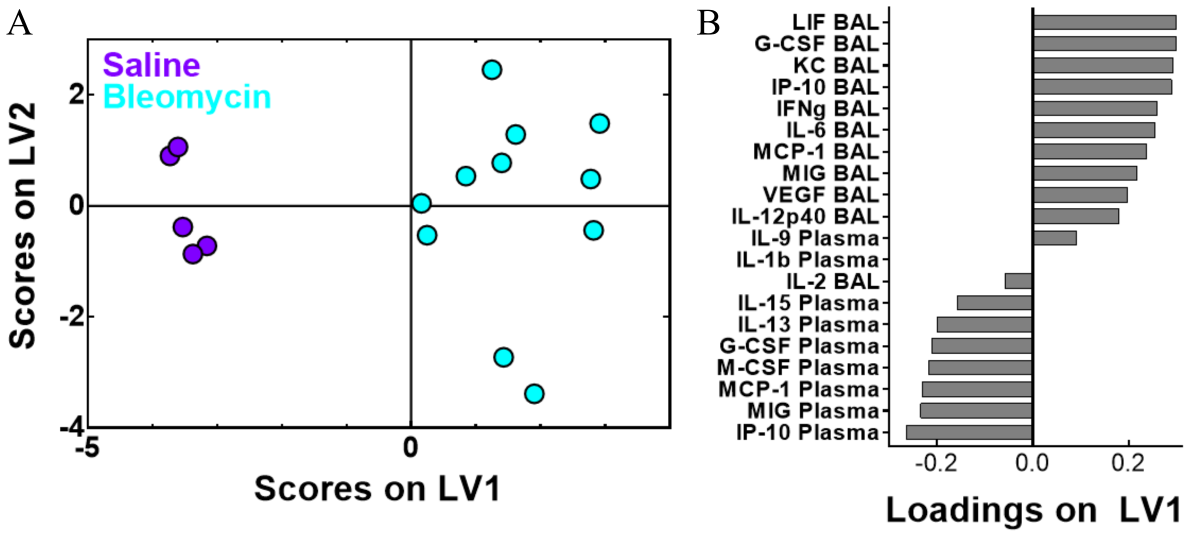


Figure E.8. PLSDA and VIP scores identified protein signatures and hubs that classified control and pulmonary fibrotic mice.

A. A PLSDA model based on the VIP-selected signature differentiated control (purple) and pulmonary fibrotic (cyan) mice with 100% cross-validation and calibration accuracy. **B.** The loadings plot indicates the weights and contributions of the VIP-signature, with positively loaded proteins being comparatively upregulated in pulmonary fibrotic mice, and negatively loaded proteins being comparatively reduced in pulmonary fibrotic mice.

pulmonary fibrotic mice had significantly more collagen formation than the control mice (P<0.0001). A PLSDA model of BAL samples (100% calibration and CV accuracy) was better able to differentiate saline- and bleomycin-treated mice than a PLSDA model of plasma samples (85.45% calibration and 75.45% CV accuracy, data not shown), and a model combining measurements from both samples also classified the groups with very high accuracy. This signature of combined plasma and BAL cytokines differentiated saline- and bleomycin-treated mice with 100% cross-validation and calibration accuracy (**Figure E.8A**). In this signature (**Figure E.8B**), there were increases in BAL chemokines (G-CSF, MCP-1, MIG and IP-10) relative to plasma chemokines in bleomycin-treated mice, suggesting that a specific gradient of chemokines (elevation in lung compared to plasma) was associated with lung fibrosis. Interestingly, both of these inflammatory cytokines signal through heterodimer receptors containing the same subunit (gp130). IL-6 had been known to be associated with bleomycin-induced pulmonary fibrosis[322], although to our knowledge, LIF being associated with fibrosis was a novel finding. Additionally, other researchers have reported that blocking gp130 signaling with the drug tunicamycin[323] caused an increase in fibrosis and lung collagen deposition in bleomycin-exposed mice, although this study was focused on the endoplasmic reticulum (ER) stress caused by tunicamycin[324]. Due to the gp130 signaling pathway being involved in many other basic cellular functions[325] besides what was discussed in ER stress study and the overall lack of novel results from our data-driven model, we decided to shift our focus on to other projects that investigated pulmonary fibrosis in the lungs of human patients.

Protein measurements in sputum samples were able to differentiate stable and AE-COPD

Sample collection was achieved thanks to a published prospective observational trial (ClinicalTrials.gov NCT00281216) has previously been described by Freeman et al.[91], and in **Chapter 5**. In brief, patients were recruited at the VAAAHS and the UMHS. All parts of the study were approved by the IRB at each location; written consent was obtained from each subject; and all parts of the study adhered to the Declaration of Helsinki. Patients were followed for up to three years and were seen at least four times a year for spirometry, clinical evaluations, questionnaires, and collection of peripheral blood and spontaneously expectorated sputum. Exacerbations in these subjects were defined if the subject reported an increase in cough, sputum, or shortness of breath, and if a study physician ordered antibiotics or oral steroids after ruling out pneumonia. Only if a diagnosis of AE-COPD was made were sputum and peripheral blood samples collected at these unexpected visits. After all data and sample collection occurred, then each subject began their treatment for their AE-COPD.

Spontaneously expectorated sputum was immediately processed in a 9:1 mixture of distilled water to Sputolysin® (EMD Millipore, Billercia, MA), and the resulting supernatant was stored at -80°C until protein measurements were made. The Luminex 200 system® (Luminex Corporation, Austin, TX) was used to measure the concentration of 32 cytokines, with ELISA being employed to measure GDF-15, IL-18, IL-23p19, and IFN-β.

Data processing involved removing samples from analysis that were missing more than 25% of the sputum protein measurements, as well as then removing proteins from inclusion in future models if more than two samples had measurements that were missing for that protein. We then used PLSDA to identify signatures of sputum cytokines that differentiated stable and AE-COPD. All data were mean-centered and variance scaled before being modeled using PLSDA. From this PLSDA model, we used VIP scores to select the cytokines that were most influential (defined as cytokines with VIP scores ≥ 1) in differentiating the clinical groups of interest, and then created a new PLSDA model based only on the VIP-selected features. Each PLSDA model was cross-validated to avoid model overfitting and to quantitatively define model accuracy. K-fold cross-validation was performed by splitting the data into seven groups and iteratively training the model on six of the groups while using the seventh group to test the model. All

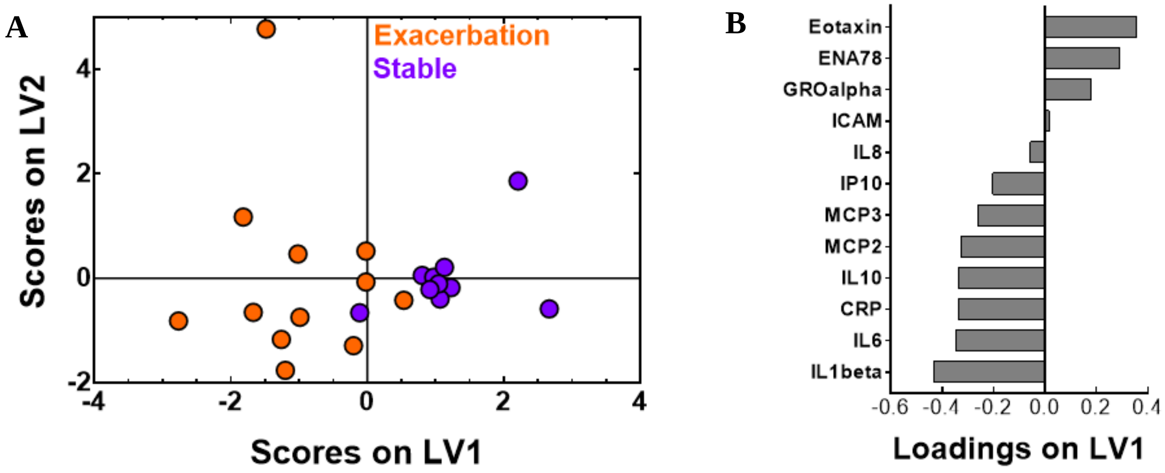


Figure E.9. A PLSDA model based on VIP-selected proteins from the sputum resulted in differentiation between stable and exacerbation measurements.

(**A**) PLSDA and VIP scores identified a signature of 12 sputum proteins that differentiated the stable (purple) from exacerbation (orange) states with 91.67% calibration and 78.33% cross-validation accuracy. (**B**) The loadings plot illustrates the protein contributions to the VIP-selected signature, with negative loadings positively associated with AE-COPD, and positive loadings comparatively reduced.

missing data points in the data were filled in by the Eigenvector software's "best guess." All final PLSDA models were orthogonalized to improve interpretability. All PLSDA models and VIP scores were created or calculated using the PLS toolbox (Eigenvector, Manson, WA) in MATLAB (MATLAB, Natick, MA).

A VIP-selected PLSDA model based on sputum proteins was able to differentiate stable and acute exacerbations of COPD (AE-COPD) with 91.67% calibration and 78.33% cross-validation accuracy (**Figure E.9A**). LV1 separated stable (purple; positive scores on LV1) and AE-COPD (orange; negative scores on LV1) (**Figure E.9B**). Many proteins that were comparatively increased during exacerbation had inflammatory functions (IL-8, IL-6, IL-1β). While this model performed with high calibration accuracy, we decided not to explore it further due to the lower cross-validation accuracy of this model compared to models of AE-COPD based on serum cytokines alone and serum and sputum cytokines in combination with blood flow markers which were discussed in **Chapter 5**.

Classification of smokers, never smokers, and COPD subjects according to clinical measurements

While we have reported success by using blood and lung proteins to differentiate COPD GOLD status in the SPIROMICS cohort, we were curious to explore if blood and lung proteins were also able to differentiate the SPIROMICS cohort COPD patients by numeric clinical measurements of interest, such as the number of reported exacerbations, FEV_1 values measured before the patients underwent the bronchoscopy, and pack years of smoking. Overall, when just looking at the Luminex plasma and Luminex BAL proteins measured separately, we did not see strong separation across latent variable 1 for any of the mentioned clinical measurements in any

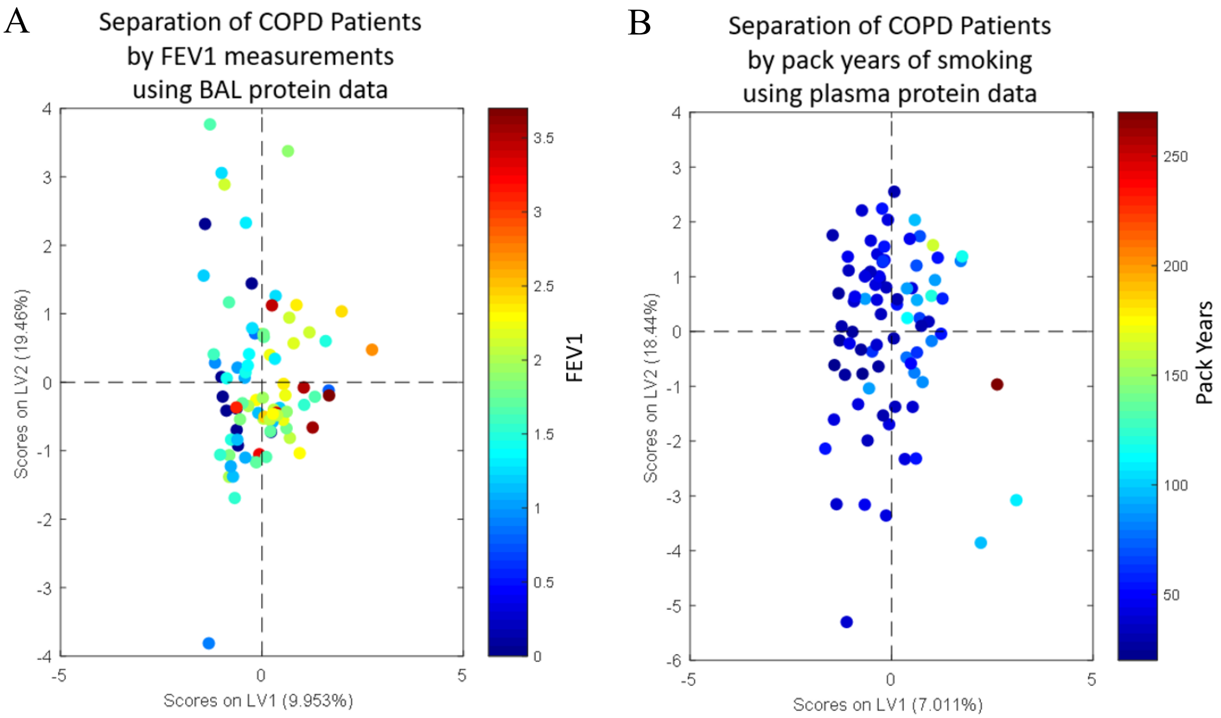


Figure E.10. Unable to resolve COPD patients based on quantitative clinical measurements using blood or BAL proteins alone.
A. PLSR model based on BAL proteins measured by Luminex technology did not separate COPD patients well by FEV1 measurements, with an R^2 calibration measurement of 0.25 and a Q^2 cross-validation measurement of 0.16, where values closer to 1 indicate a better separation. **B.** PLSR model based on plasma proteins measured by Luminex technology did not separate COPD patients well by the number of pack years smoked, with an R^2 calibration measurement of 0.42 and a Q^2 cross-validation measurement of 0.18, where measurements closer to 1 indicate a more accurate separation.

of the partial least squares regression (PLSR) models we created. **Figure E.10A** shows one of the best performing models for the Luminex BAL data (R^2 calibration measurement of 0.2478 and Q^2 cross-validation measurement of 0.1619), which involved differentiating COPD patients based on FEV1 measurements, and **Figure E.10B** shows one of the best blood Luminex models (R^2 calibration measurement of 0.4246 and Q^2 cross-validation measurement of 0.1766), which attempted to differentiate COPD patients based on number of pack years smoked. We hypothesized that these models performed so poorly for two reasons: (1) The overall small number of proteins measured by the Luminex compared to other proteomic technologies, such as SOMAmer aptamers; and (2) The fact that the SPIROMICS COPD bronchoscopy substudy purposefully tried to recruit patients with less severe COPD (e.g. GOLD stage 1 and 2) due prioritizing patient health when exposing them to a relatively invasive procedure. Due to the makeup of the patients included in the study, we did not have a large enough sample of patients with high FEV_1 or large number of pack years smoked to accurately differentiate them from the larger number of patients with moderate disease. In addition, when we included smokers and never smokers with the COPD patients, we saw even worse separation by these quantitative clinical measurements. Thus, we decided not to move forward separating patients by these values and instead focus on categorical separation by disease state or GOLD status.

Bibliography

1. Huff, R. D., Carlsten, C. & Hirota, J. A. An update on immunologic mechanisms in the respiratory mucosa in response to air pollutants. *Journal of Allergy and Clinical Immunology* **143**, 1989–2001 (2019).
2. Xie, M., Liu, X., Cao, X., Guo, M. & Li, X. Trends in prevalence and incidence of chronic respiratory diseases from 1990 to 2017. *Respir. Res.* **21**, 49 (2020).
3. Raghu, G. *et al.* An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *Am. J. Respir. Crit. Care Med.* **183**, 788–824 (2011).
4. Raghu, G. *et al.* Idiopathic pulmonary fibrosis in US Medicare beneficiaries aged 65 years and older: Incidence, prevalence, and survival, 2001-11. *Lancet Respir. Med.* **2**, 566–572 (2014).
5. Lederer, D. J. & Martinez, F. J. Idiopathic Pulmonary Fibrosis. *N. Engl. J. Med.* **378**, 1811–1823 (2018).
6. Ley, B., Brown, K. K. & Collard, H. R. Molecular biomarkers in idiopathic pulmonary fibrosis. *AJP Lung Cell. Mol. Physiol.* **307**, L681–L691 (2014).
7. Noth, I. *et al.* Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: A genome-wide association study. *Lancet Respir. Med.* **1**, 309–317 (2013).
8. Stock, C. J. *et al.* Mucin 5B promoter polymorphism is associated with idiopathic pulmonary fibrosis but not with development of lung fibrosis in systemic sclerosis or sarcoidosis. *Thorax* **68**, 436–441 (2013).
9. Horimasu, Y. *et al.* MUC5B promoter polymorphism in Japanese patients with idiopathic pulmonary fibrosis. *Respirology* **20**, 439–444 (2015).
10. Wei, R. *et al.* Association between MUC5B and TERT polymorphisms and different interstitial lung disease phenotypes. *Transl. Res.* **163**, 494–502 (2014).
11. Zhang, Q., Wang, Y., Qu, D., Yu, J. & Yang, J. The Possible Pathogenesis of Idiopathic Pulmonary Fibrosis considering MUC5B. *Biomed Res. Int.* **2019**, 9712464 (2019).
12. Snijder, J., Peraza, J., Padilla, M., Capaccione, K. & Salvatore, M. M. Pulmonary fibrosis: a disease of alveolar collapse and collagen deposition. *Expert Review of Respiratory Medicine* **13**, 615–619 (2019).
13. Stijn, W. *et al.* Multiplex protein profiling of bronchoalveolar lavage in idiopathic pulmonary fibrosis and hypersensitivity pneumonitis. *Ann. Thorac. Med.* **8**, 38–45 (2013).
14. Mayadas, T. N., Cullere, X. & Lowell, C. A. The Multifaceted Functions of Neutrophils. *Annu. Rev. Pathol. Mech. Dis.* **9**, 181–218 (2014).
15. Prasse, A. *et al.* A vicious circle of alveolar macrophages and fibroblasts perpetuates pulmonary fibrosis via CCL18. *Am. J. Respir. Crit. Care Med.* **173**, 781–792 (2006).
16. Heukels, P., Moor, C. C., von der Thüsen, J. H., Wijsenbeek, M. S. & Kool, M. Inflammation and immunity in IPF pathogenesis and treatment. *Respiratory Medicine* **147**, 79–91 (2019).
17. Xue, J. *et al.* Transcriptome-Based Network Analysis Reveals a Spectrum Model of Human Macrophage Activation. *Immunity* **40**, 274–288 (2014).

18. Murray, P. J. *et al.* Macrophage Activation and Polarization: Nomenclature and Experimental Guidelines. *Immunity* **41**, 14–20 (2014).
19. Ginhoux, F., Schultze, J. L., Murray, P. J., Ochando, J. & Biswas, S. K. New insights into the multidimensional concept of macrophage ontogeny, activation and function. *Nature Immunology* **17**, 34–40 (2016).
20. Hinz, B. *et al.* The myofibroblast: One function, multiple origins. *Am. J. Pathol.* **170**, 1807–1816 (2007).
21. Michalik, M. *et al.* Fibroblast-to-myofibroblast transition in bronchial asthma. *Cellular and Molecular Life Sciences* **75**, 3943–3961 (2018).
22. Willis, B. C. *et al.* Induction of epithelial-mesenchymal transition in alveolar epithelial cells by transforming growth factor-β1: Potential role in idiopathic pulmonary fibrosis. *Am. J. Pathol.* **166**, 1321–1332 (2005).
23. King, T. E., Pardo, A. & Selman, M. Idiopathic pulmonary fibrosis. in *The Lancet* **378**, 1949–1961 (Elsevier, 2011).
24. Tanjore, H. *et al.* Contribution of epithelial-derived fibroblasts to bleomycin-induced lung fibrosis. *Am. J. Respir. Crit. Care Med.* **180**, 657–665 (2009).
25. El Agha, E. *et al.* Mesenchymal Stem Cells in Fibrotic Disease. *Cell Stem Cell* **21**, 166–177 (2017).
26. Martinez, F. J. *et al.* Idiopathic pulmonary fibrosis. *Nature Reviews Disease Primers* **3**, 17074 (2017).
27. Kim, H. J., Perlman, D. & Tomic, R. Natural history of idiopathic pulmonary fibrosis. *Respiratory Medicine* **109**, 661–670 (2015).
28. Du Bois, R. M. *et al.* Forced vital capacity in patients with idiopathic pulmonary fibrosis: Test properties and minimal clinically important difference. *Am. J. Respir. Crit. Care Med.* **184**, 1382–1389 (2011).
29. Russell, A. M. *et al.* Daily home spirometry: An effective tool for detecting progression in idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **194**, 989–997 (2016).
30. Collard, H. R. *et al.* Acute exacerbation of idiopathic pulmonary fibrosis an international working group report. *American Journal of Respiratory and Critical Care Medicine* **194**, 265–275 (2016).
31. Leuschner, G. & Behr, J. Acute exacerbation in interstitial lung disease. *Frontiers in Medicine* **4**, (2017).
32. Schupp, J. C. *et al.* Macrophage activation in acute exacerbation of idiopathic pulmonary fibrosis. *PLoS One* **10**, e0116775 (2015).
33. Raghu, G. *et al.* Diagnosis of idiopathic pulmonary fibrosis An Official ATS/ERS/JRS/ALAT Clinical practice guideline. *Am. J. Respir. Crit. Care Med.* **198**, e44–e68 (2018).
34. Cosgrove, G. P., Bianchi, P., Danese, S. & Lederer, D. J. Barriers to timely diagnosis of interstitial lung disease in the real world: The INTENSITY survey. *BMC Pulm. Med.* **18**, 9 (2018).
35. Rosas, I. O. *et al.* MMP1 and MMP7 as potential peripheral blood biomarkers in idiopathic pulmonary fibrosis. *PLoS Med.* **5**, 0623–0633 (2008).
36. Hamai, K. *et al.* Comparative Study of Circulating MMP-7, CCL18, KL-6, SP-A, and SP-D as Disease Markers of Idiopathic Pulmonary Fibrosis. *Dis. Markers* **2016**, 4759040 (2016).
37. Ohnishi, H. *et al.* Comparative study of KL-6, surfactant protein-A, surfactantprotein-D,

and monocyte chemoattractant protein-1 as serum markersfor interstitial lung diseases. *Am. J. Respir. Crit. Care Med.* **165**, 378–381 (2002).

38. Elhai, M., Avouac, J. & Allanore, Y. Circulating lung biomarkers in idiopathic lung fibrosis and interstitial lung diseases associated with connective tissue diseases: Where do we stand? *Seminars in Arthritis and Rheumatism* 1–12 (2020). doi:10.1016/j.semarthrit.2020.01.006

39. Bauer, Y. *et al.* MMP-7 is a predictive biomarker of disease progression in patients with idiopathic pulmonary fibrosis. *ERJ Open Res.* **3**, 00074–02016 (2017).

40. Tzouvelekis, A. *et al.* Validation of the prognostic value of MMP-7 in idiopathic pulmonary fibrosis. *Respirology* **22**, 486–493 (2017).

41. Prasse, A. *et al.* Serum CC-Chemokine Ligand 18 Concentration Predicts Outcome in Idiopathic Pulmonary Fibrosis. *Am. J. Respir. Crit. Care Med.* **179**, 717–723 (2009).

42. Wakamatsu, K. *et al.* Prognostic value of serial serum KL-6 measurements in patients with idiopathic pulmonary fibrosis. *Respir. Investig.* **55**, 16–23 (2017).

43. TAKAHASHI, H. *et al.* Serum Surfactant Proteins A and D as Prognostic Factors in Idiopathic Pulmonary Fibrosis and Their Relationship to Disease Extent. *Am. J. Respir. Crit. Care Med.* **162**, 1109–1114 (2000).

44. Barlo, N. *et al.* Surfactant protein-D predicts survival in patients with idiopathic pulmonary fibrosis. *Sarcoidosis, Vasc. Diffus. lung Dis.* **26**, 155–161 (2009).

45. Raghu, G. *et al.* Idiopathic Pulmonary Fibrosis: Prospective, Case-Controlled Study of Natural History and Circulating Biomarkers. *Chest* **154**, 1359–1370 (2018).

46. Drakopanagiotakis, F., Wujak, L., Wygrecka, M. & Markart, P. Biomarkers in idiopathic pulmonary fibrosis. *Matrix Biol.* **68–69**, 404–421 (2018).

47. Idiopathic Pulmonary Fibrosis Clinical Research Network *et al.* Prednisone, Azathioprine, and N -Acetylcysteine for Pulmonary Fibrosis. *N. Engl. J. Med.* **366**, 1968–1977 (2012).

48. King, T. E. *et al.* A Phase 3 Trial of Pirfenidone in Patients with Idiopathic Pulmonary Fibrosis. *N. Engl. J. Med.* **370**, 2083–2092 (2014).

49. Richeldi, L. *et al.* Efficacy and Safety of Nintedanib in Idiopathic Pulmonary Fibrosis. *N. Engl. J. Med.* **370**, 2071–2082 (2014).

50. Maher, T. M. & Strek, M. E. Antifibrotic therapy for idiopathic pulmonary fibrosis: Time to treat. *Respiratory Research* **20**, 205 (2019).

51. Raghu, G. & Richeldi, L. Current approaches to the management of idiopathic pulmonary fibrosis. *Respiratory Medicine* **129**, 24–30 (2017).

52. Somogyi, V. *et al.* The therapy of idiopathic pulmonary fibrosis: What is next? *European Respiratory Review* **28**, 190021 (2019).

53. Murphy, S. L., Xu, J., Kochanek, K. D. & Arias, E. Mortality in the United States, 2017. *NCHS Data Brief* 1–8 (2018).

54. Riley, C. M. & Sciurba, F. C. Diagnosis and Outpatient Management of Chronic Obstructive Pulmonary Disease: A Review. *JAMA - Journal of the American Medical Association* **321**, 786–797 (2019).

55. López-Campos, J. L., Tan, W. & Soriano, J. B. Global burden of COPD. *Respirology* **21**, 14–23 (2016).

56. Soriano, J. B. *et al.* Global, regional, and national deaths, prevalence, disability-adjusted life years, and years lived with disability for chronic obstructive pulmonary disease and asthma, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Respir. Med.* **5**, 691–706 (2017).

57. Ford, E. S. *et al.* Total and State-Specific Medical and Absenteeism Costs of COPD Among Adults Aged ≥ 18 Years in the United States for 2010 and Projections Through 2020. *Chest* **147**, 31–45 (2015).
58. Mirza, S., Clay, R. D., Koslow, M. A. & Scanlon, P. D. COPD Guidelines: A Review of the 2018 GOLD Report. *Mayo Clinic Proceedings* **93**, 1488–1502 (2018).
59. Vogelmeier, C. F. *et al.* Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2017 report. *American Journal of Respiratory and Critical Care Medicine* **195**, 557–582 (2017).
60. Duffy, S. P. & Criner, G. J. Chronic Obstructive Pulmonary Disease: Evaluation and Management. *Medical Clinics of North America* **103**, 453–461 (2019).
61. Agustí, A. & Hogg, J. C. Update on the Pathogenesis of Chronic Obstructive Pulmonary Disease. *N. Engl. J. Med.* **381**, 1248–1256 (2019).
62. Lange, P. *et al.* Lung-Function Trajectories Leading to Chronic Obstructive Pulmonary Disease. *N. Engl. J. Med.* **373**, 111–122 (2015).
63. Hikichi, M., Mizumura, K., Maruoka, S. & Gon, Y. Pathogenesis of chronic obstructive pulmonary disease (COPD) induced by cigarette smoke. *Journal of Thoracic Disease* **11**, S2129–S2140 (2019).
64. Eapen, M. S. *et al.* Abnormal M1/M2 macrophage phenotype profiles in the small airway wall and lumen in smokers and chronic obstructive pulmonary disease (COPD). *Sci. Rep.* **7**, 13392 (2017).
65. Rabe, K. F. & Watz, H. Chronic obstructive pulmonary disease. *The Lancet* **389**, 1931–1940 (2017).
66. Barnes, P. J. Inflammatory mechanisms in patients with chronic obstructive pulmonary disease. *Journal of Allergy and Clinical Immunology* **138**, 16–27 (2016).
67. Taylor, A. E. *et al.* Defective macrophage phagocytosis of bacteria in COPD. *Eur. Respir. J.* **35**, 1039–1047 (2010).
68. Hogg, J. C. *et al.* The Nature of Small-Airway Obstruction in Chronic Obstructive Pulmonary Disease. *N. Engl. J. Med.* **350**, 2645–2653 (2004).
69. Gamble, E. *et al.* Airway mucosal inflammation in COPD is similar in smokers and ex-smokers: A pooled analysis. *Eur. Respir. J.* **30**, 467–471 (2007).
70. Scanlon, P. D. *et al.* Smoking cessation and lung function in mild-to-moderate chronic obstructive pulmonary disease: The lung health study. *Am. J. Respir. Crit. Care Med.* **161**, 381–390 (2000).
71. Johannesdottir, S. A. *et al.* Hospitalization with acute exacerbation of chronic obstructive pulmonary disease and associated health resource utilization: A population-based Danish cohort study. *J. Med. Econ.* **16**, 897–906 (2013).
72. Roche, N. *et al.* Predictors of outcomes in COPD exacerbation cases presenting to the emergency department. *Eur. Respir. J.* **32**, 953–961 (2008).
73. Jinjuvadia, C. *et al.* Trends in Outcomes, Financial Burden, and Mortality for Acute Exacerbation of Chronic Obstructive Pulmonary Disease (COPD) in the United States from 2002 to 2010. *COPD J. Chronic Obstr. Pulm. Dis.* **14**, 72–79 (2017).
74. Seemungal, T. A. R., Donaldson, G. C., Bhowmik, A., Jeffries, D. J. & Wedzicha, J. A. Time course and recovery of exacerbations in patients with chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* **161**, 1608–1613 (2000).
75. Wedzicha, J. A. *et al.* Management of COPD exacerbations: A European Respiratory Society/American Thoracic Society guideline. *European Respiratory Journal* **49**, 1600791

(2017).

76. Hurst, J. R. *et al.* Susceptibility to Exacerbation in Chronic Obstructive Pulmonary Disease. *N. Engl. J. Med.* **363**, 1128–1138 (2010).

77. McGarvey, L. *et al.* Characterisation of the frequent exacerbator phenotype in COPD patients in a large UK primary care population. *Respir. Med.* **109**, 228–237 (2015).

78. Wedzicha, J. A. & Seemungal, T. A. COPD exacerbations: defining their cause and prevention. *Lancet* **370**, 786–796 (2007).

79. Han, M. L. K. *et al.* Frequency of exacerbations in patients with chronic obstructive pulmonary disease: an analysis of the SPIROMICS cohort. *Lancet Respir. Med.* **5**, 619–626 (2017).

80. Vogelmeier, C. F. *et al.* Goals of COPD treatment: Focus on symptoms and exacerbations. *Respiratory Medicine* **166**, 105938 (2020).

81. Halpin, D. M. G., Miravitlles, M., Metzdorf, N. & Celli, B. Impact and prevention of severe exacerbations of COPD: A review of the evidence. *International Journal of COPD* **12**, 2891–2908 (2017).

82. Sidhaye, V. K., Nishida, K. & Martinez, F. J. Precision medicine in COPD: Where are we and where do we need to go? *Eur. Respir. Rev.* **27**, 180022 (2018).

83. Hurst, J. R. Consolidation and Exacerbation of COPD. *Med. Sci.* **6**, 44 (2018).

84. Garudadri, S. & Woodruff, P. G. Targeting chronic obstructive pulmonary disease phenotypes, endotypes, and biomarkers. *Ann. Am. Thorac. Soc.* **15**, S234–S238 (2018).

85. Barnes, N. C., Sharma, R., Lettis, S. & Calverley, P. M. A. Blood eosinophils as a marker of response to inhaled corticosteroids in COPD. *Eur. Respir. J.* **47**, 1374–1382 (2016).

86. Bafadhel, M. *et al.* Predictors of exacerbation risk and response to budesonide in patients with chronic obstructive pulmonary disease: a post-hoc analysis of three randomised trials. *Lancet Respir. Med.* **6**, 117–126 (2018).

87. Pavord, I. D. *et al.* Blood eosinophils and inhaled corticosteroid/longacting β-2 agonist efficacy in COPD. *Thorax* **71**, 118–125 (2016).

88. Christenson, S. A. *et al.* An airway epithelial IL-17A response signature identifies a steroid-unresponsive COPD patient subgroup. *J. Clin. Invest.* **129**, 169–181 (2019).

89. Hurst, J. R. *et al.* Use of Plasma Biomarkers at Exacerbation of Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Crit. Care Med.* **174**, 867–874 (2006).

90. Mutlu, L. C. *et al.* Growth Differentiation Factor-15 Is a Novel Biomarker Predicting Acute Exacerbation of Chronic Obstructive Pulmonary Disease. *Inflammation* **38**, 1805–1813 (2015).

91. Freeman, C. M. *et al.* Acute exacerbations of chronic obstructive pulmonary disease are associated with decreased CD4+ & CD8+ T cells and increased growth & differentiation factor-15 (GDF-15) in peripheral blood. *Respir. Res.* **16**, 94 (2015).

92. Bafadhel, M. *et al.* Acute exacerbations of chronic obstructive pulmonary disease: identification of biologic clusters and their biomarkers. *Am. J. Respir. Crit. Care Med.* **184**, 662–671 (2011).

93. Damera, G. *et al.* A Sputum Proteomic Signature That Associates with Increased IL-1β Levels and Bacterial Exacerbations of COPD. *Lung* **194**, 363–369 (2016).

94. Moon, J.-Y., Leitao Filho, F. S., Shahangian, K., Takiguchi, H. & Sin, D. D. Blood and sputum protein biomarkers for chronic obstructive pulmonary disease (COPD). *Expert Review of Proteomics* **15**, 923–935 (2018).

95. Keene, J. D. *et al.* Biomarkers Predictive of Exacerbations in the SPIROMICS and

COPDGene Cohorts. *Am. J. Respir. Crit. Care Med.* **195**, 473–481 (2017).

96. Miller, B. E. *et al.* Plasma Fibrinogen Qualification as a Drug Development Tool in Chronic Obstructive Pulmonary Disease. Perspective of the Chronic Obstructive Pulmonary Disease Biomarker Qualification Consortium. *Am. J. Respir. Crit. Care Med.* **193**, 607–613 (2016).
97. Vasilescu, D. M. *et al.* Noninvasive imaging biomarker identifies small airway damage in severe chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* **200**, 575–581 (2019).
98. Ideker, T., Galitski, T. & Hood, L. A New Approach To Decoding Life: Systems Biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343–372 (2001).
99. Starchenko, A. & Lauffenburger, D. A. In vivo systems biology approaches to chronic immune/inflammatory pathophysiology. *Current Opinion in Biotechnology* **52**, 9–16 (2018).
100. Norman, K. C., Moore, B. B., Arnold, K. B. & O'Dwyer, D. N. Proteomics: Clinical and research applications in respiratory diseases. *Respirology* **23**, 993–1003 (2018).
101. Benedict, K. F. & Lauffenburger, D. A. Insights into Proteomic Immune Cell Signaling and Communication via Data-Driven Modeling. in *Current topics in microbiology and immunology* **363**, 201–233 (2012).
102. Arnold, K. B. *et al.* Increased levels of inflammatory cytokines in the female reproductive tract are associated with altered expression of proteases, mucosal barrier proteins, and an influx of HIV-susceptible target cells. *Mucosal Immunol.* **9**, 194–205 (2016).
103. Liebenberg, L. J. P. *et al.* Genital-systemic chemokine gradients and the risk of HIV acquisition in women. *J. Acquir. Immune Defic. Syndr.* **74**, 318–325 (2017).
104. Banchereau, R. *et al.* Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. *Cell* **165**, 551–565 (2016).
105. Chiche, L. *et al.* Modular transcriptional repertoire analyses of adults with systemic lupus erythematosus reveal distinct type i and type ii interferon signatures. *Arthritis Rheumatol.* **66**, 1583–1595 (2014).
106. Rao, D. A. *et al.* Pathologically expanded peripheral T helper cell subset drives B cells in rheumatoid arthritis. *Nature* **542**, 110–114 (2017).
107. Jones, D. S. *et al.* Profiling drugs for rheumatoid arthritis that inhibit synovial fibroblast activation. *Nat. Chem. Biol.* **13**, 38–45 (2017).
108. Ammar, R., Sivakumar, P., Jarai, G. & Thompson, J. R. A robust data-driven genomic signature for idiopathic pulmonary fibrosis with applications for translational model selection. *PLoS One* **14**, e0215565 (2019).
109. Maher, T. M. *et al.* An epithelial biomarker signature for idiopathic pulmonary fibrosis: an analysis from the multicentre PROFILE cohort study. *Lancet Respir. Med.* **5**, 946–955 (2017).
110. White, E. S. *et al.* Plasma surfactant protein-D, matrix metalloproteinase-7, and osteopontin index distinguishes idiopathic pulmonary fibrosis from other idiopathic interstitial pneumonias. *Am. J. Respir. Crit. Care Med.* **194**, 1242–1251 (2016).
111. Todd, J. L. *et al.* Peripheral blood proteomic profiling of idiopathic pulmonary fibrosis biomarkers in the multicentre IPF-PRO Registry. *Respir. Res.* **20**, 227 (2019).
112. Zemans, R. L. *et al.* Multiple biomarkers predict disease severity, progression and mortality in COPD. *Respir. Res.* **18**, 117 (2017).
113. Martinez, F. J. *et al.* The Clinical Course of Patients with Idiopathic Pulmonary Fibrosis.

Ann. Intern. Med. **142**, 963–967 (2005).

114. Richards, T. J. *et al.* Peripheral Blood Proteins Predict Mortality in Idiopathic Pulmonary Fibrosis. *Am. J. Respir. Crit. Care Med.* **185**, 67–76 (2012).
115. Huang, Y. *et al.* A functional genomic model for predicting prognosis in idiopathic pulmonary fibrosis. *BMC Pulm. Med.* **15**, 147 (2015).
116. Herazo-Maya, J. D. *et al.* Peripheral Blood Mononuclear Cell Gene Expression Profiles Predict Poor Outcome in Idiopathic Pulmonary Fibrosis. *Sci. Transl. Med.* **5**, 205ra136 (2013).
117. Yang, I. V *et al.* The peripheral blood transcriptome identifies the presence and extent of disease in idiopathic pulmonary fibrosis. *PLoS One* **7**, e37708 (2012).
118. Seibold, M. A. *et al.* A common MUC5B promoter polymorphism and pulmonary fibrosis. *N. Engl. J. Med.* **364**, 1503–1512 (2011).
119. O'Dwyer, D. N. *et al.* The toll-like receptor 3 L412F polymorphism and disease progression in idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **188**, 1442–1450 (2013).
120. Peljto, A. L. *et al.* Association between the MUC5B promoter polymorphism and survival in patients with idiopathic pulmonary fibrosis. *JAMA - J. Am. Med. Assoc.* **309**, 2232–2239 (2013).
121. Yang, S. *et al.* Participation of miR-200 in pulmonary fibrosis. *Am. J. Pathol.* **180**, 484–493 (2012).
122. Liu, G. *et al.* miR-21 mediates fibrogenic activation of pulmonary fibroblasts and lung fibrosis. *J. Exp. Med.* **207**, 1589–1597 (2010).
123. Yang, G. *et al.* Discovery and validation of extracellular/circulating microRNAs during idiopathic pulmonary fibrosis disease progression. *Gene* **562**, 138–144 (2015).
124. Goodwin, A. T. & Jenkins, G. Molecular endotyping of pulmonary fibrosis. *Chest* **149**, 228–237 (2016).
125. Brownell, R. *et al.* Precision medicine: The new frontier in idiopathic pulmonary fibrosis. *American Journal of Respiratory and Critical Care Medicine* **193**, 1213–1218 (2016).
126. Foster, M. W. *et al.* Quantitative proteomics of bronchoalveolar lavage fluid in idiopathic pulmonary fibrosis. *J. Proteome Res.* **14**, 1238–1249 (2015).
127. Korfei, M. *et al.* Comparative proteome analysis of lung tissue from patients with idiopathic pulmonary fibrosis (IPF), non-specific interstitial pneumonia (NSIP) and organ donors. *J. Proteomics* **85**, 109–128 (2013).
128. Rottoli, P. *et al.* Cytokine profile and proteome analysis in bronchoalveolar lavage of patients with sarcoidosis, pulmonary fibrosis associated with systematic sclerosis and idiopathic pulmonary fibrosis. *Proteomics* **5**, 1423–1430 (2005).
129. Landi, C. *et al.* A system biology study of BALF from patients affected by idiopathic pulmonary fibrosis (IPF) and healthy controls. *Proteomics - Clin. Appl.* **8**, 932–950 (2014).
130. Gold, L. *et al.* Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery. *PLoS One* **5**, e15004 (2010).
131. Hathout, Y. *et al.* Large-scale serum protein biomarker discovery in Duchenne muscular dystrophy. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7153–7158 (2015).
132. Sattlecker, M. *et al.* Alzheimer's disease biomarker discovery using SOMAscan multiplexed protein technology. *Alzheimer's Dement.* **10**, 724–734 (2014).
133. De Groote, M. A. *et al.* Elucidating Novel Serum Biomarkers Associated with Pulmonary

Tuberculosis Treatment. *PLoS One* **8**, e61002 (2013).

134. Ostroff, R. M. *et al.* Early Detection of Malignant Pleural Mesothelioma in Asbestos-Exposed Individuals with a Noninvasive Proteomics-Based Surveillance Tool. *PLoS One* **7**, e46091 (2012).
135. Ganz, P. *et al.* Development and validation of a protein-based risk score for cardiovascular outcomes among patients with stable coronary heart disease. *JAMA - J. Am. Med. Assoc.* **315**, 2532–2541 (2016).
136. Ashley, S. L. *et al.* Six-SOMAmer Index Relating to Immune, Protease and Angiogenic Functions Predicts Progression in IPF. *PLoS One* **11**, e0159878 (2016).
137. Han, M. K. *et al.* Lung microbiome and disease progression in idiopathic pulmonary fibrosis: an analysis of the COMET study. *Lancet Respir. Med.* **2**, 548–556 (2014).
138. Molyneaux, P. L. *et al.* The role of bacteria in the pathogenesis and progression of idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **190**, 906–913 (2014).
139. Knippenberg, S. *et al.* Streptococcus pneumoniae triggers progression of pulmonary fibrosis through pneumolysin. *Thorax* **70**, 636–646 (2015).
140. Kawai, T. & Akira, S. Toll-like Receptors and Their Crosstalk with Other Innate Receptors in Infection and Immunity. *Immunity* **34**, 637–650 (2011).
141. Tanaka, C. *et al.* Inducible costimulator ligand regulates bleomycin-induced lung and skin fibrosis in a mouse model independently of the inducible costimulator/inducible costimulator ligand pathway. *Arthritis Rheum.* **62**, 1723–1732 (2010).
142. Moore, B. B. & Moore, T. A. Viruses in idiopathic pulmonary fibrosis etiology and exacerbation. in *Annals of the American Thoracic Society* **12**, S186–S192 (American Thoracic Society, 2015).
143. Grimminger, F., Günther, A. & Vancheri, C. The role of tyrosine kinases in the pathogenesis of idiopathic pulmonary fibrosis. *Eur. Respir. J.* **45**, 1426–1433 (2015).
144. Hilberg, F. *et al.* BIBF 1120: Triple angiokinase inhibitor with sustained receptor blockade and good antitumor efficacy. *Cancer Res.* **68**, 4774–4782 (2008).
145. Wollin, L., Maillet, I., Quesniaux, V., Holweg, A. & Ryffel, B. Antifibrotic and anti-inflammatory activity of the tyrosine kinase inhibitor nintedanib in experimental models of lung fibrosis. *J. Pharmacol. Exp. Ther.* **349**, 209–220 (2014).
146. Koch, S. & Claesson-Welsh, L. Signal transduction by vascular endothelial growth factor receptors. *Cold Spring Harbor Perspectives in Medicine* **2**, a006502 (2012).
147. Dey, J. H. *et al.* Targeting fibroblast growth factor receptors blocks PI3K/AKT signaling, induces apoptosis, and impairs mammary tumor outgrowth and metastasis. *Cancer Res.* **70**, 4151–4162 (2010).
148. Nethery, D. E. *et al.* Expression of mutant human epidermal receptor 3 attenuates lung fibrosis and improves survival in mice. *J. Appl. Physiol.* **99**, 298–307 (2005).
149. Faress, J. A. *et al.* Bleomycin-induced pulmonary fibrosis is attenuated by a monoclonal antibody targeting HER2. *J. Appl. Physiol.* **103**, 2077–2083 (2007).
150. Buckley, S. *et al.* Increased alveolar soluble Annexin V promotes lung inflammation and fibrosis. *Eur. Respir. J.* **46**, 1417–1429 (2015).
151. Crooks, M. G., Fahim, A., Naseem, K. M., Morice, A. H. & Hart, S. P. Increased platelet reactivity in idiopathic pulmonary fibrosis is mediated by a plasma factor. *PLoS One* **9**, e111347 (2014).
152. Zorzetto, M. *et al.* Complement receptor 1 gene polymorphisms are associated with idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **168**, 330–334 (2003).

153. Gu, H. *et al.* Crosstalk between TGF-β1 and complement activation augments epithelial injury in pulmonary fibrosis. *FASEB J.* **28**, 4223–4234 (2014).

154. Gu, H. *et al.* Contribution of the anaphylatoxin receptors, C3aR and C5aR, to the pathogenesis of pulmonary fibrosis. *FASEB J.* **30**, 2336–2350 (2016).

155. Herro, R., Da Silva Antunes, R., Aguilera, A. R., Tamada, K. & Croft, M. Tumor necrosis factor superfamily 14 (LIGHT) controls thymic stromal lymphopoietin to drive pulmonary fibrosis. *J. Allergy Clin. Immunol.* **136**, 757–768 (2015).

156. Gurrieri, C. *et al.* 3-(2,4-dichlorophenyl)-4-(1-methyl-1H-indol-3-yl)-1H-pyrrole-2,5-dione (SB216763), a glycogen synthase kinase-3 inhibitor, displays therapeutic properties in a mouse model of pulmonary inflammation and fibrosis. *J. Pharmacol. Exp. Ther.* **332**, 785–94 (2010).

157. Fujimoto, H. *et al.* Thrombin-activatable fibrinolysis inhibitor and protein C inhibitor in interstitial lung disease. *Am. J. Respir. Crit. Care Med.* **167**, 1687–1694 (2003).

158. Kobayashi, H. *et al.* Protein C anticoagulant system in patients with interstitial lung disease. *Am. J. Respir. Crit. Care Med.* **157**, 1850–1854 (1998).

159. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1994).

160. Hochberg, Y. & Benjamini, Y. More powerful procedures for multiple significance testing. *Stat. Med.* **9**, 811–818 (1990).

161. Bonferroni, C. . *Il calcolo delle assicurazioni su gruppi di teste*. (Studi in Onore del Professore Salvatore Ortu Carboni, 1935).

162. Bonferroni, C. . Teoria statistica delle classi e calcolo delle probabilita. *Pubbl. del R Inst. Super. di Sci. Econ. e Commer. di Firenze* **8**, 3–62 (1936).

163. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).

164. Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091–1093 (2009).

165. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).

166. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics Enrichment Tools: Paths Toward the Comprehensive Functional Analysis of Large Gene Lists - PubMed. *Nucleic Acids Res.* **31**, 1–13 (2009).

167. Qian, J., Hastie, T., Friedman, J., Tibshirani, R. & Simon, N. Glmnet for Matlab. (2013).

168. Lin, C., Borensztajn, K. & Spek, C. A. Targeting coagulation factor receptors - protease-activated receptors in idiopathic pulmonary fibrosis. *J. Thromb. Haemost.* **15**, 597–607 (2017).

169. Yusen, R. D. *et al.* The Registry of the International Society for Heart and Lung Transplantation: Thirty-third Adult Lung and Heart–Lung Transplant Report—2016; Focus Theme: Primary Diagnostic Indications for Transplant. *J. Hear. Lung Transplant.* **35**, 1170–1184 (2016).

170. Jenkins, R. G. *et al.* Longitudinal change in collagen degradation biomarkers in idiopathic pulmonary fibrosis: An analysis from the prospective, multicentre PROFILE study. *Lancet Respir. Med.* **3**, 462–472 (2015).

171. Chien, J. W. *et al.* Serum lysyl oxidase-like 2 levels and idiopathic pulmonary fibrosis disease progression. *Eur. Respir. J.* **43**, 1430–1438 (2014).

172. Bouros, D. *et al.* Design, Rationale, Methodology, and Aims of a Greek Prospective

Idiopathic Pulmonary Fibrosis Registry: Investigating Idiopathic Pulmonary Fibrosis in Greece (INDULGE IPF). *Respiration* **96**, 41–47 (2018).

173. Pleasants, R. & Tighe, R. M. Management of Idiopathic Pulmonary Fibrosis. *Ann. Pharmacother.* **53**, 1238–1248 (2019).
174. O'Dwyer, D. N. *et al.* The peripheral blood proteome signature of idiopathic pulmonary fibrosis is distinct from normal and is associated with novel immunological processes. *Sci. Rep.* **7**, 46560 (2017).
175. Norman, K. C. *et al.* Inference of Cellular Immune Environments in Sputum and Peripheral Blood Associated with Acute Exacerbations of COPD. *Cell. Mol. Bioeng.* **12**, 165–177 (2019).
176. Lau, K. S. *et al.* In vivo systems analysis identifies spatial and temporal aspects of the modulation of TNF-α-induced apoptosis and proliferation by MAPKs. *Sci. Signal.* **4**, ra16 (2011).
177. Jaffar, J. *et al.* Fibulin-1 predicts disease progression in patients with idiopathic pulmonary fibrosis. *Chest* **146**, 1055–1063 (2014).
178. Song, J. W. *et al.* Blood biomarkers MMP-7 and SP-A: Predictors of outcome in idiopathic pulmonary fibrosis. *Chest* **143**, 1422–1429 (2013).
179. Beckwith-Hall, B. M. *et al.* Nuclear Magnetic Resonance Spectroscopic and Principal Components Analysis Investigations into Biochemical Effects of Three Model Hepatotoxins. *Chem. Res. Toxicol.* **11**, 260–272 (1998).
180. Sahu, A. & Lambris, J. D. Structure and biology of complement protein C3, a connecting link between innate and acquired immunity. *Immunol. Rev.* **180**, 35–48 (2001).
181. Foley, J. H. *et al.* Interplay between fibrinolysis and complement: Plasmin cleavage of iC3b modulates immune responses. *J. Thromb. Haemost.* **13**, 610–618 (2015).
182. Okamoto, T. *et al.* The relationship between complement C3 expression and the MUC5B genotype in pulmonary fibrosis. *Am. J. Physiol. Cell. Mol. Physiol.* **315**, L1–L10 (2018).
183. Phan, S. H. & Thrall, R. S. Inhibition of bleomycin-induced pulmonary fibrosis by cobra venom factor. *Am. J. Pathol.* **107**, 25–28 (1982).
184. Haslam, P. & Baughman, R. Report of ERS Task Force: guidelines for measurement of acellular components and standardization of BAL. *Eur Respir J* **14**, 245–248 (1999).
185. Lee, J.-U. *et al.* Upregulation of interleukin-33 and thymic stromal lymphopoietin levels in the lungs of idiopathic pulmonary fibrosis. *BMC Pulm. Med.* **17**, 39 (2017).
186. Shinoda, H. *et al.* Elevated CC Chemokine Level in Bronchoalveolar Lavage Fluid Is Predictive of a Poor Outcome of Idiopathic Pulmonary Fibrosis. *Respiration* **78**, 285–292 (2009).
187. Gharsalli, H. *et al.* The utility of bronchoalveolar lavage in the evaluation of interstitial lung diseases: A clinicopathological perspective. *Seminars in Diagnostic Pathology* **35**, 280–287 (2018).
188. Kebbe, J. & Abdo, T. Interstitial lung disease: The diagnostic role of bronchoscopy. *Journal of Thoracic Disease* **9**, S996–S1010 (2017).
189. Meyer, K. C. *et al.* An official American Thoracic Society clinical practice guideline: The clinical utility of bronchoalveolar lavage cellular analysis in interstitial lung disease. *American Journal of Respiratory and Critical Care Medicine* **185**, 1004–1014 (2012).
190. Willems, S. *et al.* Multiplex protein profiling of bronchoalveolar lavage in idiopathic pulmonary fibrosis and hypersensitivity pneumonitis. *Ann. Thorac. Med.* **8**, 38–45 (2013).
191. Ronan, N. *et al.* Tissue and Bronchoalveolar Lavage Biomarkers in Idiopathic Pulmonary

Fibrosis Patients on Pirfenidone. *Lung* **196**, 543–552 (2018).

192. Carleo, A. *et al.* Comparative proteomic analysis of bronchoalveolar lavage of familial and sporadic cases of idiopathic pulmonary fibrosis. *J. Breath Res.* **10**, 026007 (2016).
193. Vasakova, M. *et al.* Bronchoalveolar lavage fluid cellular characteristics, functional parameters and cytokine and chemokine levels in interstitial lung diseases. *Scand. J. Immunol.* **69**, 268–274 (2009).
194. Nishikiori, H. *et al.* Distinct compartmentalization of SP-A and SP-D in the vasculature and lungs of patients with idiopathic pulmonary fibrosis. *BMC Pulm. Med.* **14**, 196 (2014).
195. Kucejko, W., Chyczewska, E., Naumnik, W. & Ossolińska, M. Concentration of surfactant protein D, Clara cell protein CC-16 and IL-10 in bronchoalveolar lavage (BAL) in patients with sarcoidosis, hypersensivity pneumonitis and idiopathic pulmonary fibrosis. *Folia Histochem. Cytobiol.* **47**, 225–230 (2009).
196. Gui, X. *et al.* Prognostic value of IFN-γ sCD163, CCL2 and CXCL10 involved in acute exacerbation of idiopathic pulmonary fibrosis. *Int. Immunopharmacol.* **70**, 208–215 (2019).
197. Norman, K. C. *et al.* Identification of a unique temporal signature in blood and BAL associated with IPF progression. *Sci. Rep.* **10**, 12049 (2020).
198. Gordon, S. & Taylor, P. R. Monocyte and macrophage heterogeneity. *Nature Reviews Immunology* **5**, 953–964 (2005).
199. Byrne, A. J., Maher, T. M. & Lloyd, C. M. Pulmonary Macrophages: A New Therapeutic Pathway in Fibrosing Lung Disease? *Trends in Molecular Medicine* **22**, 303–316 (2016).
200. Henry, M. T. *et al.* Matrix metalloproteinases and tissue inhibitor of metalloproteinase-1 in sarcoidosis and IPF. *Eur. Respir. J.* **20**, 1220–1227 (2002).
201. Desai, O., Winkler, J., Minasyan, M. & Herzog, E. L. The role of immune and inflammatory cells in idiopathic pulmonary fibrosis. *Frontiers in Medicine* **5**, (2018).
202. Hetzel, M., Bachem, M., Anders, D., Trischler, G. & Faehling, M. Different Effects of Growth Factors on Proliferation and Matrix Production of Normal and Fibrotic Human Lung Fibroblasts. *Lung* **183**, 225–237 (2005).
203. Furuie, H., Yamasaki, H., Suga, M. & Ando, M. Altered accessory cell function of alveolar macrophages: a possible mechanism for induction of Th2 secretory profile in idiopathic pulmonary fibrosis. *Eur. Respir. J.* **10**, 787–794 (1997).
204. Park, S.-W. *et al.* Interleukin-13 and Its Receptors in Idiopathic Interstitial Pneumonia: Clinical Implications for Lung Function. *J. Korean Med. Sci.* **24**, 614–620 (2009).
205. Arnold, K. B., Szeto, G. L., Alter, G., Irvine, D. J. & Lauffenburger, D. A. CD4+ T cell-dependent and CD4+ T cell-independent cytokine-chemokine network changes in the immune responses of HIV-infected individuals. *Sci. Signal.* **8**, ra104 (2015).
206. Strieter, R., Miller, E., Kurdowska, A., Reid, P. & Donnelly, S. Measurement of cytokines in bronchoalveolar lavage fluid. *Rep. ERS Task Force Guidel. measruement acellular components Recomm. Stand. bronchoalveolar lavage (BAL). Eur. Respir. Rev.* **9**, 106–112 (1999).
207. Kochanek, K. D., Murphy, S. L., Xu, J. & Tejada-Vera, B. Deaths: Final Data for 2014. *Natl. vital Stat. reports* **65**, 1–122 (2016).
208. Rodriguez-Roisin, R. Toward a consensus definition for COPD exacerbations. *Chest* **117**, 398S-401S (2000).
209. Toy, E. L., Gallagher, K. F., Stanley, E. L., Swensen, A. R. & Duh, M. S. The Economic Impact of Exacerbations of Chronic Obstructive Pulmonary Disease and Exacerbation

Definition: A Review. *COPD J. Chronic Obstr. Pulm. Dis.* **7**, 214–228 (2010).

210. Papi, A. *et al.* Infections and Airway Inflammation in Chronic Obstructive Pulmonary Disease Severe Exacerbations. *Am. J. Respir. Crit. Care Med.* **173**, 1114–1121 (2006).

211. Santos, S. *et al.* Treatment of patients with COPD and recurrent exacerbations: the role of infection and inflammation. *Int. J. Chron. Obstruct. Pulmon. Dis.* **11**, 515–525 (2016).

212. Agustí, A. *et al.* Persistent systemic inflammation is associated with poor clinical outcomes in copd: A novel phenotype. *PLoS One* **7**, e37483 (2012).

213. Albert, R. K. *et al.* Azithromycin for Prevention of Exacerbations of COPD. *N. Engl. J. Med.* **365**, 689–698 (2011).

214. Martinez, F. J. *et al.* Effect of roflumilast on exacerbations in patients with severe chronic obstructive pulmonary disease uncontrolled by combination therapy (REACT): a multicentre randomised controlled trial. *Lancet* **385**, 857–866 (2015).

215. Wedzicha, J. A. *et al.* Indacaterol–Glycopyrronium versus Salmeterol–Fluticasone for COPD. *N. Engl. J. Med.* **374**, 2222–2234 (2016).

216. Cane, J. L. *et al.* Matrix metalloproteinases -8 and -9 in the Airways, Blood and Urine During Exacerbations of COPD. *COPD J. Chronic Obstr. Pulm. Dis.* **13**, 26–34 (2016).

217. Thomsen, M. *et al.* Inflammatory Biomarkers and Exacerbations in Chronic Obstructive Pulmonary Disease. *JAMA* **309**, 2353–2361 (2013).

218. Vaitkus, M. *et al.* Reactive Oxygen Species in Peripheral Blood and Sputum Neutrophils During Bacterial and Nonbacterial Acute Exacerbation of Chronic Obstructive Pulmonary Disease. *Inflammation* **36**, 1485–1493 (2013).

219. Oba, Y. & Lone, N. A. Efficacy and safety of roflumilast in patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis. *Ther. Adv. Respir. Dis.* **7**, 13–24 (2013).

220. Chang, C. *et al.* Dynamics of inflammation resolution and symptom recovery during AECOPD treatment. *Sci. Rep.* **4**, 5516 (2014).

221. Krommidas, G. *et al.* Plasma leptin and adiponectin in COPD exacerbations: Associations with inflammatory biomarkers. *Respir. Med.* **104**, 40–46 (2010).

222. Müller, B. & Tamm, M. Biomarkers in acute exacerbation of chronic obstructive pulmonary disease: among the blind, the one-eyed is king. *Am. J. Respir. Crit. Care Med.* **174**, 848–849 (2006).

223. Karadag, F., Karul, A. B., Cildag, O., Yilmaz, M. & Ozcan, H. Biomarkers of Systemic Inflammation in Stable and Exacerbation Phases of COPD. *Lung* **186**, 403–409 (2008).

224. Cosio, M. G., Saetta, M. & Agustí, A. Immunologic Aspects of Chronic Obstructive Pulmonary Disease. *N. Engl. J. Med.* **360**, 2445–2454 (2009).

225. Curtis, J. L., Freeman, C. M. & Hogg, J. C. The Immunopathogenesis of Chronic Obstructive Pulmonary Disease: Insights from Recent Research. *Proc. Am. Thorac. Soc.* **4**, 512–521 (2007).

226. McCubbrey, A. L., Sonstein, J., Ames, T. M., Freeman, C. M. & Curtis, J. L. Glucocorticoids relieve collectin-driven suppression of apoptotic cell uptake in murine alveolar macrophages through downregulation of SIRPα. *J. Immunol.* **189**, 112–119 (2012).

227. Wold, S., Johansson, E. & Cocchi, M. PLS-partial least squares projections to latent structures. in *3D QSAR in Drug Design: Volume 1: Theory Methods and Applications* (ed. Kubinyi, H.) 523–550 (Kluwer Academic Publishers, 1993).

228. Witkowska, A. M. & Borawska, M. H. Soluble intercellular adhesion molecule-1

(sICAM-1): an overview. *Eur. Cytokine Netw.* **15**, 91–98 (2004).

229. Agouridakis, P. *et al.* The predictive role of serum and bronchoalveolar lavage cytokines and adhesion molecules for acute respiratory distress syndrome development and outcome. *Respir. Res.* **3**, 25 (2002).

230. Leeuwenberg, J. F. *et al.* E-selectin and intercellular adhesion molecule-1 are released by activated human endothelial cells in vitro. *Immunology* **77**, 543–549 (1992).

231. Yung, S. C. & Farber, J. M. Chemokines. in *Handbook of Biologically Active Peptides* (ed. Kastin, A. J.) 656–663 (Elsevier, 2013). doi:10.1016/B978-0-12-385095-9.00089-0

232. Hunter, C. A. & Jones, S. A. IL-6 as a keystone cytokine in health and disease. *Nat. Immunol.* **16**, 448–457 (2015).

233. Dinarello, C. A. Interleukin-1 in the pathogenesis and treatment of inflammatory diseases. *Blood* **117**, 3720–3732 (2011).

234. Douni, E. & Kollias, G. A critical role of the p75 tumor necrosis factor receptor (p75TNF-R) in organ inflammation independent of TNF, lymphotoxin alpha, or the p55TNF-R. *J. Exp. Med.* **188**, 1343–52 (1998).

235. Freeman, C. M. *et al.* Lung Dendritic Cell Expression of Maturation Molecules Increases with Worsening Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Crit. Care Med.* **180**, 1179–1188 (2009).

236. Freeman, C. M. & Curtis, J. L. Lung Dendritic Cells: Shaping Immune Responses Throughout COPD Progression. *Am. J. Respir. Cell Mol. Biol.* **56**, 152–159 (2017).

237. Andelid, K. *et al.* Systemic signs of neutrophil mobilization during clinically stable periods and during exacerbations in smokers with obstructive pulmonary disease. *Int. J. Chron. Obstruct. Pulmon. Dis.* **10**, 1253–1263 (2015).

238. Barnes, P. J. The Cytokine Network in Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Cell Mol. Biol.* **41**, 631–638 (2009).

239. Chen, Y.-W. R., Leung, J. M. & Sin, D. D. A Systematic Review of Diagnostic Biomarkers of COPD Exacerbation. *PLoS One* **11**, e0158843 (2016).

240. Mishra, A. *et al.* A Critical Role for P2X $_7$ Receptor–Induced VCAM-1 Shedding and Neutrophil Infiltration during Acute Lung Injury. *J. Immunol.* **197**, 2828–2837 (2016).

241. El-Deek, S. E., Makhlouf, H. A., Saleem, T. H., Mandour, M. A. & Mohamed, N. A. Surfactant Protein D, Soluble Intercellular Adhesion Molecule-1 and High-Sensitivity C-Reactive Protein as Biomarkers of Chronic Obstructive Pulmonary Disease. *Med. Princ. Pract.* **22**, 469–474 (2013).

242. Hollander, C., Sitkauskiene, B., Sakalauskas, R., Westin, U. & Janciauskiene, S. M. Serum and bronchial lavage fluid concentrations of IL-8, SLPI, sCD14 and sICAM-1 in patients with COPD and asthma. *Respir. Med.* **101**, 1947–1953 (2007).

243. Gerritsen, W. B. M., Asin, J., Zanen, P., van den Bosch, J. M. M. & Haas, F. J. L. M. Markers of inflammation and oxidative stress in exacerbated chronic obstructive pulmonary disease patients. *Respir. Med.* **99**, 84–90 (2005).

244. Aaron, C. P. *et al.* Intercellular adhesion molecule 1 and progression of percent emphysema: The MESA Lung Study. *Respir. Med.* **109**, 255–264 (2015).

245. Kwiatkowska, S., Noweta, K., Zieba, M., Nowak, D. & Bialasiewicz, P. Enhanced exhalation of matrix metalloproteinase-9 and tissue inhibitor of metalloproteinase-1 in patients with COPD exacerbation: a prospective study. *Respiration* **84**, 231–241 (2012).

246. Chakrabarti, S. & Patel, K. D. Regulation of matrix metalloproteinase-9 release from IL-8-stimulated human neutrophils. *J. Leukoc. Biol.* **78**, 279–288 (2005).

247. Röpcke, S. *et al.* Repeatability of and Relationship between Potential COPD Biomarkers in Bronchoalveolar Lavage, Bronchial Biopsies, Serum, and Induced Sputum. *PLoS One* **7**, e46207 (2012).

248. Masciantonio, M. G., Lee, C. K. S., Arpino, V., Mehta, S. & Gill, S. E. The Balance Between Metalloproteinases and TIMPs. in *Progress in molecular biology and translational science* **147**, 101–131 (2017).

249. Navratilova, Z., Kolek, V. & Petrek, M. Matrix Metalloproteinases and Their Inhibitors in Chronic Obstructive Pulmonary Disease. *Arch. Immunol. Ther. Exp. (Warsz).* **64**, 177–193 (2016).

250. Dentener, M. A. *et al.* Systemic anti-inflammatory mediators in COPD: increase in soluble interleukin 1 receptor II during treatment of exacerbations. *Thorax* **56**, 721–726 (2001).

251. Groenewegen, K. H., Dentener, M. A. & Wouters, E. F. M. Longitudinal follow-up of systemic inflammation after acute exacerbations of COPD. *Respir. Med.* **101**, 2409–2415 (2007).

252. Peters, V. A., Joesting, J. J. & Freund, G. G. IL-1 receptor 2 (IL-1R2) and its role in immune regulation. *Brain. Behav. Immun.* **32**, 1–8 (2013).

253. Couper, D. *et al.* Design of the Subpopulations and Intermediate Outcomes in COPD Study (SPIROMICS). *Thorax* **69**, 492–495 (2014).

254. Wells, J. M. *et al.* Safety and Tolerability of Comprehensive Research Bronchoscopy in Chronic Obstructive Pulmonary Disease. Results from the SPIROMICS Bronchoscopy Substudy. *Ann. Am. Thorac. Soc.* **16**, 439–446 (2019).

255. Gonçalves, I. *et al.* Clinical and molecular markers in COPD. *Pulmonology* **24**, 250–259 (2018).

256. Brusselle, G. *et al.* Blood eosinophil levels as a biomarker in COPD. *Respiratory Medicine* **138**, 21–31 (2018).

257. Rennard, S. I. & Drummond, M. B. Early chronic obstructive pulmonary disease: Definition, assessment, and prevention. *The Lancet* **385**, 1778–1788 (2015).

258. Siafakas, N., Bizymi, N., Mathioudakis, A. & Corlateanu, A. EARLY versus MILD Chronic Obstructive Pulmonary Disease (COPD). *Respiratory Medicine* **140**, 127–131 (2018).

259. Siafakas, N., Corlateanu, A. & Fouka, E. Phenotyping Before Starting Treatment in COPD? *COPD J. Chronic Obstr. Pulm. Dis.* **14**, 367–374 (2017).

260. Barnes, P. J. Inflammatory endotypes in COPD. *Allergy* **74**, 1249–1256 (2019).

261. Woodruff, P. G. *et al.* Clinical Significance of Symptoms in Smokers with Preserved Pulmonary Function. *N. Engl. J. Med.* **374**, 1811–1821 (2016).

262. Martinez, F. J. *et al.* At the root: Defining and halting progression of early chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine* **197**, 1540–1551 (2018).

263. Vestbo, J. *et al.* Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE). *Eur. Respir. J.* **31**, 869–873 (2008).

264. Regan, E. A. *et al.* Genetic epidemiology of COPD (COPDGene) study design. *COPD J. Chronic Obstr. Pulm. Dis.* **7**, 32–43 (2010).

265. Eapen, M. S. *et al.* Abnormal M1/M2 macrophage phenotype profiles in the small airway wall and lumen in smokers and chronic obstructive pulmonary disease (COPD). *Sci. Rep.* **7**, 13392 (2017).

266. Wen, Y. *et al.* Assessment of airway inflammation using sputum, BAL, and endobronchial

biopsies in current and ex-smokers with established COPD. *Int. J. Chron. Obstruct. Pulmon. Dis.* **5**, 327–334 (2010).

267. Polverino, F. *et al.* A disintegrin and metalloproteinase domain-8: A novel protective proteinase in chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* **198**, 1254–1267 (2018).

268. Papakonstantinou, E. *et al.* Acute exacerbations of COPD are associated with significant activation of matrix metalloproteinase 9 irrespectively of airway obstruction, emphysema and infection. *Respir. Res.* **16**, 78 (2015).

269. Kunz, L. I. Z. *et al.* Smoking status and anti-inflammatory macrophages in bronchoalveolar lavage and induced sputum in COPD. *Respir. Res.* **12**, 34 (2011).

270. Moré, J. M. *et al.* Smoking reduces surfactant protein D and phospholipids in patients with and without chronic obstructive pulmonary disease. *BMC Pulm. Med.* **10**, 53 (2010).

271. Blidberg, K. *et al.* Adhesion molecules in subjects with COPD and healthy non-smokers: A cross sectional parallel group study. *Respir. Res.* **14**, 47 (2013).

272. Drost, E. M. *et al.* Oxidative stress and airway inflammation in severe exacerbations of COPD. *Thorax* **60**, 293–300 (2005).

273. Che, K. F. *et al.* The neutrophil-mobilizing cytokine interleukin-26 in the airways of long-term tobacco smokers. *Clin. Sci.* **132**, 959–983 (2018).

274. Halper-Stromberg, E. *et al.* Bronchoalveolar lavage fluid from copd patients reveals more compounds associated with disease than matched plasma. *Metabolites* **9**, 157 (2019).

275. Pastor, M. D. *et al.* Identification of proteomic signatures associated with lung cancer and COPD. *J. Proteomics* **89**, 227–237 (2013).

276. Pastor, M. D. *et al.* Identification of oxidative stress related proteins as biomarkers for lung cancer and chronic obstructive pulmonary disease in bronchoalveolar lavage. *Int. J. Mol. Sci.* **14**, 3440–3455 (2013).

277. Plymoth, A. *et al.* Rapid proteome analysis of bronchoalveolar lavage samples of lifelong smokers and never-smokers by micro-scale liquid chromatography and mass spectrometry. *Clin. Chem.* **52**, 671–679 (2006).

278. Plymoth, A. *et al.* Protein expression patterns associated with progression of chronic obstructive pulmonary disease in bronchoalveolar lavage of smokers. *Clin. Chem.* **53**, 636–644 (2007).

279. Yang, M. *et al.* Long-term smoking alters abundance of over half of the proteome in bronchoalveolar lavage cell in smokers with normal spirometry, with effects on molecular pathways associated with COPD. *Respir. Res.* **19**, 40 (2018).

280. Tkacova, R., McWililams, A., Lam, S. & Sin, D. D. Integrating Lung and Plasma Expression of Pneumo-Proteins in Developing Biomarkers in COPD: A Case Study of Surfactant Protein D. *Med. Sci. Monit.* **16**, CR540-4 (2010).

281. Gasiuniene, E., Lavinskiene, S., Sakalauskas, R. & Sitkauskiene, B. Levels of IL-32 in Serum, Induced Sputum Supernatant, and Bronchial Lavage Fluid of Patients with Chronic Obstructive Pulmonary Disease. *COPD J. Chronic Obstr. Pulm. Dis.* **13**, 569–575 (2016).

282. Sauleda, J. *et al.* Pulmonary and systemic hepatocyte and keratinocyte growth factors in patients with chronic obstructive pulmonary disease. *Int. J. COPD* **3**, 719–725 (2008).

283. Létuvé, S. *et al.* YKL-40 Is Elevated in Patients with Chronic Obstructive Pulmonary Disease and Activates Alveolar Macrophages. *J. Immunol.* **181**, 5167–5173 (2008).

284. Um, S. J., Lam, S., Coxson, H., Man, S. F. P. & Sin, D. D. Budesonide/formoterol

enhances the expression of pro surfactant protein-B in lungs of COPD Patients. *PLoS One* **8**, e83881 (2013).

285. Roche, N. Stable COPD Treatment: Where are We? *COPD J. Chronic Obstr. Pulm. Dis.* **15**, 123–129 (2018).

286. Kawayama, T. *et al.* Responsiveness of blood and sputum inflammatory cells in Japanese COPD patients, non-COPD smoking controls, and non-COPD nonsmoking controls. *Int. J. COPD* **11**, 295–303 (2016).

287. Bowler, R. P. *et al.* Integrative omics approach identifies interleukin-16 as a biomarker of emphysema. *Omi. A J. Integr. Biol.* **17**, 619–626 (2013).

288. Baralla, A. *et al.* Plasma Proteomic Signatures in Early Chronic Obstructive Pulmonary Disease. *PROTEOMICS - Clin. Appl.* **12**, e1700088 (2018).

289. Freeman, C. M. *et al.* Design of a multi-center immunophenotyping analysis of peripheral blood, sputum and bronchoalveolar lavage fluid in the Subpopulations and Intermediate Outcome Measures in COPD Study (SPIROMICS). *J. Transl. Med.* **13**, 19 (2015).

290. Akdis, M. *et al.* Interleukins (from IL-1 to IL-38), interferons, transforming growth factor β, and TNF-α: Receptors, functions, and roles in diseases. *J. Allergy Clin. Immunol.* **138**, P984-1010 (2016).

291. Martinez, F. J. Idiopathic interstitial pneumonias: Usual interstitial pneumonia versus nonspecific interstitial pneumonia. *Proceedings of the American Thoracic Society* **3**, 81–95 (2006).

292. Rennard, S. I. *et al.* Estimation of volume of epithelial lining fluid recovered by lavage using urea as marker of dilution. *J. Appl. Physiol.* **60**, 532–538 (1986).

293. Willemse, B. W. M. *et al.* Effect of 1-year smoking cessation on airway inflammation in COPD and asymptomatic smokers. *Eur. Respir. J.* **26**, 835–845 (2005).

294. Brandsma, C.-A., Van den Berge, M., Hackett, T.-L., Brusselle, G. & Timens, W. Recent advances in chronic obstructive pulmonary disease pathogenesis: from disease mechanisms to precision medicine. *Journal of Pathology* **250**, 624–635 (2020).

295. Wedzicha, J. A., Singh, R. & Mackay, A. J. Acute COPD exacerbations. *Clinics in Chest Medicine* **35**, 157–163 (2014).

296. Tangedal, S. *et al.* Sputum microbiota and inflammation at stable state and during exacerbations in a cohort of chronic obstructive pulmonary disease (COPD) patients. *PLoS One* **14**, e0222449 (2019).

297. Saraswat, M. *et al.* Label-free plasma proteomics identifies haptoglobin-related protein as candidate marker of idiopathic pulmonary fibrosis and dysregulation of complement and oxidative pathways. *Sci. Rep.* **10**, 1–11 (2020).

298. Miller, R. D., Kueppers, F. & Offord, K. P. Serum concentrations of C3 and C4 of the complement system in patients with chronic obstructive pulmonary disease. *J. Lab. Clin. Med.* **95**, 266–271 (1980).

299. Chauhan, S., Gupta, M., Goyal, A. & Dasgupta, D. Alterations in immunoglobulin & complement levels in chronic obstructive pulmonary disease. *Indian J. Med. Res.* **92**, 241–245 (1990).

300. Sun, W. *et al.* Common Genetic Polymorphisms Influence Blood Biomarker Measurements in COPD. *PLOS Genet.* **12**, e1006011 (2016).

301. Raffield, L. M. *et al.* Comparison of Proteomic Assessment Methods in Multiple Cohort Studies. *Proteomics* **20**, e1900278 (2020).

302. SomaLogic, I. *Correlation of SOMAmer® Reagents in the SOMAscan® Assay and*

Commercially Available Immunoassays. (2016).

303. Ohtani, Y. *et al.* Chronic summer-type hypersensitivity pneumonitis initially misdiagnosed as idiopathic interstitial pneumonia. *Intern. Med.* **47**, 857–862 (2008).
304. O'Brien, E. C. *et al.* Rationale for and design of the idiopathic pulmonary fibrosis–PRospective outcomes (IPF-PRO) registry. *BMJ Open Respir. Res.* **3**, e000108 (2016).
305. Moore, B. B. *et al.* Animal Models of Fibrotic Lung Disease. *Am. J. Respir. Cell Mol. Biol.* **49**, 167–179 (2013).
306. Degryse, A. L. *et al.* Repetitive intratracheal bleomycin models several features of idiopathic pulmonary fibrosis. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **299**, L442-452 (2010).
307. Pérez-Rial, S., Girón-Martínez, Á. & Peces-Barba, G. Animal models of chronic obstructive pulmonary disease. *Archivos de Bronconeumologia* **51**, 121–127 (2015).
308. Gaschler, G. J. *et al.* Bacteria challenge in smoke-exposed mice exacerbates inflammation and skews the inflammatory profile. *Am. J. Respir. Crit. Care Med.* **179**, 666–675 (2009).
309. Nie, Y.-C. *et al.* Characteristic comparison of three rat models induced by cigarette smoke or combined with LPS: To establish a suitable model for study of airway mucus hypersecretion in chronic obstructive pulmonary disease. *Pulm. Pharmacol. Ther.* **25**, 349–356 (2012).
310. Kang, M.-J. *et al.* Cigarette smoke selectively enhances viral PAMP- and virus-induced pulmonary innate immune and remodeling responses in mice. *J. Clin. Invest.* **118**, 2771–2784 (2008).
311. Strieter, R. M. & Mehrad, B. New mechanisms of pulmonary fibrosis. *Chest* **136**, 1364–1370 (2009).
312. King, T. E. *et al.* Effect of interferon gamma-1b on survival in patients with idiopathic pulmonary fibrosis (INSPIRE): a multicentre, randomised, placebo-controlled trial. *Lancet* **374**, 222–228 (2009).
313. Argelaguet, R. *et al.* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
314. Majewski, S. *et al.* Epithelial alarmin levels in exhaled breath condensate in patients with idiopathic pulmonary fibrosis: A pilot study. *Clin. Respir. J.* **13**, 652–656 (2019).
315. Hao, W., Li, M., Zhang, C., Zhang, Y. & Wang, P. Inflammatory mediators in exhaled breath condensate and peripheral blood of healthy donors and stable COPD patients. *Immunopharmacol. Immunotoxicol.* **41**, 224–230 (2019).
316. Corhay, J. L. *et al.* Increased of exhaled breath condensate neutrophil chemotaxis in acute exacerbation of COPD. *Respir. Res.* **15**, 115 (2014).
317. Hayton, C. *et al.* Breath biomarkers in idiopathic pulmonary fibrosis: A systematic review. *Respiratory Research* **20**, 7 (2019).
318. Yue, M. *et al.* Measurement of Short-chain Fatty Acids in Respiratory Samples. *Am. J. Respir. Crit. Care Med.* (2020). doi:10.1164/rccm.201909-1840le
319. Erb-Downward, J. R. *et al.* Critical Relevance of Stochastic Effects on Low-Bacterial-Biomass 16S rRNA Gene Analysis. *MBio* **11**, e00258-20 (2020).
320. Magnini, D. *et al.* Idiopathic Pulmonary Fibrosis: Molecular Endotypes of Fibrosis Stratifying Existing and Emerging Therapies. *Respiration* **93**, 379–395 (2017).
321. Kazerooni, E. A. *et al.* Thin-section CT obtained at 10-mm increments versus limited three-level thin-section CT for idiopathic pulmonary fibrosis: Correlation with pathologic scoring. *Am. J. Roentgenol.* **169**, 977–983 (1997).

322. Saito, F. *et al.* Role of interleukin-6 in bleomycin-induced lung inflammatory changes in mice. *Am. J. Respir. Cell Mol. Biol.* **38**, 566–571 (2008).

323. Matsuo, R. *et al.* The Inhibition of N-Glycosylation of Glycoprotein 130 Molecule Abolishes STAT3 Activation by IL-6 Family Cytokines in Cultured Cardiac Myocytes. *PLoS One* **9**, e111097 (2014).

324. Lawson, W. E. *et al.* Endoplasmic reticulum stress enhances fibrotic remodeling in the lungs. *Proc. Natl. Acad. Sci.* **108**, 10562–10567 (2011).

325. Yoshida, K. *et al.* Targeted disruption of gp130, a common signal transducer for the interleukin 6 family of cytokines, leads to myocardial and hematological disorders. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 407–11 (1996).