SUPPLEMENTAL METHODS AND DATA

Discovery of Distinct Immune Phenotypes Using Machine Learning in Pulmonary Arterial Hypertension

Andrew J. Sweatt MD [1,2], Haley K. Hedlin PhD [3], Vidhya Balasubramanian MS [3],
Andrew Hsi BS [2], Lisa K. Blum PhD [4], William H. Robinson PhD [4],
Francois Haddad MD [5,6], Peter M. Hickey MB ChB [7], Robin A. Condliffe MD [8],
Allan Lawrie PhD [7], Mark R. Nicolls MD [1,2,9], Marlene Rabinovitch MD [2,10],
Purvesh Khatri PhD [9,11], Roham T. Zamanian MD [1,2]

[1] Division of Pulmonary and Critical Care Medicine, Stanford University (Stanford, CA, USA);
[2] Vera Moulton Wall Center for Pulmonary Vascular Disease, Stanford University (Stanford, CA, USA);
[3] Quantitative Sciences Unit, Stanford University (Stanford, CA, USA);
[4] Division of Immunology and Rheumatology, Stanford University (Stanford, CA, USA);
[5] Division of Cardiovascular Medicine, Stanford University (Stanford, CA, USA);
[6] Stanford Cardiovascular Institute, Stanford University (Stanford, CA, USA);
[7] Department of Infection, Immunity, and Cardiovascular Disease, University of Sheffield (Sheffield, UK);
[8] Sheffield Pulmonary Vascular Disease Unit, Royal Hallamshire Hospital (Sheffield, UK);
[9] Institute for Immunity, Transplantation, and Infection, Stanford University (Stanford, CA, USA);
[10] Department of Pediatric Cardiology, Stanford University (Stanford, CA, USA);
[11] Division of Biomedical Informatics Research, Stanford University (Stanford, CA, USA)

CONTENTS

SUPPLEMENTAL METHODS

Sample collection, processing, and biorepository storage

Discovery cohort: Stanford University

Pulmonary arterial hypertension (PAH). PAH plasma samples (n=281) were obtained from the Stanford Pulmonary Hypertension Biobank (SPHB) (Stanford University, USA), which is a comprehensive tissue bank that includes plasma, serum, peripheral blood mononuclear cells, exhaled breath condensate, and urine from patients with all forms of pulmonary hypertension (Stanford University IRB #14083). Initiated in 2007, the SPHB has captured samples from over 600 well-characterized subjects who were recruited at the time of evaluation in Stanford pulmonary hypertension clinic. Each sample is linked to comprehensive clinical data captured in the Vera Moulton Wall Center Pulmonary Hypertension database (see below). After informed consent was obtained, study plasma samples were collected from patients during right heart catheterization in the fasting state. Peripheral blood was drawn from the antecubital fossa into EDTA vacutainers under standard sterile precautions. Collection tubes were immediately placed upright into a rack at room temperature. Within 30 minutes, the sample was inverted several times to mix components, and subsequently centrifuged at 1300 rpm for 10 minutes. The plasma layer was then carefully removed by pipette without disturbing the buffy coat, and transferred to Eppendorf tubes in 200 µL aliquots. Aliquoted samples were secured and stored upright at -80°C in the SPHB.

Healthy controls. Control plasma samples (n=88) were acquired from the Stanford Cardiovascular Institute Biomarker and Phenotype Core Laboratory biorepository (Stanford University IRB #40869). Samples were collected from healthy volunteers between 2009 and 2013, as part of the Stanford Healthy Aging Study (Stanford University IRB #20942). Peripheral venous blood was drawn in the fasting state from the antecubital fossa, and subsequently processed and stored utilizing the same protocol and conditions that were applied for PAH samples (see above). To establish health in these individuals, initial screening involved a comprehensive questionnaire and clinical assessment with blood pressure and anthropometric measurements. Subjects were excluded for any unexplained chronic cardiopulmonary symptoms (dyspnea, cough, angina, palpitations, orthopnea, lightheadedness), symptomatic atherosclerotic disease, symptomatic heart failure, history of atrial fibrillation, hypertension, chronic pulmonary disease, diabetes mellitus requiring therapy, current smoking, hyperlipidemia requiring therapy, body mass index ≥35 kg/m^2, history of atopy, chronic systemic inflammatory disease, recent infectious disease, prior or current malignancy, and Alzheimer's disease. Subjects had basic laboratory studies to rule out dyslipidemia, elevated NT-proBNP (≥ 300 pg/mL), abnormal liver function tests, and renal dysfunction (stage 3 or greater). The participants were then screened with vascular ultrasound for abdominal aorta aneurysm (≥5 cm) and asymptomatic carotid or femoral atherosclerosis (≥20% diameter stenosis). Transthoracic echocardiography was also employed to screen and exclude those with subclinical left ventricular dysfunction (ejection fraction ≤50%), subclinical valve disease (categorized as mild or greater), any right ventricular enlargement or dysfunction, or estimated right ventricular systolic pressure ≥30 mmHg.

Validation cohort: University of Sheffield

PAH. Plasma samples (n=104) were obtained from the Sheffield Pulmonary Vascular Disease Unit (SPVDU) pulmonary hypertension biobank (Sheffield Teaching Hospitals Foundation trust, University of Sheffield, UK). A single-center prospective observational study, the SPVDU pulmonary hypertension biobank (STH15222) includes samples from more than 600 adult subjects who were recruited with a suspected diagnosis of pulmonary hypertension. The validation cohort in our study included consecutive patients who had samples collected for the SPVDU biobank between 2009-2012. All patients met hemodynamic diagnostic criteria for PAH, which included mean pulmonary arterial pressure ≥25 mmHg, pulmonary arterial wedge pressure ≤15 mmHg, and pulmonary vascular resistance >240 dynes/sec/cm^{-5}. Following consent, blood was drawn from a peripheral upper extremity vein during right heart catheterization into EDTA vacutainers, mixed thoroughly by inverting several times, and placed onto ice prior to centrifugation (at 1500 rpm for 15 minutes at 4°C). The plasma layer was then carefully removed, and banked in 250 µl aliquots at -80°C. Each sample in the Sheffield HTA-approved biorepository was linked to the patient's electronic phenotype record in a bespoke MySQL database, which was in turn linked by patient ID to clinical data in hospital electronic records. All samples were collected, stored and assayed with appropriate informed consent and approval from North Sheffield Research Ethics Committee (08/H1308/193+5) via the Sheffield NIHR Clinical Research Facility.

Clinical data collection

Discovery cohort: Stanford University

Demographic, clinical, hemodynamic, and outcomes data was obtained from the Vera Moulton Wall Center (VMWC) Pulmonary Hypertension Database. Established in 2000, the VMWC database has enrolled consecutive patients evaluated at Stanford University with hemodynamically confirmed PAH (Stanford University IRB#12338). This observational relational database captures 371 unique variables in a longitudinal manner.

Clinical data was extracted for analysis at the time of proteomic sample collection. Data was omitted if not available within one month of plasma sampling. Demographic variables included age, sex, and race. PAH etiologies were categorized according to the World Health Organization classification scheme.[1] For each subject, we collected the dates of symptom onset (patient reported) and PAH diagnosis (including diagnoses that preceded referral to Stanford). We obtained right heart catheterization hemodynamic data and non-invasive disease markers (New York Heart Association functional class, six-minute walk distance, NT-proBNP, diffusion capacity of the lung for carbon monoxide, echocardiographic variables, and serum creatinine). Extracted variables were used to calculate a composite REVEAL risk score,[2] and patients missing variables necessary for calculation had data imputed from cohort medians (no subject had more than 2 variables missing). In addition, we collected background PAH therapies (phosphodiesterase-5 inhibitors, endothelin receptor antagonists, prostanoids, guanylate cyclase stimulators) and adjunct agents (loop diuretics, aldosterone antagonists, anticoagulants, digoxin, calcium channel blockers).

Certain clinical variables included in our analysis were not available in the VMWC pulmonary hypertension database, including comorbidities and medications with potential cytokine relevance. These variables were manually extracted from electronic patient medical records. Comorbidities included coronary artery disease (prior stent placement or coronary artery bypass grafting), cerebrovascular accident (prior ischemic stroke), essential hypertension (requiring active anti-hypertensive therapy), chronic kidney disease (stage 3 or greater, glomerular filtration rate <60 ml/min/1.73 m^3), diabetes mellitus (requiring active oral glucose lowering therapy or insulin), hyperlipidemia (requiring active lipid-lowering therapy), history of thyroiditis (Graves' disease or Hashimoto's), active tobacco abuse (patient reported), inflammatory bowel disease (requiring active topical or systemic therapy), specific connective tissue disease diagnoses (systemic sclerosis, systemic lupus erythematosus, mixed connective tissue disease, rheumatoid arthritis, sjogren syndrome, dermatomyositis/polymyositis, psoriatic arthritis), obstructive sleep apnea (formal diagnosis by polysomnography with apnea-hypopnea index >5 events/hour), chronic obstructive pulmonary disease (compatible clinical history and spirometry with FEV1/FVC <0.70), mild radiographic interstitial lung disease (reticulation, chronic or recurrent ground glass opacities, absence of honeycombing, relatively preserved lung function with total lung capacity >60% of predicted), and asthma (requiring at least rescue inhaler ± maintenance therapy). These pulmonary comorbidities were deemed to not be significant contributors to pulmonary hypertension, given that treating physicians established a

Group 1 PAH diagnosis. Additional background medications were collected including immune modulating agents (prednisone, mycophenolate mofetil, hydroxychloroquine, leflunomide, azathioprine, tacrolimus, bortezomib, anakinra, rituximab, or cyclophosphamide), non-steroidal anti-inflammatory drugs (aspirin or other NSAIDs), and allergy agents (H1 blockers, leukotriene antagonists).

The VMWC database and electronic medical records were also used to identify patients who died from any cause or underwent lung/heart-lung transplantation prior to database lock (August 2016).

Validation cohort: University of Sheffield

Clinical data were obtained from a bespoke SQL database, which links University of Sheffield hospital records and survival data from National Health Service (NHS) Digital. Demographic data, PAH etiologies, and non-invasive PAH clinical markers (WHO functional class, incremental shuttle walk test, diffusion capacity of the lung for carbon monoxide) were available for nearly all subjects within a week of plasma collection, as it had been obtained during the index clinical encounter at the University of Sheffield. Moreover, hemodynamic measurements were acquired on the day of plasma collection. Long-term survival data was even available for patients lost to follow-up at the University of Sheffield, since NHS Digital updates were incorporated into the Sheffield pulmonary hypertension database in real-time. Clinical data was collected with appropriate informed consent and approval from North Sheffield Research Ethics Committee (08/H1308/193+5) via the Sheffield NIHR Clinical Research Facility.

Proteomic immunoassay measurements

Overview of Bio-Plex® multiplex immunoassay platform. The Bio-Plex® multiplex immunoassay is a magnetic bead-based flow cytometric platform that is built on Luminex® xMAP™ technology. A magnetic bead mixture provides the substrate for numerous parallel sandwich immunoassays within each sample. Individual beads are each coupled to an analyte-specific capture antibody and have a unique fluorescent spectral address. After a protein of interest is captured in sample, labeled detection antibody binds a different epitope to sandwich the analyte. Beads are flown single-file through a dual-laser instrument, which (i) identifies an analyte according to the spectral address of the bead and (ii) quantifies bound analyte by measuring fluorescence intensity from the detection antibody reporter label. Hundreds of beads specific to each measured protein are in each given sample, thus median fluorescence intensity (MFI) is determined from single-bead measurements to quantify abundance of a protein. The MFI of reporter signal is proportional to the concentration of an analyte in sample.

Study protocol for Bio-Plex® immunoassay. Prior to setting up our assays, the Luminex 200™ plate reader instrument was calibrated according to manufacturer instructions. This process employed Bio-Plex calibration beads to standardize fluorescent signal detection. To prepare experimental samples, frozen biobanked plasma aliquots were passively thawed to room temperature and diluted four-fold in assay buffer (1 volume plasma to 3 volumes buffer). These plasma samples were assayed within 30 minutes of reaching room temperature, as described below. To prepare a magnetic capture bead mixture, bead stock solution (20x) was vortexed at medium speed for 30 seconds and then diluted 20-fold in assay buffer. The preparation of standards involved first adding 500μL of standard diluent to each stock vial of lyophilized standard, which contained known concentrations of analytes measured by our assay. The reconstituted standard was vortexed and incubated on ice for 30 minutes. Thereafter, we performed four-fold serial dilution to prepare a series of eight total standards.

After preparation of the samples, capture bead mixture, and standards, the immunoassay was carried out on a 96-well plate. First, we added 75 μL of standard to eight wells (one standard dilution per well), 75 μL of assay buffer to one well (‘blank’ well without sample later used to measure background fluorescence from non-specific binding), and a 75 μL pre-diluted experimental sample to the remaining wells. Next, we added 25 μL of capture bead mixture to all plate wells. The plate was sealed, placed on a shaker for two hours (800 rpm), and incubated at 4°C overnight. After passive re-warming the next day, solution was removed and the plate was washed by magnetic separation with the Bio-Plex Pro™ wash station (200 μL buffer, 3 cycles). Biotinylated detection antibody stock solution (10x) was diluted 10-fold in assay buffer and added to each well (25 μL), followed by two-hour incubation on a shaker (800 rpm), solution removal, and three magnetic separation washes. Next, streptavidin-phycoerythrin stock solution (100x) was diluted 100-fold, incubated in each well (50 μL) for 40 minutes on a shaker (500 rpm), removed, and the plate was washed again. Finally, after addition of reading buffer (100 μL) and ten-minute incubation on a shaker (800 rpm), the plate was read by a Luminex 200™ dual-laser detection instrument. Data acquisition was set to a 50 bead count minimum per analyte per well. Data was processed and presented with Bio-plex Manager™ software.

Adjustment for non-biological variation with empirical Bayes methodology

Discovery cohort PAH samples were run across nine assay plates in two experimental batches, which were balanced with respect to patient demographics and underlying WHO Group 1 subtypes. Principal components analysis (PCA) was applied to visualize non-biological variation due to batch effects. PCA is a dimensionality reduction technique that allows visualization of high-dimensional data, as numerous measured variables are represented by one point in space for each sample. We then implemented empirical Bayes methodology to adjust for plate and batch effects (‘ComBat’ parametric algorithm, ‘sva’ R software package).[3, 4] This batch correction approach is robust to outliers, well-validated, and used widely for high-dimensional molecular datasets [5]. After ‘ComBat’ adjustment, PCA was repeated to confirm improved proteomic homogeneity across plates and batches. Validation cohort samples were assayed in a separate batch from the discovery cohort runs, thus the ‘ComBat’ algorithm was also applied to correct for non-biological variation between cohorts. The impact of these adjustments on batch effects is shown in **Figure S1**.

Discovery of proteomic-based PAH subgroups with unsupervised machine learning

Overview of unsupervised consensus clustering. Consensus clustering is not a standalone method, but rather a resampling-based analysis approach used in conjunction with various clustering algorithms.[6] The method involves iterative runs of a specified algorithm on multiple subsamples from the original dataset. Across clustering iterations, the stability of cluster assignments is analyzed with quantitative and graphical tools. Consensus clustering is well-suited for unsupervised classification, as it offers objective criteria to determine the number of clusters and define cluster boundaries. Conversely, when clustering algorithms are applied in isolation, there may be a subjective component to defining cluster boundaries (hierarchal clustering), or the number of clusters must be known or decided *a priori (*k-means, k-medoids, self-organizing maps, and others). While many algorithms are sensitive to the algorithm search starting point within data, consensus clustering sidesteps this with random re-sampling runs.

Consensus clustering implementation. To sort our PAH patients into clusters based on circulating proteomic immune profiles, we carried out consensus clustering with the ConsensusClusterPlus R software package (version 1.38).[7] Consensus clustering was executed with 1,000 random patient resampling iterations. Patient subsamples were sized at 95% of the original cohort, and all measured proteins were resampled in each iteration. Analysis occurred over a cluster number (k) range from k=2-20. For each k assessed*,* cluster boundaries resulting in the best cluster assignment agreement across resampling runs were determined. In other words, a set of 19 possible consensus clustering schemes was generated (one per k from k=2-20). A symmetrical *consensus matrix* was also constructed for each k, which was an n x n matrix (n=281 PAH patients in our case) of pairwise *consensus index* values. Each consensus index value quantified the proportion of resampling runs that two patients clustered together (of the runs in which both were subsampled), and could range from 0 (never clustered together) to 1 (always clustered together).

Identifying the 'true' number of dataset clusters by consensus clustering. C*onsensus matrices* were leveraged to objectively evaluate the stability of cluster assignments across resampling runs for each *k*, and this provided the basis for identifying the 'true' cluster number (*k* optima) and ultimate cluster boundaries. We measured cluster consensus and determined the *k* optima by way of consensus matrix heatmaps (**Online Figure IIA**), cumulative consensus distribution functions (**Online Figure IIB**), average consensus index values by cluster (**Online Figure IIC**), and the proportion of ambiguously clustered pairs (PAC) metric (**Online Table II**). Each of these graphical and quantitative tools are further detailed in the corresponding supplemental figure and table legends.

Determining the optimal clustering algorithm and distance metric. Consensus clustering can be implemented with a variety of specific clustering algorithms and distance metrics. We therefore sought to objectively determine which algorithm-distance combination produced the most valid consensus clusters in our PAH proteomic dataset. Evaluation of external validity was not possible, as (i) clustering was unsupervised without *a priori* knowledge of group membership and (ii) the circulating proteomic immune milieu had not been previously investigated as multi-parameter biological system in PAH. We instead used internal validity statistics to appraise the consensus clusters generated by various algorithm-distance combinations. Measures of internal cluster validity do not require any prior knowledge about the data or clusters, as they are solely based on information intrinsic to the data. In the discovery cohort, our consensus clustering approach was carried out 15 times in total, as three common clustering algorithms (k-medoids, k-means, agglomerative hierarchical) were each executed with five different distance metrics (Euclidean, Canberra, Pearson correlation, Spearman correlation, and Manhattan). We then determined which algorithm-distance permutation yielded consensus clusters with the best measures of internal validity, as calculated with the 'fpc' R software package ('cluster.stats' function).[8] For certain statistics, low numerical values reflected better clustering validity– average distance within clusters, average sum of squares within clusters, G3 coefficient,[9] average within-cluster to between-cluster distance ratio, widest within-cluster gap, connectivity,[10] and the proportion of ambiguously clustered pairs.[11] In contrast, higher values indicated better cluster validity for other metrics– average distance between clusters, average silhouette width,[12] G2 coefficient,[13] Dunn index,[14] Dunn index 2,[14] normalized Pearson gamma,[14] Calinski and Harabasz index,[15, 16] and the separation index,[16] Ultimately, internal validity measures were the best for clusters resulting from the k-medoids algorithm with Euclidean distance (as detailed in **Online Table III**), and this combination defined our final proteomic-based PAH clusters. These input parameters were also used for clustering in the validation cohort. The k-medoids algorithm partitions a multi-dimensional dataset in space, by identifying boundaries that minimize the sum of distances between samples in each cluster and their respective cluster centers.[17]

Proteomic network analysis by machine learned PAH immune cluster (discovery cohort only)

To investigate our immune-relevant proteomic panel as a multivariable system and explore protein-protein relationships within each machine learned PAH cluster, we constructed undirected weighted partial correlation networks with force-directed graphical modeling. Partial correlation networks, which are also known as Gaussian graphical models and belong to a more general class of statistical models termed pairwise Markov random fields,[18, 19] have been used to infer biological relationships in several 'omics domains.[20-25] *The overall approach, which was carried out separately for each proteomic-based PAH cluster discovered by unsupervised consensus clustering, involved:* **(i)** obtaining partial correlations between all possible protein-protein pairs, **(ii)** construction of a saturated partial correlation network with force-directed graphical modeling (all pairwise protein-protein relationships retained), **(iii)** identification of proteins with high measures of network centrality, and **(iv)** sparse core immune network selection with removal of the more likely false positive 'spurious' protein-protein connections by regularization.

Obtaining partial correlations between all protein-protein pairs. A simple correlation coefficient quantifies the strength of linear relationship between two variables, but does not consider the influence of other variables on this relationship. Conversely, partial correlation measures the degree of association between two variables with the effect of other controlling variables removed. In other words, partial correlation is a measure of conditional dependence and is equivalent to how well two variables predict each other in multiple regression.

To obtain the partial correlation coefficients for all possible pairwise protein-protein interactions in our proteomic panel (48 measured proteins= 1,128 pairwise partial correlations per PAH cluster), we applied node-wise regression.[26, 27] Each pairwise partial correlation coefficient was the result of conditioning on all other measured proteins (we controlled for interactions with all covariates). Multiple regression was performed, in which y_1 is predicted from all other variables:

$$y_1 = \beta_{10} + \beta_{12}y_2 + \beta_{13}y_3 + \cdots + \varepsilon_1,$$

Followed by a similar regression model for y_2:

$$y_2 = \beta_{20} + \beta_{21}y_1 + \beta_{23}y_3 + \cdots + \varepsilon_2,$$

And then so on for $y_3, y_4, y_5 \ . \ . \ . to \ y_n$ ($n = 48$, the number of variables in our proteomic panel)

The partial correlation coefficient between y_i and y_j was proportional to the regression slope predicting y_i from y_j, or the regression slope predicting y_j from y_i:

$$\mathrm{Cor}\left(y_i, y_j \mid \boldsymbol{y}_{-(i,j)}\right) = \frac{\beta_{ij}\ \mathrm{Var}\ (\varepsilon_j)}{\mathrm{Var}\ (\varepsilon_i)} = \frac{\beta_{ji}\ \mathrm{Var}\ (\varepsilon_i)}{\mathrm{Var}\ (\varepsilon_j)}$$

We utilized these pairwise partial correlation coefficients to create a symmetric *weighted adjacency matrix* for each cluster, with measured proteins as rows and columns. Matrix cell values were the signed partial correlation coefficients of each protein-protein pair. This weighted adjacency matrix served as the basis for construction of network models.

Visualization of saturated partial correlation network with force-directed graphical modeling. Visualizing partial correlations in a network structure highlights the unique variance between variables, and provides the ability to map out multicollinearity and predictive mediation. After we obtained partial correlation coefficients, force-directed graphical modeling was employed to construct a saturated network of all 1,128 possible pairwise protein-protein interactions. Each measured immune-relevant protein was represented by a network *node*, and undirected *edges* connected nodes to indicate partial correlations between all protein pairs. Network edges were weighted according to the magnitude of each partial correlation coefficient.

To construct these force-directed network graphs for each machine learned PAH cluster, we applied the Fruchterman-Reingold algorithm [28]. The 'qgraph' R software package (version 1.43) was utilized for implementation, and the aforementioned adjacency matrices served as input.[29] This force-directed graphical modeling approach treated partial correlation networks as physical systems, in which both attractive and repulsive 'forces' acted on each node. The aim of the Fruchterman-Reingold algorithm was to find the graphical configuration that achieved physical equilibrium, such that the net force acting on each node was zero or near zero (the configuration with minimal energy). Nodes behaved like atomic charged particles that naturally repulsed each other, while network edges were spring-like forces that attracted nodes according to partial correlation weights. The Fruchterman-Reingold algorithm modeled forces on each node with the following equations (analogous to Coulomb's and Hooke's laws)[28]:

$F_r(d) = -\frac{l^2}{d}$, where F_r is the repulsive force, d is distance between nodes, and l is the optimal distance between nodes based on overall graph area

$F_a(d) = \frac{d^2}{l}$, where F_a is the attractive force

The algorithm iteratively moved nodes until reaching the configuration with minimal energy. All force vectors acting on nodes were summed to determine the direction and distance that a node moved at each iteration. The distance moved was limited to a maximum value, and this limit (termed 'temperature') decreased with successive algorithm iterations ('cooling') until no further movement occurred.

Identifying proteins with high measures of network centrality. In network analysis and force-directed graphical theory, centrality measures are used to quantify how central or 'important' nodes are within the overall structure of network interactions. To identify the central nodes in each PAH cluster network, we calculated *strength*, *closeness*, and *betweenness* centralities for all nodes. These three well-established indices, which have been used in biological networks to detect nodes that most affect network topology,[30] were defined and calculated as follows:

Strength (or degree) centrality is the sum of absolute partial correlation weights for all edge linkages of a node. Nodes involved in a large number of strong network interactions are more central according to the measure[31]:

$S_i = C_D^w(i) = \sum_j^N w_{ij}$, where i is the focal node, j represents all other nodes, N is the total number of nodes, w is the weighted adjacency matrix, and cell w_{ij} is defined by the partial correlation weight

Closeness centrality is the reciprocal of the average shortest path length between the node and all other nodes in a graph, thus a node with high closeness centrality is spatially close to other nodes in the network model[32, 33]:

$C_C^{w\alpha}(i) = \left[\sum_{j=1}^N d^{w\alpha}(i,j)\right]^{-1}$, where $d^{w\alpha}(i,j)$ is shortest path length between node i and nodes j:

$d^{w\alpha}(i,j) = \min\left(\frac{1}{(w_{ij})^\alpha} + \ldots + \frac{1}{(w_{hj})^\alpha}\right)$, where h is the intermediary nodes on paths between node i and j, and α is a positive tuning parameter set $= 1$

Betweenness centrality quantifies how often a node bridges the shortest path between nodes[32, 33]:

$C_B^{w\alpha}(i) = \frac{g_{jk}^{w\alpha}(i)}{g_{jk}^{w\alpha}}$, where $g_{jk}^{w\alpha}$ is number of weighted shortest paths (tuned at $\alpha = 1$) between two nodes, and $g_{jk}^{w\alpha}(i)$ is the number of these paths that go through node i

Centrality indices were calculated for each protein node in all four PAH cluster networks, and measures were z-score standardized relative to other nodes in the same network. Proteins were then grouped in each cluster according to their pattern of network centrality and plasma expression, as either high centrality-high expression, high centrality-low expression, high expression-low centrality, or low centrality-low expression. Proteins were considered to have high network centrality if at least 2 of 3 centrality indices were above the cluster mean (z-score >0). High plasma expression was defined as a mean cluster level greater than the overall PAH cohort.

Selection of sparse core immune networks by regularization. Estimating parameters from data is always associated with some degree of noise, thus partial correlation coefficients are never exactly zero even if two variables are conditionally independent. In the context of a weighted partial correlation network, these 'spurious' or 'false positive' partial correlations would be represented by weak edges in a saturated model. To reduce the number of false positive proteomic relationships and construct more interpretable force-directed network models for each PAH cluster (which we termed 'sparse core networks'), graphical least absolute shrinkage and selection operator (LASSO) regularization was applied.[34] In the context of weighted partial correlation network modeling, the graphical LASSO algorithm achieves regularization by limiting (or 'penalizing') partial correlation coefficients so that their absolute sum is below a certain value.[35] In turn, all estimated coefficients shrink and some become exactly zero. The LASSO aims to set spurious partial correlation weights to zero and generate a sparse adjacency matrix, which results in a force-directed graphical model that is no longer fully saturated. The LASSO algorithm involves a tuning parameter λ (lambda) that controls the sparsity of the regularized network (how many edges are removed). When λ is set to a low value, fewer edges are removed at the cost of retaining more spurious edges. Conversely, a high λ value often removes more spurious edges, but can lead to a larger number of false negatives.

To select sparse core immune network models for each proteomic-based PAH cluster, graphical LASSO regularization was implemented with the 'glasso' (version 1.8) and 'qgraph' ('EBICglasso' function, version 1.43) software packages in R.[29, 36] A collection of 100 possible sparse networks was estimated for each cluster, as LASSO regularization was carried out over 100 different λ values. The tested range spanned from a maximum λ, equaling the largest correlation weight (corresponding network with no retained edges), to a minimum λ that was 0.01* maximum λ. In order to choose the λ value that produced the optimal sparse network, we applied the Extended Bayesian Information Criterion (EBIC).[37] LASSO regularization with EBIC model selection has been shown to perform with high specificity (rarely estimates edges that are not in the 'true' network).[38] The EBIC includes a manually specified hyperparameter γ (gamma) that controls whether simpler or more complex models are preferred. The γ is usually set between 0 (more edges) and 0.5 (fewer edges).[38] Even when $\gamma=0$ (reduces to standard BIC), the selected network is significantly more sparse than a partial correlation network that is not regularized. To select a core immune network model for each machine learned PAH cluster, we set $\gamma=0.1$ to err on the side of discovery.

SUPPLEMENTAL TABLES

Table I. Measured cytokines, chemokines, and growth factors (with aliases)

IL-1α	Interleukin 1 alpha (hematopoetin 1, 1L1F1)
IL-1β	Interleukin 1 beta (leukocytic pyrogen/endogenous mediator, mononuclear cell factor, lymphocyte activating factor,
IL-1RA	Interleukin 1 receptor antagonist (IL1RN, DIRA, ICIL-1RA, IL1F3, IRAP, MVCD4)
IL-2	Interleukin 2 (TCGF, lymphokine)
IL-2RA	Interleukin 2 receptor alpha (CD25, IDDM10, IL2R, TCGFR, p55, IMD41)
IL-3	Interleukin 3 (MCGF, MULTI-CSF)
IL-4	Interleukin 4 (BCGF-1, BSF-1)
IL-5	Interleukin 5 (EDF, TRF)
IL-6	Interleukin 6 (BSF2, HSF, IFNB2, BSF-2, CDF, IFN-beta2)
IL-7	Interleukin 7
IL-8	Interleukin 8 (CXCL8, GCP-1, LECT, LUCT, LYNAP, MDNCF, MONAP, NAF, NAP-1)
IL-9	Interleukin 9 (HP40, P40)
IL-10	Interleukin 10 (human cytokine synthesis inhibitory factor, CSIF, GVHDS, TGIF)
IL-12 (p40)	Interleukin 12 subunit p40 (IL-12 beta, CLMF, CLMF2, IMD28, NKSF, NKSF2, IMD29)
IL-12 (p70)	Interleukin 12 subunit p70 (IL-12 alpha)
IL-13	Interleukin 13 (P600)
IL-15	Interleukin 15
IL-16	Interleukin 16 (LCF, NPRprIL-16)
IL-17	Interleukin 17
IL-18	Interleukin 18 (interferon gamma inducing factor, IGIF, IL-1g, IL1F4)
IFN-α2	Interferon alpha 2 (LeIF A)
IFN-γ	Interferon gamma (IFI, IFG)
TNF-α	Tumor necrosis factor alpha (cachexin, cachetin, DIF, TNFSF2)
TNF-β	Tumor necrosis factor beta (lymphotoxin alpha, LTA, TNFSF1, TNLG1E)
CCL2	Monocyte chemotactic protein 1 (MCP-1, GDCF-2, HC11, HSMCR30, MCAF, SCYA2, SMC-CF)
CCL3	Macrophage inflammatory protein 1-alpha (MIP1α, G0S19-1, LD78ALPHA, SCYA3)
CCL4	Macrophage inflammatory protein 1-beta (MIP1β, ACT2, HC21, LAG-1, SCYA2, SCYA4)
CCL5	Regulated on Activation Normal T-cell Expressed and Secreted (RANTES, D17S136E, SCYA5, SIS-delta, TCP228, eoCP)
CCL7	Monocyte chemotactic protein 3 (MCP-3, FIC, MARC, NC28, SCYA6, SCYA7)
CCL11	Eotaxin (eotaxin-1, eosinophilic chemotactic protein, SCYA11)
CCL27	Cutaneous T cell attracting cytokine (CTACK, SCYA27, ALP, ILC, skinkine, Eskine, PESKY)
CXCL1	GRO alpha oncogene (GROα, FSP, MGSA, MGSA-a, NAP-3, SCYB1)
CXCL9	Monokine induced by gamma interferon (MIG, CMK, Humig, SCYB9, crg-10)
CXCL10	Interferon gamma-induced protein 10 (IP10, C7, IFI10, INP10, SCYB10, crg-2, gIP-10, mob-1)
CXCL12	Stromal cell-derived factor 1 alpha (SDF1α, IRH, PBSF, SCYB12, SDF1, TLSF, TPAR1)
G-CSF	Granulocyte colony stimulating factor (colony stimulating factor 3, CSF3, C17orf33, CSF3OS)
GM-CSF	Granulocyte macrophage colony stimulating factor (colony stimulating factor 2, CSF2)
M-CSF	Macrophage colony stimulating factor (colony stimulating factor 1, CSF1)
FGF	Basic fibroblast growth factor (bFGF, FGF2, FGF-beta)
PDGF-β	Platelet derived growth factor beta-beta homodimer
VEGF	Vascular endothelial growth factor (VEGF-A, vascular permeability factor, VPF, MVCD1)
HGF	Hepatocyte growth factor (scatter factor, DFNB39, F-TCF, HGFB, HPTA, SF)
LIF	Leukemia inhibitory factor (CDF, DIA, HILDA, MLPL1)
β-NGF	Beta nerve growth factor (NGFB, Beta-HSAN5)
SCF	Stem cell factor (Kit-ligand, steel factor, FPH2, FPHH, KL-1, MGF, SF, SHEP7, DCUA, DFNA69)
SCGF-β	Stem cell growth factor beta (C-type lectin domain family 11, CLEC11A, LSLCL, CLECSF3, P47)
TRAIL	TNF-related apoptosis-inducing ligand (TNFSF10, APO2L, CD253, TL2, TNLG6A)
MIF	Macrophage migration inhibitory factor (glycosolation-inhibiting factor, MMIF, GIF, GLIF)

Table II. Discovery cohort consensus clustering: determining the cluster number that optimizes consensus (*k* optima) via the proportion of ambiguously clustered pairs metric. As introduced in the supplemental methods section, the proportion of ambiguously clustered pairs (PAC) is a simple objective measure used to infer the *k* optima (in addition to the tools detailed in Online Figure II). The PAC metric has been shown to outperform other methods,[11] and is defined as the fraction of sample pairs with *consensus index* values falling between 0.1 and 0.9. Because the consensus index is the proportion of resampling runs in which two patients are assigned to the same cluster (0= never cluster together, 1= always cluster together), the PAC range of interest (0.1-0.9) corresponds to pairs of samples that had ambiguous clustering across runs (sometimes clustered together and sometimes did not cluster together). These ambiguously clustered pairs contribute to the mid-portion of the consensus cumulative distribution function curve (see Figure IIB) – a higher PAC translates to a consensus CDF curve with a sloped middle segment, whereas a low PAC results in a flat mid-portion. In turn, the *k* optima is the *k* with the lowest PAC value. As shown below, evaluation of PAC values across *k* inferred a four-cluster structure in our dataset (*k*=4 optima).

Cluster number (*k*)	Proportion of ambiguously clustered pairs (PAC)
k = 2	0.189
k = 3	0.286
***k* = 4**	**0.079**
k = 5	0.209
k = 6	0.224
k = 7	0.208
k = 8	0.217

Table III. Discovery cohort: cluster validity statistics for various consensus clustering algorithm-distance metric combinations. Cluster internal validity statistics are shown for all executed instances of consensus clustering, each of which involved a different clustering algorithm-distance metric combination. For each validity statistic, the top five ranking algorithm-distance combinations are highlighted (refer to legend for color-coding scheme).

		Statistics to be minimized							Statistics to be maximized							
Consensus clustering: algorithm and distance metric combinations	***k*-optima** (cluster number with the best clustering stability)	Proportion of ambiguously clustered pairs (PAC)	Average distance within clusters	Average sum of squares within clusters	G3 coefficient	Average within-cluster to between-cluster distance ratio	Widest within-cluster gap	Connectivity	Average distance between clusters	Average silhouette width	G2 coefficient	Dunn index	Dunn index 2	Normalized Pearson gamma	Calinski and Harabasz index	Separation index
k-medoids + Euclidean	*k*=4	0.079	744.6	356175	0.168	0.681	1479	185.9	1094	0.136	0.567	0.123	0.808	0.402	53.6	333.4
k-medoids + Canberra	*k*=4	0.083	805.2	379255	0.245	0.783	1568	258.0	1029	0.030	0.361	0.079	0.754	0.223	33.4	321.8
k-medoids + Pearson	*k*=5	0.248	864.3	419191	0.304	0.848	1601	279.9	1019	0.045	0.258	0.084	0.760	0.166	23.6	283.4
k-medoids + Spearman	*k*=5	0.213	876.2	435457	0.313	0.867	1551	397.5	1011	0.002	0.210	0.078	0.776	0.135	20.2	279.0
k-medoids + Manhattan	*k*=4	0.076	782.0	362651	0.223	0.756	1482	234.7	1035	0.052	0.409	0.113	0.790	0.253	38.1	319.5
k-means + Euclidean	*k*=4	0.236	755.9	365827	0.187	0.704	1479	166.2	1073	0.101	0.522	0.121	0.831	0.354	49.7	333.2
k-means + Canberra	*k*=4	0.213	794.0	449459	0.227	0.742	1645	205.6	1070	0.061	0.452	0.068	0.530	0.317	23.3	309.7
k-means + Pearson	*k*=4	0.053	886.0	431465	0.334	0.864	1479	257.9	1024	0.050	0.213	0.089	0.813	0.157	28.1	290.0
k-means + Spearman	*k*=4	0.110	892.2	440774	0.345	0.866	1583	248.6	1030	0.004	0.206	0.089	0.689	0.162	25.6	291.3
k-means + Manhattan	*k*=5	0.155	754.5	358330	0.170	0.654	1507	207.5	1093	0.100	0.635	0.120	0.630	0.395	42.5	326.4
Hierarchical + Euclidean	*NF*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Hierarchical + Canberra	*NF*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Hierarchical + Pearson	*k*=4	0.258	901.2	433476	0.358	0.875	1482	206.1	1030	0.005	0.184	0.078	0.806	0.154	27.5	285.0
Hierarchical + Spearman	*k*=4	0.263	896.9	441335	0.351	0.872	1551	227.6	1029	0.012	0.202	0.085	0.793	0.156	25.4	290.7
Hierarchical + Manhattan	*NF*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

For each cluster validity statistic, algorithm-distance combinations with the best objective performance are denoted (color spectrum indicates rank): 1 2 3 4 5

NF= No k-optima found for these consensus clustering algorithm-distance combinations

Table IV. Discovery cohort: proteomic immune cluster membership by PAH etiology subtype. The number and percentage of patients with each underlying PAH etiology is shown by proteomic-based cluster. There was no significant difference in the overall distribution of etiologies within clusters (p=0.110 by Fisher's exact test). Across-cluster comparisons for each etiology are also shown. Idiopathic PAH and congenital heart disease-associated PAH are somewhat overrepresented in cluster 3 and cluster 1, respectively.

	Overall (n=281)	**Cluster 1 (n=58)**	**Cluster 2 (n=109)**	**Cluster 3 (n=77)**	**Cluster 4 (n=37)**	**p-value ***
WHO etiologic subtype, n (%)						
Connective tissue disease-APAH	87 (31.0)	20 (34.5)	31 (28.4)	24 (31.2)	12 (32.4)	0.873
Idiopathic PAH	84 (29.9)	11 (19.0)	31 (28.4)	31 (40.2)	11 (29.7)	0.061
Drugs and toxins-APAH	49 (17.4)	11 (19.0)	18 (16.5)	11 (14.3)	9 (24.3)	0.592
Congenital heart disease-APAH	38 (13.5)	6 (10.3)	23 (21.1)	6 (7.8)	3 (8.1)	0.031
Portopulmonary hypertension	19 (6.8)	8 (13.7)	5 (4.6)	4 (5.2)	2 (5.4)	0.167
Heritable PAH	4 (1.4)	2 (3.4)	1 (0.9)	1 (1.3)	-	-

* Comparison across clusters by Pearson's chi-square test, or Fisher's exact test when chi-square assumptions were not met.

Table V. Discovery cohort: timing of plasma sampling by proteomic immune cluster, relative to PAH diagnosis and symptom onset. Tabular comparison of the time intervals from PAH diagnosis to plasma collection, symptom onset (patient reported) to plasma collection, and symptom onset to PAH diagnosis.

	Overall (n=281)	**Cluster 1 (n=58)**	**Cluster 2 (n=109)**	**Cluster 3 (n=77)**	**Cluster 4 (n=37)**	**p-value ***
PAH diagnosis to plasma sampling, years, median (IQR)	1.0 (0.0, 4.7)	1.1 (0.0, 5.4)	1.3 (0.0, 5.3)	0.9 (0.0, 3.5)	0.0 (0.0, 4.6)	0.193
Interval < 1 year, n (%)	137 (48.8)	25 (43.1)	60 (45.0)	40 (51.9)	23 (62.2)	0.226
Interval < 2 years, n (%)	167 (59.4)	32 (55.2)	63 (57.8)	48 (62.3)	24 (64.9)	0.734
Symptom onset to plasma sampling, years, median (IQR)	3.0 (1.1, 5.8)	3.3 (1.1, 6.4)	3.4 (1.5, 6.4)	2.1 (1.0, 4.9)	2.3 (0.5, 5.0)	0.067
Interval < 2 years, n (%)	106 (37.7)	22 (37.9)	33 (30.3)	33 (42.9)	18 (48.6)	0.150
Interval < 5 years, n (%)	194 (69.0)	37 (63.8)	70 (64.2)	60 (77.9)	27 (73.0)	0.169
Symptom onset to PAH diagnosis, years, median (IQR)	0.8 (0.4, 1.7)	0.7 (0.3, 1.5)	0.9 (0.4, 2.0)	0.9 (0.4, 1.5)	0.6 (0.4, 1.2)	0.343
Interval < 1 year, n (%)	159 (56.6)	34 (58.6)	56 (51.2)	43 (55.8)	26 (70.3)	0.247

* The Kruskal-Wallis test was used to compare continuous time interval variables (each had a skewed distribution), and the chi-square test was applied for binary variables.

Table VI. Discovery cohort: prognostic clinical biomarkers by protemic immune cluster at the time of plasma sampling. New York Heart Association (NYHA) functional class, six-minute walk distance (6MWD), N-terminal prohormone of B-type natriuretic peptide (NT-proBNP), diffusion capacity of the lung for carbon monoxide (DLCO), and the multi-parameter REVEAL risk score are compared across clusters.

	Cluster 1 (n=58)	**Cluster 2 (n=109)**	**Cluster 3 (n=77)**	**Cluster 4 (n=37)**	**p-value ***
NYHA functional class, n (%)					
Class I or II	18 (31.0)	44 (40.4)	31 (40.3)	13 (35.1)	0.627
Class IV	13 (22.4)	11 (10.1)	4 (5.2)	3 (8.1)	0.020 †
6MWD ‡, meters, median (IQR)	378 (305, 443)	390 (306, 485)	446 (340, 534)	404 (268, 449)	0.247
>440 meters, n (%)	15 (26.3)	37 (35.9)	37 (51.4)	15 (44.1)	0.026 †
<165 meters, n (%)	3 (5.3)	4 (3.9)	8 (11.1)	3 (8.8)	0.251
NT-proBNP, pg/ml, median (IQR)	746 (98, 2540)	312 (96, 1499)	254 (80, 850)	415 (112, 1187)	0.041 †
<300 pg/ml, n (%)	22 (37.9)	53 (48.6)	47 (61.0)	17 (45.9)	0.050 †
>1500 pg/ml, n (%)	23 (39.7)	27 (24.8)	11 (14.3)	8 (21.6)	0.008 †
DLCO ‡, % predicted, median (IQR)	57 (48, 70)	67 (47, 81)	66 (56, 90)	79 (64, 86)	0.072
>80% of predicted, n (%)	9 (17.0)	29 (27.6)	21 (30.4)	16 (45.7)	0.033 †
<40% of predicted, n (%)	6 (11.3)	16 (15.2)	7 (10.1)	4 (11.4)	0.795
REVEAL risk score, n (%)					
Low-average risk (score 1-8)	32 (55.2)	75 (68.8)	61 (79.2)	25 (67.6)	0.030 †
High-very high risk (score >10)	21 (36.2)	25 (22.9)	10 (13.0)	10 (27.0)	0.017 †

* Comparison across clusters with the Kruskal-Wallis test (continuous variables) or chi-square test (categorical variables)

† Post-hoc pairwise cluster comparisons were performed (Dunn's test for continuous variables, and independent two-cluster chi-square tests for categorical data), with Benjamini-Hochberg (BH) adjustment of p-values for multiple testing.
Cluster differences that achieved statistical significance (p<0.05) after BH adjustment:
NYHA FC IV (Cluster 1 vs 3 p=0.022)
6MWD >440 meters (Cluster 1 vs 3 p=0.039)
NT-proBNP median (Cluster 1 vs 3 p=0.046)
NT-proBNP <300 pg/ml (Cluster 1 vs 3 p=0.046)
NT-proBNP >1500 pg/ml (Cluster 1 vs 3 p=0.007)
DLCO >80% of predicted (Cluster 1 vs 4 p=0.043)
REVEAL score low-average risk (Cluster 1 vs 3 p=0.027)
REVEAL score high-very high risk (Cluster 1 vs 3 p=0.012)

‡ Missing data: 6MWD (Cluster 1 [n=1], Cluster 2 [n=6], Cluster 3 [n=5], Cluster 4 [n=3])
DLCO (Cluster 1 [n=5], Cluster 2 [n=4], Cluster 3 [n=8], Cluster 4 [n=2])

Abbreviations: NYHA= New York Heart Association, 6MWD= six-minute walk distance, NT-proBNP= N-terminal pro b-type natriuretic peptide, DLCO= diffusion capacity of lung for carbon dioxide.

Table VII. Discovery cohort: echocardiographic findings by proteomic immune cluster. Data shown is captured from transthoracic echocardiograms performed within 3 months of plasma sample collection (available for all patients).

	Cluster 1 (n=58)	Cluster 2 (n=109)	Cluster 3 (n=77)	Cluster 4 (n=37)	p-value *
Pericardial effusion, n (%)	12 (20.7)	15 (13.8)	7 (9.1)	7 (18.9)	0.239
RVSP ‡, mmHg, median (IQR)	82 (64, 94)	71 (54, 84)	65 (47, 80)	75 (49, 92)	0.009 †
RV FAC ‡, %, median (IQR)	24.0 (21.2, 28.4)	26.7 (22.1, 31.7)	29.1 (24.8, 33.5)	26.1 (22.7, 30.4)	0.032 †
TAPSE ‡, cm, median (IQR)	1.5 (1.2, 1.8)	1.8 (1.5, 2.1)	1.9 (1.5, 2.1)	1.6 (1.3, 1.9)	0.134
≤ 1.6 cm, n (%)	28 (53.8)	28 (32.2)	18 (30.0)	17 (53.1)	0.011 †

* Comparison across clusters with the Kruskal-Wallis test (continuous variables) or chi-square test (categorical variables)

† Post-hoc pairwise cluster comparisons were performed (Dunn's test for continuous variables, and independent two-cluster chi-square tests for categorical data), with Benjamini-Hochberg (BH) adjustment of p-values for multiple testing.

Cluster differences that achieved statistical significance ($p<0.05$) after BH adjustment:
RVSP (Cluster 1 vs 2 p=0.032, Cluster 1 vs 3 p=0.001)
RVFAC (Cluster 1 vs 3 p=0.021)
TAPSE ≤1.6 cm (Cluster 1 vs 2 p=0.040, Cluster 1 vs 3 p=0.040)

‡ Missing data: RVSP (Cluster 1 [n=2], Cluster 2 [n=12], Cluster 3 [n=1], Cluster 4 [n=3]),
RV FAC and TAPSE (Cluster 1 [n=6], Cluster 2 [n=22], Cluster 3 [n=17], Cluster 4 [n=5])

Abbreviations: RVSP= estimated right ventricular systolic pressure, RV FAC= right ventricular fractional area change, TAPSE= tricuspid annular plane systolic excursion.

Table VIII. Discovery cohort: hemodynamic disease severity by proteomic immune cluster. Hemodynamic measurements obtained on the day of plasma sample collection are indicated for each cluster (median and interquartile range for all data).

	Cluster 1 (n=58)	Cluster 2 (n=109)	Cluster 3 (n=77)	Cluster 4 (n=37)	p-value *
mPAP, mmHg	52 (45, 60)	46 (34, 60)	45 (35, 54)	49 (32, 58)	0.043 †
PVR ‡ , dynes·sec/cm^5	824 (584, 1216)	704 (424, 1160)	648 (408, 960)	664 (400, 1168)	0.284
Cardiac index ‡, ml/min/m^2	2.2 (1.9, 2.5)	2.1 (1.7, 2.4)	2.0 (1.8, 2.4)	2.2 (1.7, 2.3)	0.549
Right atrial pressure ‡, mmHg	11 (5, 14)	8 (5, 11)	7 (5, 10)	6 (4, 8)	0.023 †
PAWP, mmHg	12 (9, 14)	10 (8, 14)	11 (8, 14)	10 (8, 14)	0.515
RVEDP, mmHg	14 (8, 20)	12 (8, 15)	11 (8, 14)	10 (7, 13)	0.044 †
Heart rate, beats/min	80 (72, 95)	76 (69, 86)	76 (68, 86)	74 (66, 87)	0.104
Systolic BP, mmHg	118 (107, 131)	120 (111, 135)	124 (109, 144)	119 (111, 130)	0.521

* The Kruskal-Wallis test was used to compare hemodynamic indices across clusters, as each variable had a skewed distribution.

† Post-hoc pairwise cluster comparisons were performed by applying Dunn's test, with Benjamini-Hochberg (BH) adjustment of p-values for multiple testing. Cluster differences that achieved statistical significance ($p<0.05$) after BH adjustment:
mPAP (Cluster 1 vs 3 p=0.005, Cluster 1 vs 2 p=0.045)
Right atrial pressure (cluster 1 vs 4 p=0.002, cluster 1 vs 3 p=0.030, cluster 1 vs 2 p=0.041)
RVEDP (Cluster 1 vs 4 p=0.030)

‡ Missing data: PVR and right atrial pressure (Cluster 1 [n=1], Cluster 2 [n=1]), cardiac index (Cluster 1 [n=2], Cluster 2 [n=2])

Abbreviations: mPAP= mean pulmonary arterial pressure, PVR= pulmonary vascular resistance, PAWP= pulmonary arterial wedge pressure, RVEDP= right ventricular end diastolic pressure, BP= systemic blood pressure

Table IX. Discovery cohort: immunoassay plates and experimental batches by PAH proteomic immune cluster. PAH samples were run on nine assay pates across two experimental batches. As shown in the table below, machine learned clusters had comparable sample percentages from each immunoassay plate and experimental batch. This finding suggests that cluster assignments were not driven by residual plate or batch effects, and provides additional evidence (beyond the principal component analyses shown in Online Figure I) that batch adjustments carried out in the data pre-processing stage were effective.

	Overall (n=281)	**Cluster 1 (n=58)**	**Cluster 2 (n=109)**	**Cluster 3 (n=77)**	**Cluster 4 (n=37)**
Batch A					
Plate 1, n (%)	14 (5.0)	2 (3.4)	4 (3.7)	4 (5.2)	4 (10.8)
Plate 2	16 (5.7)	4 (6.9)	6 (5.5)	4 (6.5)	1 (2.7)
Plate 3	39 (13.9)	10 (17.2)	15 (13.8)	10 (13.0)	4 (10.8)
Plate 4	43 (15.3)	9 (15.5)	17 (15.6)	9 (11.7)	8 (21.6)
Plate 5	28 (10.0)	9 (15.5)	9 (8.3)	6 (7.8)	4 (10.8)
Total	***140 (49.8)***	***34 (58.6)***	***51 (46.8)***	***34 (44.2)***	***21 (56.8)***
Batch B					
Plate 6	42 (14.9)	6 (10.3)	20 (18.3)	12 (15.6)	4 (10.8)
Plate 7	35 (12.5)	6 (10.3)	16 (14.7)	9 (11.7)	4 (10.8)
Plate 8	42 (14.9)	7 (12.1)	15 (13.8)	14 (18.2)	6 (16.2)
Plate 9	22 (7.8)	5 (8.6)	7 (6.4)	8 (10.4)	2 (5.4)
Total	***141 (50.2)***	***24 (41.4)***	***58 (53.2)***	***43 (55.8)***	***16 (43.2)***

Table X. Comparison of discovery and validation cohort patient characteristics. Demographics, PAH subtype, treatment status, non-invasive disease metrics, and hemodynamics were captured at the time of proteomic sampling (within one month).

	Discovery cohort: Stanford (n=281)	**Validation cohort:** Sheffield (n=104)	**p-value** *
Demographics			
Age, years, median (IQR)	52 (41–63)	60 (51–66)	<0.001
Female sex, n (%)	206 (73.3)	65 (62.5)	0.039
Race, n (%)			
White	155 (55.2)	98 (94.2)	<0.001
Asian	49 (17.4)	6 (5.8)	
Hispanic	43 (15.3)	0 (0.0)	
Black	19 (6.8)	0 (0.0)	
Other	15 (5.3)	0 (0.0)	
PAH subtypes and treatment status			
PAH etiology, n (%)			
Connective tissue disease-APAH	87 (31.0)	37 (35.6)	<0.001
Idiopathic PAH	84 (29.9)	53 (51.0)	
Drugs and toxins-APAH	49 (17.4)	0 (0.0)	
Congenital heart disease-APAH	38 (13.5)	7 (6.7)	
Portopulmonary hypertension	19 (6.8)	7 (6.7)	
Hereditary PAH	4 (1.4)	0 (0.0)	
Treatment naïve, n (%)	123 (43.8)	104 (100.0)	<0.001
Non-invasive disease metrics			
NYHA/WHO functional classification †, n (%)			
Class I	17 (6.0)	2 (2.0)	0.008
Class II	89 (31.7)	18 (18.2)	
Class III	141 (50.2)	68 (68.7)	
Class IV	34 (12.1)	11 (11.1)	
DLCO, % of predicted †, median (IQR)	69 (50–84)	46 (35–56)	<0.001
Hemodynamics			
Mean PAP, mmHg, median (IQR)	47 (36–59)	50 (40–60)	0.333
PVR †, dynes·sec/cm^5, median (IQR)	720 (424–1104)	688 (376–872)	0.028
Right atrial pressure †, mmHg, median (IQR)	8 (5–12)	10 (7–15)	<0.001
PAWP, mmHg, median (IQR)	10 (8–14)	11 (8 –14)	0.649
Outcome events during follow-up			
Deaths, n (%)	62 (22.1)	52 (50.0)	-
Transplantation (lung or heart-lung), n (%)	17 (6.0)	-	-
Events per 100 patient-years of follow-up	8.67	14.36	-

* Two-group comparison by t-test (continuous dependent variables, normal), Mann-Whitney U test (continuous dependent variables, non-normal), chi-square test (categorical dependent variable), or Fisher's exact test (chi-square assumptions not met).

† Missing data: functional class (validation n=5), six-minute walk, DLCO (discovery n=19, validation n=16), PVR (discovery n=2, validation n=3), right atrial pressure (discovery n=2).

Table XI. Validation cohort: Clinical comparison of PAH proteomic immune clusters at plasma sampling.

	Cluster 1 (n=26)	Cluster 2 (n=33)	Cluster 3 (n=36)	Cluster 4 (n=9)	p-value *
Age, years, median (IQR)	61 (57–69)	59 (49–62)	62 (49–68)	56 (42–60)	0.074
Female sex, n (%)	15 (57.8)	20 (60.6)	26 (72.2)	4 (44.4)	0.390
PAH etiology, n (%)					
Idiopathic PAH	11 (42.3)	21 (63.6)	15 (41.7)	6 (66.7)	0.188
CTD-APAH	10 (38.5)	7 (21.2)	18 (50.0)	2 (22.2)	
PoPH	2 (7.7)	4 (12.1)	1 (2.8)	0 (0.0)	
CHD-APAH	3 (11.5)	1 (3.0)	2 (5.6)	1 (11.1)	
NYHA functional class†, n (%)					
Class I	0 (0.0)	0 (0.0)	1 (3.0)	1 (11.1)	0.035
Class II	2 (8.3)	4 (12.1)	11 (33.3)	1 (11.1)	
Class III	17 (70.8)	27 (81.8)	19 (57.6)	5 (55.6)	
Class IV	5 (20.8)	2 (6.1)	2 (6.1)	2 (22.2)	
Incremental shuttle walk distance†, meters, median (IQR)	120 (80–210)	180 (75–330)	170 (110–290)	230 (148-323)	0.338
DLCO†, % predicted, median (IQR)	41 (27–48)	49 (41–58)	46 (37–58)	45 (29–47)	0.126
>80% of predicted, n (%)	0 (0.0)	1 (3.4)	2 (6.7)	0 (0.0)	0.826
<40% of predicted, n (%)	9 (41.0)	7 (24.1)	11 (36.7)	3 (42.9)	0.540
Hemodynamics, median (IQR)					
Mean PAP, mmHg	49 (40–55)	53 (44–63)	45 (32–61)	48 (41–58)	0.385
PVR†, dynes·sec/cm^5	608 (392–856)	664 (392–920)	568 (328–888)	544 (360–872)	0.747
Cardiac index†, L/min/m^2	2.7 (2.2–3.1)	2.7 (2.3–3.4)	2.9 (2.4–3.3)	3.1 (2.5–3.7)	0.774
RAP, mmHg	12 (9–16)	10 (7–14)	9 (5–14)	7 (5–9)	0.045
PAWP, mmHg	12 (9–15)	11 (8–13)	10 (7–13)	11 (10–13)	0.224

* Comparison across clusters with the Kruskal-Wallis test (continuous variables), chi-square test (categorical variables), or Fisher's exact test (categorical variables when chi-square assumptions not met)

† Missing data: NYHA functional class (cluster 1 [n=2], cluster 3 [n=3]), incremental shuttle walk distance (cluster 1 [n=5], cluster 2 [n=2], cluster 3 [n=5], cluster 4 [n=1]), DLCO (cluster 1 [n=4], cluster 2 [n=3], cluster 3 [n=6], cluster 4 [n=2]), PVR (cluster 1 [n=1], cluster 3 [n=1], cluster 4 [n=1]), cardiac index (cluster 3 [n=2], cluster 4 [n=1])

Table XII. Cox proportional hazards modeling of survival in pooled discovery and validation cohort (n=385).
[A] Univariate analysis. The five-year risk of death (hazard ratio and 95% confidence interval) associated with the following variables is shown: immune cluster status (each cluster vs. remainder of cohort), age, sex, PAH etiology (CTD-APAH or PoPH), treatment status (incident vs prevalent PAH), functional class (III or IV), and hemodynamic metrics. Multivariable models were constructed to evaluate risk associated with **[B]** immune cluster 1 and **[C]** immune cluster 3, after adjustment for covariates that were selected *a priori* (age, sex, PAH etiology) or significant in univariate analysis (incident PAH, functional class, and right atrial pressure had $p<0.05$). P-values shown reflect the Wald test statistic. Cox regression was applied with the Efron estimator to handle ties,[39] and the proportional hazards assumption was tested by evaluating relationships between Schoenfeld residuals and time for each variable.

[A] Univariate analysis

	Hazard Ratio (95% CI)	**p-value**
Immune cluster (vs. rest of cohort)		
Cluster 1	**2.52 (1.76 – 3.63)**	**< 0.0001**
Cluster 2	0.91 (0.63 – 1.31)	0.601
Cluster 3	**0.42 (0.27 – 0.67)**	**0.0002**
Cluster 4	1.02 (0.60 – 1.72)	0.944
Age (per year older)	1.02 (1.00 – 1.03)	0.032
Male sex	1.18 (0.81 – 1.70)	0.392
CTD-APAH or PoPH	1.64 (1.25 – 2.16)	0.024
Incident PAH (treatment naïve)	1.32 (1.02 – 1.75)	0.046
NYHA functional class III or IV	2.18 (1.42 – 3.36)	< 0.0001
mPAP (per 1 mmHg increase)	1.01 (0.99 – 1.02)	0.167
RAP (per 1 mmHg increase)	1.05 (1.02 – 1.08)	0.0005
PVR (per 80 dynes·sec/cm^5 increase)	1.01 (0.98 – 1.04)	0.408

[B] Multivariable analysis of immune cluster 1 risk.

	Hazard Ratio (95% CI)	**p-value**
Immune Cluster 1 (vs. rest of cohort)	**2.19 (1.51 – 3.18)**	**< 0.0001**
Age (per year older)	1.01 (0.99 – 1.02)	0.139
Male sex	1.23 (0.83 – 1.83)	0.297
CTD-APAH or PoPH	1.41 (1.05 – 2.14)	0.040
Incident PAH (treatment naïve)	1.40 (0.95 – 2.06)	0.089
NYHA functional class III or IV	2.01 (1.27 – 3.19)	0.003
RAP (per 1 mmHg increase)	1.04 (1.01 – 1.08)	0.010

[C] Multivariable analysis of immune cluster 3 risk

	Hazard Ratio (95% CI)	**p-value**
Immune Cluster 3 (vs. rest of cohort)	**0.39 (0.25 – 0.63)**	**0.0001**
Age (per year older)	1.01 (0.99 – 1.02)	0.134
Male sex	1.23 (0.83 – 1.81)	0.299
CTD-APAH or PoPH	1.38 (1.02 – 2.02)	0.047
Incident PAH (treatment naïve)	1.36 (0.93 – 2.01)	0.111
NYHA functional class III or IV	2.12 (1.34 – 3.36)	0.001
RAP (per 1 mmHg increase)	1.05 (1.02 – 1.09)	0.003

Table XIII. Standard curve-derived cytokine concentrations by PAH immune cluster. Median analyte concentrations (with 25-75% IQR) are displayed in pg/mL for PAH patients overall (pooled discovery and validation cohorts), healthy controls, and immune clusters identified in each cohort. Colored asterisks indicate cytokines that had central network importance in clusters (according to the proteomic network analysis detailed elsewhere). Concentration levels were determined by incorporating a standard dilution series on each assay plate (contained known analyte concentrations), plotting detected MFI vs known concentration for standards, and deriving a five-parameter logistic regression model to predict concentrations from detected MFI for experimental samples. Relative to standard curve-derived concentration values, MFI immunoassay data does not require detection limit censoring, more accurately quantifies analytes with low abundance, is less susceptible to non-biological variation, and has greater statistical power in downstream analysis.[40] Therefore, all other study analyses we report are based on MFI. *This table is not intended to provide a differential statistical analysis of analyte levels across PAH clusters, but instead presented to offer a general reference of standard curve-derived protein concentrations.*

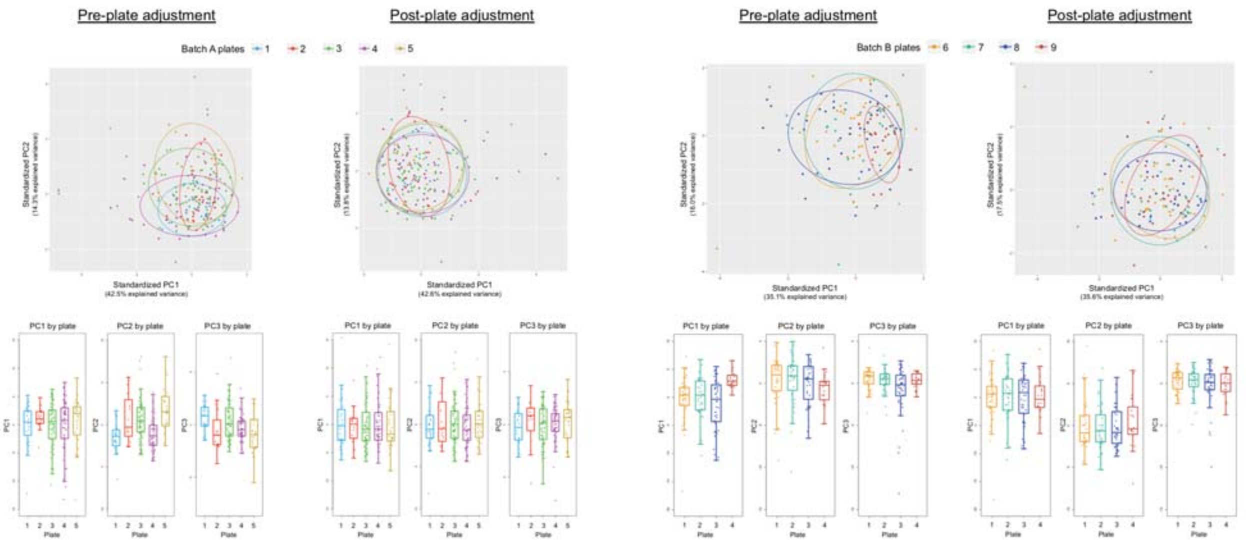
Cytokine pg/mL, median (IQR)	**Cluster 1** Discovery (n=58) Validation (n=26)	**Cluster 2** Discovery (n=109) Validation (n=33)	**Cluster 3** Discovery (n=77) Validation (n=36)	**Cluster 4** Discovery (n=37) Validation (n=9)	**PAH overall** Discovery + validation (n=385)	**Healthy controls** (n=88)
IL–18 *	156 (70–223) 166 (115–263)	61 (26–104) 74 (0–188)	62 (33–111) 40 (0–125)	62 (47–112) 112 (12–234)	74 (38–141)	51 (45–58)
IL–16	400 (326–549) 495 (414–594)	308 (250–381) 320 (249–480)	331 (269–400) 338 (225–496)	339 (279–438) 264 (211–308)	335 (262–459)	270 (223–316)
SCF	126 (102–179) 152 (127–208)	92 (76–110) 94 (73–114)	94 (74–105) 76 (56–116)	83 (67–96) 76 (74–86)	97 (76–123)	87 (75–101)
CXCL9 *	712 (438–2041) 1269 (879–1821)	551 (345–1006) 402 (224–842)	474 (238–714) 568 (290–972)	518 (304–777) 659 (291–919)	568 (328–1023)	316 (203–570)
CTACK	743 (573–902) 984 (742–1185)	675 (515–867) 632 (499–715)	567 (469–735) 654 (524–767)	586 (478–692) 504 (462–583)	654 (513–836)	492 (404–630)
HGF	684 (454–1102) 696 (427–1324)	432 (346–565) 497 (249–812)	398 (354–529) 473 (256–714)	404 (328–588) 267 (200–393)	453 (348–714)	287 (231–336)
SCGF–β	28464 (19272–42055) 23186 (17733–36193)	16521 (12013–27375) 22891 (18296–32209)	17362 (11698–22380) 16397 (11375–24167)	19571 (13777–29624) 15417 (8911–19351)	19541 (13179–28519)	9044 (6275–11739)
CXCL12	592 (545–641) 524 (487–623)	501 (441–539) 484 (442–517)	513 (455–563) 498 (449–542)	532 (482–585) 433 (420–483)	515 (455–577)	398 (369–428)
β–NGF	12.2 (10.9–13.6) 10.0 (8.9–12.1)	9.2 (8.0–10.3) 9.5 (7.8–10.5)	9.7 (8.5–11.1) 10.9 (9.2–12.2)	10.3 (8.0–11.4) 11.3 (9.6–11.7)	10.1 (8.4–11.6)	8.1 (7.1–8.9)
TNF–β *	2.3 (0.5–4.5) 1.4 (0.7–2.3)	0.0 (0.0–0.4) 0.0 (0.0–0.4)	0.3 (0.0–0.9) 0.6 (0.0–2.0)	0.3 (0.0–1.8) 4.4 (1.1–6.9)	0.3 (0.00–1.5)	0.0 (0.0–0.0)
IL–1α **	0.5 (0.0–0.9) 0.0 (0.0–0.0)	0.0 (0.0–0.0) 0.0 (0.0–0.0)	1.5 (0.5–2.4) 1.0 (0.0–2.1)	2.4 (1.8–4.3) 4.3 (3.8–6.8)	0.0 (0.0–1.4)	0.0 (0.0–0.0)
LIF	24.3 (15.3–36.0) 19.6 (7.5–30.2)	8.1 (4.6–15.4) 13.9 (7.5–27.0)	15.2 (6.2–24.3) 7.5 (4.6–18.2)	16.5 (7.1–26.8) 7.5 (4.6–34.0)	13.2 (6.1–25.2)	10.0 (0.0–18.4)
IFN–α	132 (122–142) 125 (109–133)	112 (103–120) 110 (102–119)	118 (109–126) 122 (110–133)	119 (103–125) 128 (119–136)	118 (108–128)	107 (97–111)
IL–3 *	228 (202–256) 183 (147–227)	160 (132–189) 141 (119–175)	180 (143–215) 184 (167–223)	175 (127–203) 201 (189–243)	176 (140–217)	158 (133–175)
CCL7 **	26 (21–32) 23 (14–29)	14 (8.3–19) 16 (8.9–28)	23 (16–27) 22 (12–27)	20 (12–27) 27 (21–30)	19 (12–26)	14 (11–15)
M–CSF	14.4 (10.3–19.3) 14.8 (8.8–21.5)	4.5 (3.0–8.6) 6.7 (4.4–11.0)	7.2 (4.0–11.1) 6.4 (2.8–11.3)	5.9 (4.5–12.3) 8.9 (6.9–15.9)	7.5 (4.0–13.1)	2.6 (0.0–4.6)
IL–12(p40)	91 (52–168) 76 (51–123)	13 (0.0–29) 17 (0.0–28)	15 (0.0–25) 0.0 (0.0–13)	0.0 (0.0–44) 0.0 (0.0–57)	17 (0.0–53)	0.0 (0.0–36)
IL–2Ra	343 (245–496) 423 (241–550)	205 (137–255) 88 (13–179)	241 (168–303) 263 (141–379)	208 (107–281) 361 (309–468)	238 (150–345)	189 (164–207)
TRAIL *	120 (94–167) 122 (44–155)	74 (37–104) 81 (15–128)	88 (58–129) 84 (28–132)	86 (64–126) 118 (53–189)	89 (62–135)	89 (70–112)
CCL2 *	44 (33–74) 51 (25–86)	41 (26–56) 27 (0–57)	52 (36–59) 52 (36–83)	65 (39–86) 102 (61–140)	47 (30–70)	22 (17–30)
GM–CSF *	69 (35–130) 74 (37–114)	57 (18–108) 41 (0–71)	80 (43–143) 117 (89–146)	99 (55–183) 229 (226–258)	71 (36–95)	40 (28–60)

Table XIII (continued). Standard curve-derived cytokine concentrations by PAH immune cluster.

Cytokine pg/mL, median (IQR)	**Cluster 1** Discovery (n=58) Validation (n=26)	**Cluster 2** Discovery (n=109) Validation (n=33)	**Cluster 3** Discovery (n=77) Validation (n=36)	**Cluster 4** Discovery (n=37) Validation (n=9)	**PAH overall** Discovery + validation (n=385)	**Healthy controls** (n=88)
CCL11 *	86 (58–123) 103 (55–184)	62 (26–101) 53 (8–105)	93 (71–142) 98 (57–156)	149 (80–188) 213 (191–461)	85 (50–142)	28 (16–40)
IL–2	14 (9.0–23) 16 (12–18)	14 (8.9–25) 15 (8.5–17)	34 (18–45) 25 (16–33)	61 (14–73) 77 (67–84)	16 (14–36)	0.0 (0.0–3.9)
IL–1Ra	253 (44–519) 270 (18–645)	265 (54–529) 42 (0–439)	894 (326–1530) 1422 (1065–1797)	1886 (300–3062) 2515 (2323–3774)	449 (44–1185)	284 (180–403)
IL–15 **	0 (0.0–2.9) 3 (0.0–3.8)	0.0 (0.0–2.3) 0.0 (0.0–0.0)	15 (8.4–34) 10 (5.4–24)	43 (34–52) 49 (42–52)	3.3 (0.0–12)	0.0 (0.0–0.0)
IL–6 *	28 (18–40) 31 (18–43)	25 (17–35) 19 (12–24)	43 (26–55) 38 (34–46)	65 (32–85) 61 (54–93)	31 (20–49)	20 (16–30)
IL–10 *	21 (12–32) 25 (22–28)	22 (14–30) 22 (18–25)	43 (26–52) 34 (28–38)	67 (24–95) 59 (39–67)	27 (20–43)	17 (12–20)
IL–8 *	42 (36–61) 42 (29–66)	44 (33–59) 29 (20–42)	71 (51–87) 82 (71–95)	94 (52–125) 125 (114–151)	52 (37–80)	32 (26–46)
IFN–γ	253 (155–358) 221 (163–329)	272 (193–366) 163 (124–254)	426 (304–510) 494 (457–573)	635 (353–837) 968 (748–1030)	321 (209–486)	202 (153–262)
IL–4 *	10.6 (7.2–13.2) 10.0 (7.7–12.9)	11.1 (8.1–13.9) 7.4 (5.6–9.9)	16.5 (12.3–19.3) 18.5 (15.6–20.7)	21.8 (12.9–26.0) 26.9 (20.1–30.9)	12.6 (8.6–18.0)	6.2 (4.8–8.2)
IL–1β *	6.3 (4.5–8.4) 6.1 (5.2–7.3)	6.6 (5.2–9.1) 4.8 (4.0–6.1)	10.2 (7.3–12.3) 11.7 (9.8–13.0)	15.8 (6.8–19.3) 16.6 (14.2–19.9)	7.8 (5.5–11.7)	5.1 (3.7–6.9)
IL–5	10.5 (7.6–14.5) 10.7 (8.2–12.0)	11.2 (8.2–14.9) 8.4 (7.4–10.9)	18.2 (13.2–21.7) 18.1 (15.8–19.7)	26.3 (14.3–35.9) 25.0 (20.2–31.5)	13.2 (9.1–19.3)	10.2 (7.7–13.6)
G–CSF *	335 (259–406) 321 (131–426)	321 (253–378) 218 (202–305)	397 (307–453) 430 (367–492)	504 (410–576) 599 (526–675)	362 (272–452)	205 (149–266)
IL–7 **	31 (21–43) 34 (24–39)	32 (20–46) 21 (18–34)	47 (33–61) 47 (38–55)	70 (32–84) 67 (61–87)	37 (24–55)	27 (21–35)
IL–13 *	17.3 (10.3–23.2) 14.7 (13.0–17.0)	14.5 (8.3–20.5) 13.1 (10.7–16.6)	22.0 (14.4–29.8) 18.5 (16.9–22.2)	27.4 (16.6–35.5) 20.8 (18.4–28.2)	16.6 (12.4–24.2)	10.6 (7.4–13.1)
CCL3 *	9.1 (4.1–11.4) 9.3 (8.6–10.0)	8.5 (5.2–10.9) 8.4 (7.5–8.9)	12.9 (9.1–17.4) 11.1 (9.4–12.6)	20.7 (10.6–27.3) 15.8 (12.5–19.8)	9.6 (7.6–13.5)	7.5 (6.0–9.6)
FGFb *	119 (75–138) 101 (90–121)	108 (82–142) 87 (76–99)	162 (122–191) 150 (133–173)	223 (127–270) 214 (206–314)	124 (91–172)	101 (81–133)
IL–17 *	64 (15–104) 75 (63–84)	60 (12–99) 20 (0.0–49)	124 (67–194) 117 (66–199)	210 (84–352) 179 (138–326)	78 (39–161)	24 (3.2–51)
TNF–α *	93 (48–137) 90 (78–115)	88 (56–131) 76 (65–96)	152 (94–200) 155 (138–212)	209 (114–381) 277 (205–427)	114 (73–174)	77 (50–111)
IL–12(p70) **	48 (20–81) 68 (42–73)	53 (29–85) 33 (15–54)	109 (80–145) 81 (72–96)	195 (137–257) 185 (168–227)	73 (38–117)	58 (44–86)
IL–9 *	22 (10–40) 24 (13–37)	18 (8–35) 13 (7–18)	48 (30–64) 55 (40–63)	92 (32–114) 77 (74–154)	31 (12–56)	13 (10–21)
CCL4 *	90 (65–120) 94 (80–101)	69 (57–96) 75 (66–81)	77 (60–100) 87 (68–97)	102 (75–130) 101 (87–128)	81 (63–102)	45 (33–59)
CXCL10 *	1168 (741–1978) 1661 (1180–2293)	880 (565–1309) 787 (391–1771)	741 (546–1316) 707 (438–996)	1026 (730–1631) 1483 (558–2448)	950 (559–1513)	429 (312–563)
MIF *	715 (598–917) 640 (531–816)	534 (451–644) 467 (354–622)	565 (517–687) 732 (575–832)	599 (502–714) 733 (620–931)	593 (489–733)	504 (496–555)
CXCL1	50 (32–92) 41 (18–53)	30 (3–76) 27 (0.0–51)	43 (19–71) 58 (50–89)	52 (24–76) 137 (110–243)	44 (18–78)	18 (8.7–30)
VEGF *	47 (34–62) 44 (31–54)	40 (25–60) 25 (18–34)	65 (40–94) 71 (59–85)	84 (50–108) 116 (103–245)	50 (30–77)	30 (20–39)
PDGF–β *	466 (210–878) 456 (382–591)	430 (253–866) 416 (365–474)	754 (350–1090) 704 (560–927)	910 (505–1223) 1210 (876–1600)	571 (353–928)	258 (164–364)
CCL5 *	6147 (1881–7702) 4862 (3086–5822)	6394 (3993–8916) 4162 (2612–5949)	6911 (4736–8942) 8011 (6476–10104)	6259 (5107–7811) 12032 (9331–15010)	6176 (3772–8707)	2863 (2380–3543)

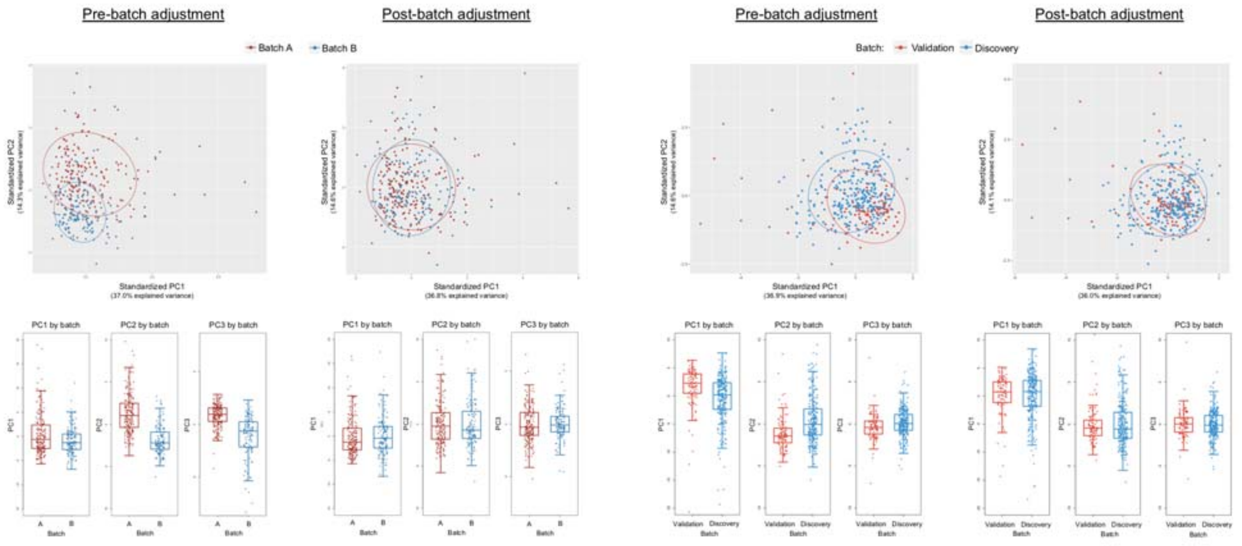
SUPPLEMENTAL FIGURES

Figure I. Adjustment for plate and batch effects. Principal component analysis (PCA) plots show the impact of 'ComBat' algorithm adjustment on **[A]** plate effect in discovery cohort batch A (five plates), **[B]** plate effect in discovery cohort batch B (four plates), and **[C]** batch effect in discovery cohort (two batches), and **[D]** batch effect between the discovery and validation cohort. Both before and after ComBat adjustment, scatter plots of the first two principal components are displayed (PC1 vs PC2). In these plots, individual patient samples are represented by dots and color-coded according to their assay plate or batch of origin. Ellipses surrounding each colored group indicate 68% confidence regions (one standard deviation assuming a normal distribution). In addition, boxplots show distributions of the first three principal components (PC1, PC2, PC3) for each assay plate/batch before and after adjustment. Post-ComBat adjustment the plates and batches demonstrated more proteomic homogeneity, as evidenced by increased overlap in PCA scatter plots and similar distributions across the PC1-PC3 boxplots.



[A] Adjustment for plate effect within discovery cohort batch A

[B] Adjustment for plate effect within discovery cohort batch B



[C] Adjustment for discovery cohort batch effect

[D] Adjustment for batch effect between discovery and validation cohorts

Figure II. Discovery cohort consensus clustering: measuring consensus and determining the number of clusters (*k optima*) (k-medoids algorithm, Euclidean distance). **[A]** Heatmaps of consensus matrices for *k*=3 to *k*=8. Each heatmap represents the symmetric *consensus matrix* for a given *k*, with the same patients as rows and columns. Patients are grouped by consensus cluster membership, which is specified by horizontal colored bars atop each heatmap. The heatmaps show pairwise *consensus index* values over a spectrum from 0 (white= never clustered together across resampling runs) to 1 (dark blue= always clustered together). Perfect cluster assignment consensus across runs would appear as solid cluster-specific dark blue blocks along the diagonal on a white-only background. Thus, the highest degree of cluster stability was observed for *k*=4. **[B]** Consensus cumulative distribution functions (CDFs) for each *k* (from *k*=2 to *k*=8). Each consensus CDF is estimated by a histogram containing 100 bins and visually represents how entries in the respective *consensus matrix* are cumulatively distributed within the 0-1 range. The plot provides graphically represents the clustering stability across resampling runs for each *k,* and can be utilized to determine the *k* optima. If perfect clustering consensus were achieved at a given *k* (consensus matrix only containing 0's and 1's), the consensus CDF would be purely bimodal with a flat horizontal mid-portion between modes at 0 and 1 (a step at 0, flat line between 0 and 1, and another step at 1). The area under CDFs will increase across values of *k* until the *k* optima is reached, while beyond *k* optima the area under CDFs will not markedly increase. Based on the CDF plot shown, our PAH proteomic dataset appears to have a *k*=4 optima. The consensus CDF at *k*=4 has the most bimodal appearance, and from *k*=4 to *k*=5 the area under CDFs do not increase significantly. **[C]** Average intra-cluster *consensus index* values. For the consensus clusters generated at each *k,* a bar plot shows the mean consensus index value between all pairs of patients within each cluster. Each colored bar represents a cluster. Clusters at the *k*=4 optima had the highest average consensus values.

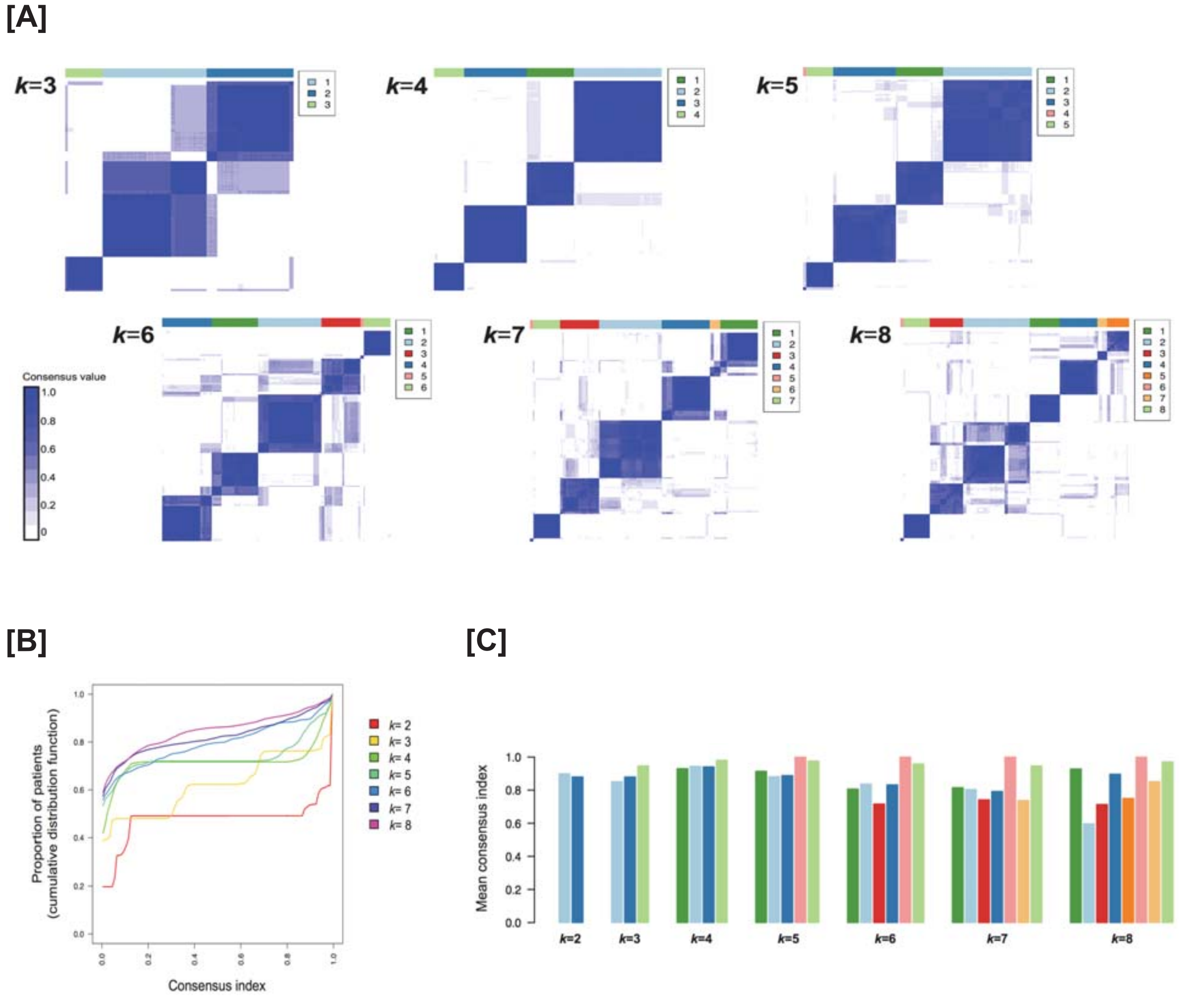


Figure III. Discovery cohort: saturated proteomic networks by PAH cluster. Protein-protein relationships are displayed for each PAH cluster with weighted partial correlation networks constructed as force-directed graphs. While the 'sparse core networks' highlighted in Figure 3A of the main manuscript reflected removal of 'spurious' false positive edges by regularization, each saturated network below retains all possible protein-protein interactions (1128 edges). Network nodes correspond to individual proteins, and their size is proportional to the mean plasma expression level in the cluster. Protein-protein network edges have weights that are proportional to the magnitude of partial correlation (red= positive, blue= negative).

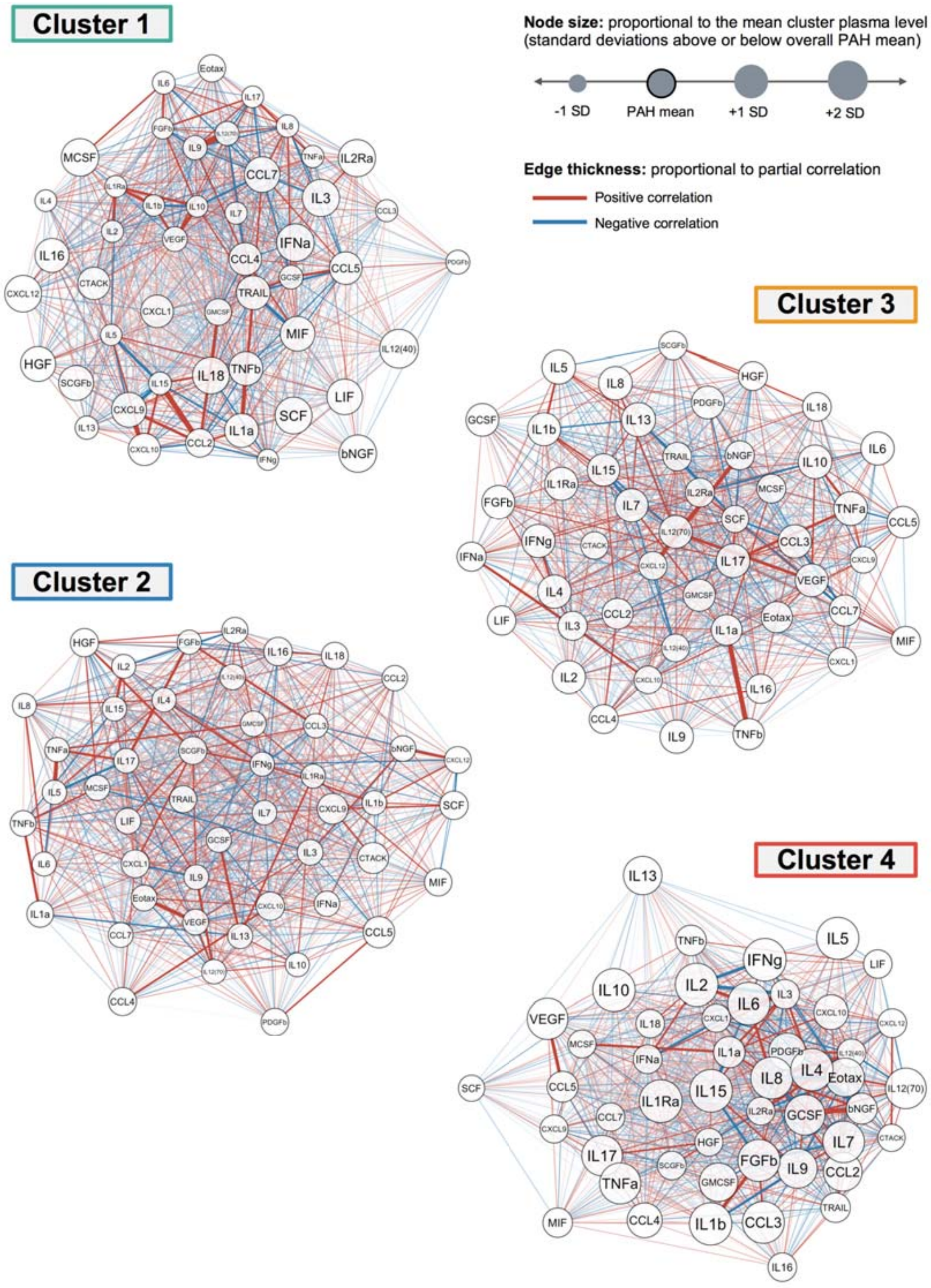


Figure IV. Discovery cohort: network node centrality measures and plasma expression within each PAH cluster. Paneled dot plots show calculated indices of network centrality (strength, closeness, and betweenness as defined in supplemental methods) and plasma expression levels for individual proteins in the proteomic partial correlation networks of **[A]** cluster 1, **[B]** cluster 2, **[C]** cluster 3, and **[D]** cluster 4. Centrality indices for each protein node are displayed as z-scores relative to other nodes in the cluster network. The mean plasma expression levels for each protein by cluster are z-scored relative to the overall PAH cohort. Proteins are grouped in each cluster according to the pattern of network centrality and plasma expression (see inset legend). Nodes highlighted in red have high network centrality measures (at least 2 of 3 indices above the cluster mean) and are upregulated (greater than overall PAH cohort mean).

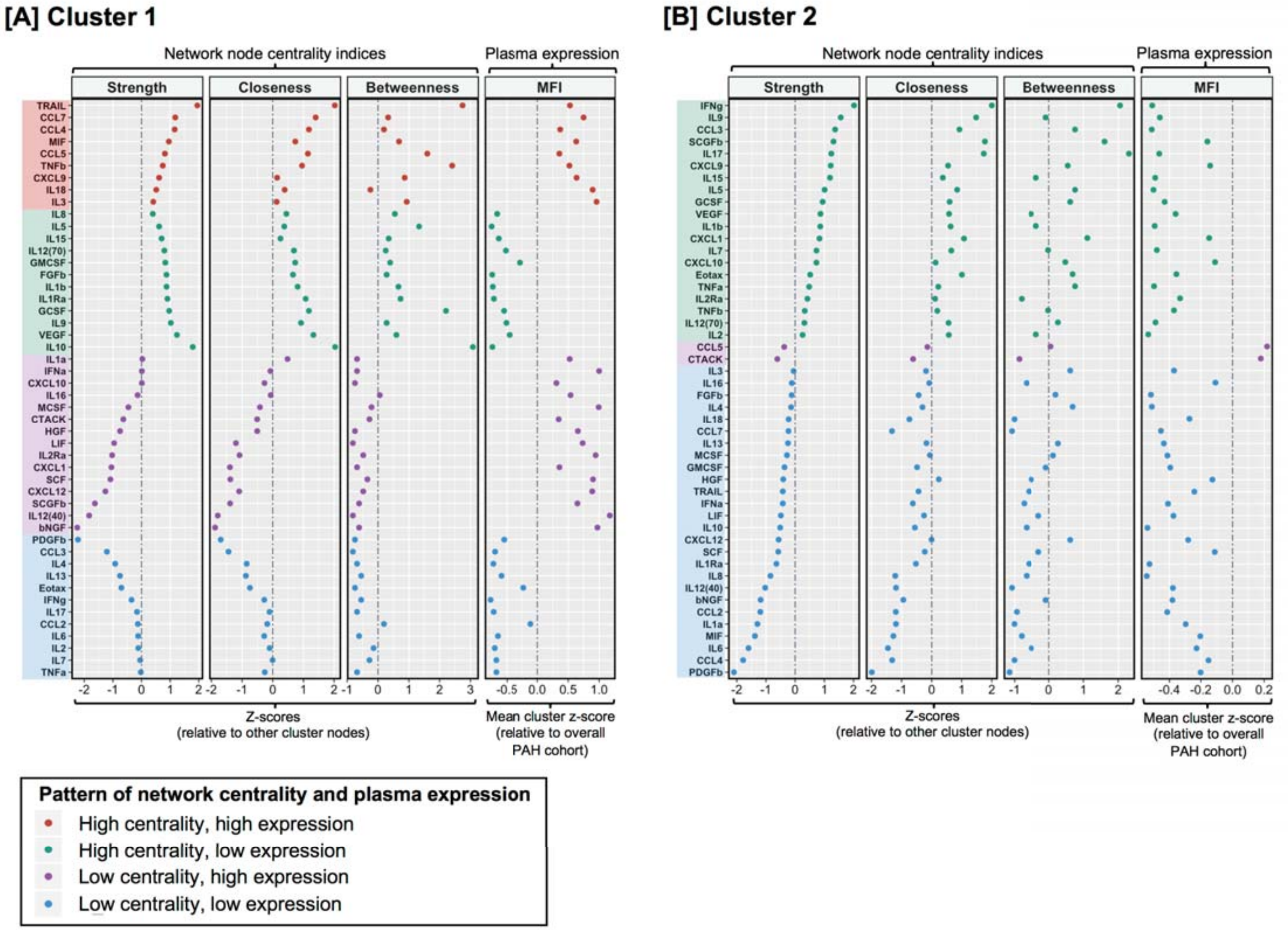


Figure IV (continued). Discovery cohort: network node centrality measures and plasma expression within each PAH cluster.

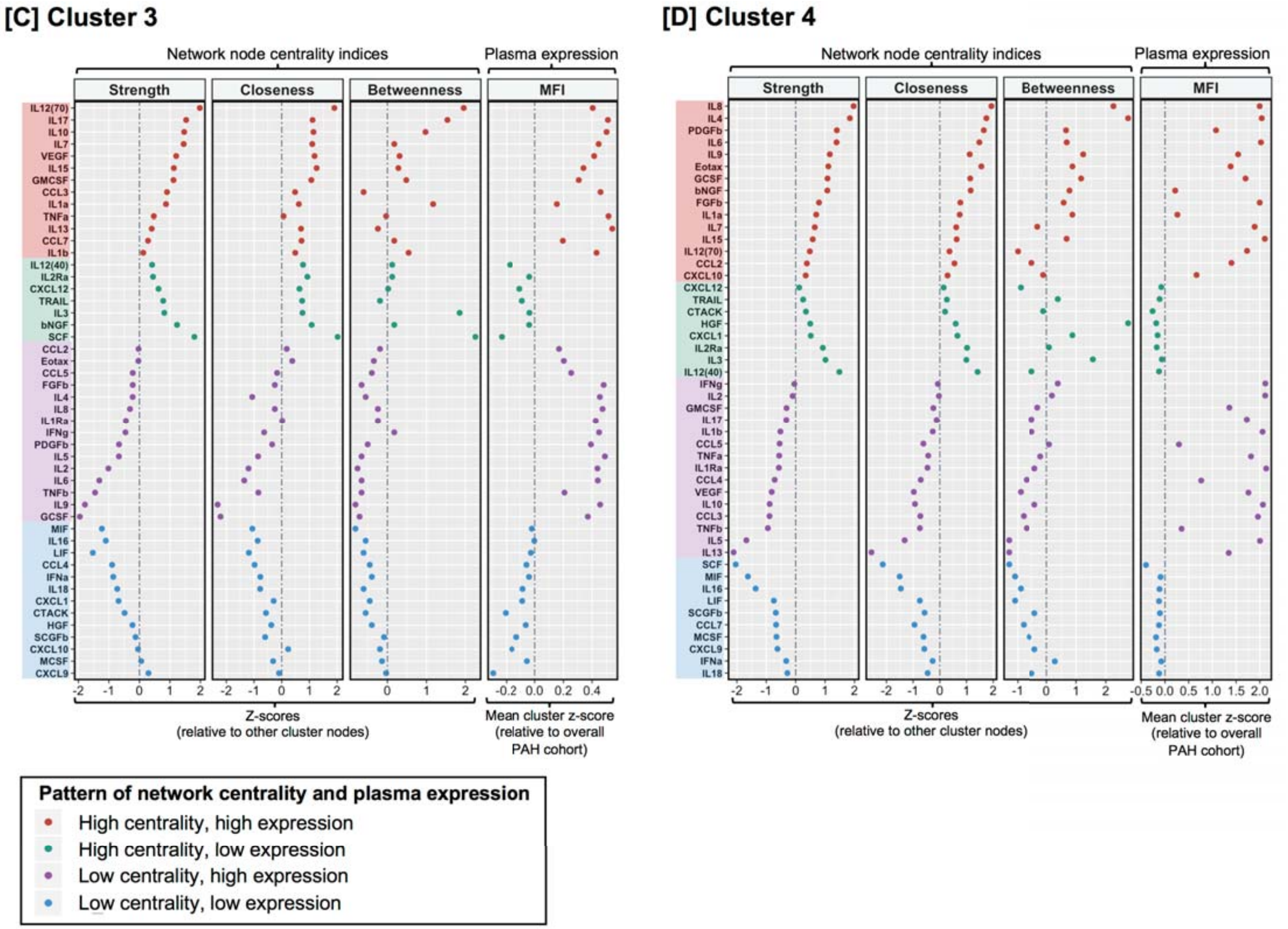
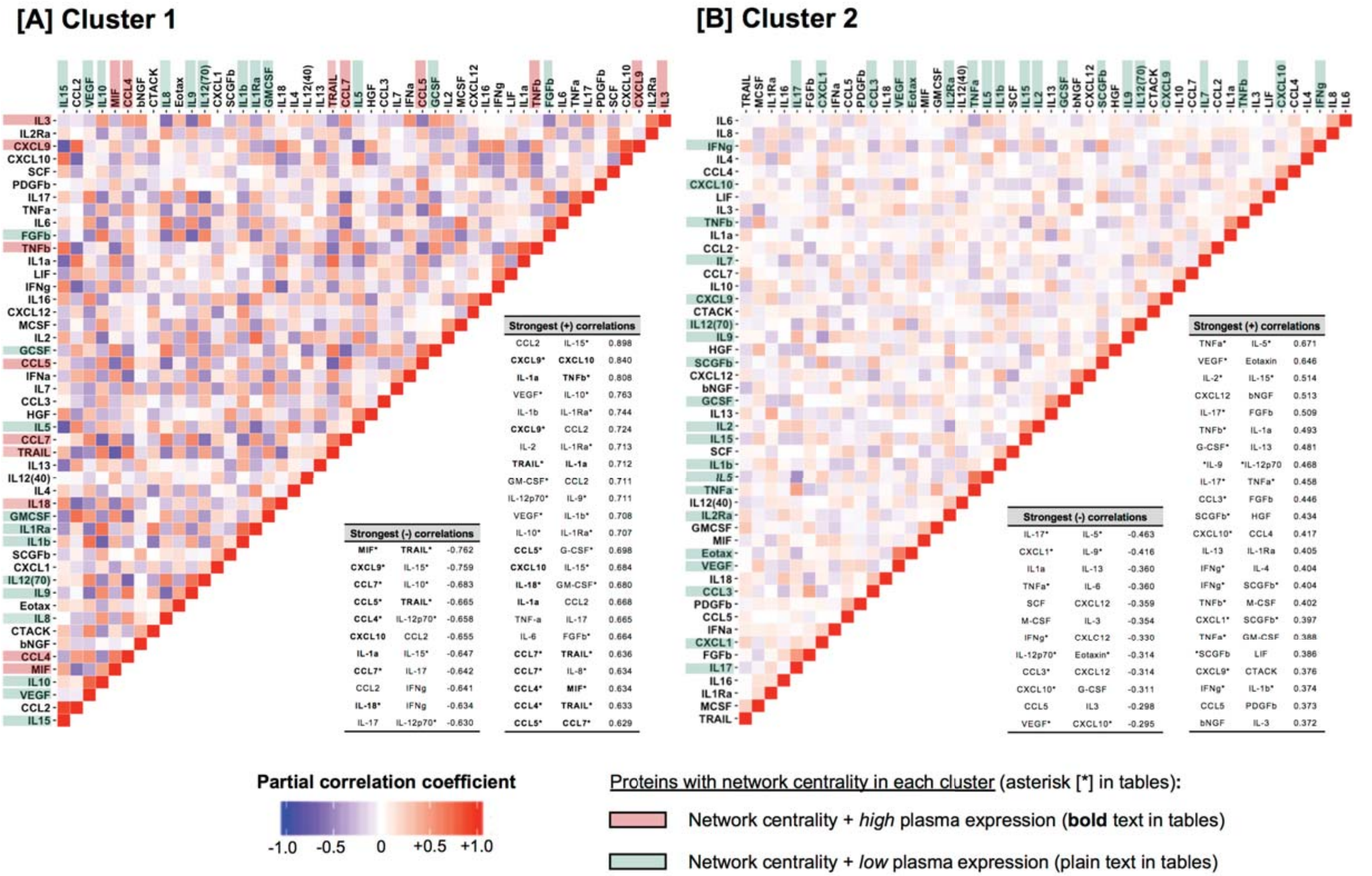


Figure V. Discovery cohort: pairwise protein-protein relationships within each PAH cluster. Partial correlation heatmaps are shown for **[A]** cluster 1, **[B]** cluster 2, **[C]** cluster 3, and **[D]** cluster 4. Each individual matrix cell represents the partial correlation coefficient between two specified proteins (their direct interaction conditioned on all other measured variables). The heatmaps are accompanied by tables that indicate the protein pairs with the strongest positive and negative correlations by cluster. For each cluster, these protein-protein relationships provided the basis for construction of partial correlation networks as force-directed graphs (Online Figure III, Figure 3A). Proteins with network centrality (as determined in Online Figure IV) are highlighted (see inset legend).



[A] Cluster 1

Strongest (+) correlations		
CCL2	IL-15*	0.898
CXCL9*	**CXCL10**	0.840
IL-1a	**TNFb***	0.808
VEGF*	IL-10*	0.763
IL-1b	IL-1Ra*	0.744
CXCL9*	CCL2	0.724
IL-2	IL-1Ra*	0.713
TRAIL*	**IL-1a**	0.712
GM-CSF*	CCL2	0.711
IL-12p70*	IL-9*	0.711
VEGF*	IL-1b*	0.708
IL-10*	IL-1Ra*	0.707
CCL5*	G-CSF*	0.698
CXCL10	IL-15*	0.684
IL-18*	GM-CSF*	0.680
IL-1a	CCL2	0.668
TNF-a	IL-17	0.665
IL-6	FGFb*	0.664
CCL7*	**TRAIL***	0.636
CCL7*	IL-8*	0.634
CCL4*	**MIF***	0.634
CCL4*	**TRAIL***	0.633
CCL5*	**CCL7***	0.629

Strongest (-) correlations		
MIF*	**TRAIL***	-0.762
CXCL9*	IL-15*	-0.759
CCL7*	IL-10*	-0.683
CCL5*	**TRAIL***	-0.665
CCL4*	IL-12p70*	-0.658
CXCL10	CCL2	-0.655
IL-1a	IL-15*	-0.647
CCL7*	IL-17	-0.642
CCL2	IFNg	-0.641
IL-18*	IFNg	-0.634
IL-17	IL-12p70*	-0.630

[B] Cluster 2

Strongest (+) correlations		
TNFa*	IL-5*	0.671
VEGF*	Eotaxin	0.646
IL-2*	IL-15*	0.514
CXCL12	bNGF	0.513
IL-17*	FGFb	0.509
TNFb*	IL-1a	0.493
G-CSF*	IL-13	0.481
*IL-9	*IL-12p70	0.468
IL-17*	TNFa*	0.458
CCL3*	FGFb	0.446
SCGFb*	HGF	0.434
CXCL10*	CCL4	0.417
IL-13	IL-1Ra	0.405
IFNg*	IL-4	0.404
IFNg*	SCGFb*	0.404
TNFb*	M-CSF	0.402
CXCL1*	SCGFb*	0.397
TNFa*	GM-CSF	0.388
*SCGFb	LIF	0.386
CXCL9*	CTACK	0.376
IFNg*	IL-1b*	0.374
CCL5	PDGFb	0.373
bNGF	IL-3	0.372

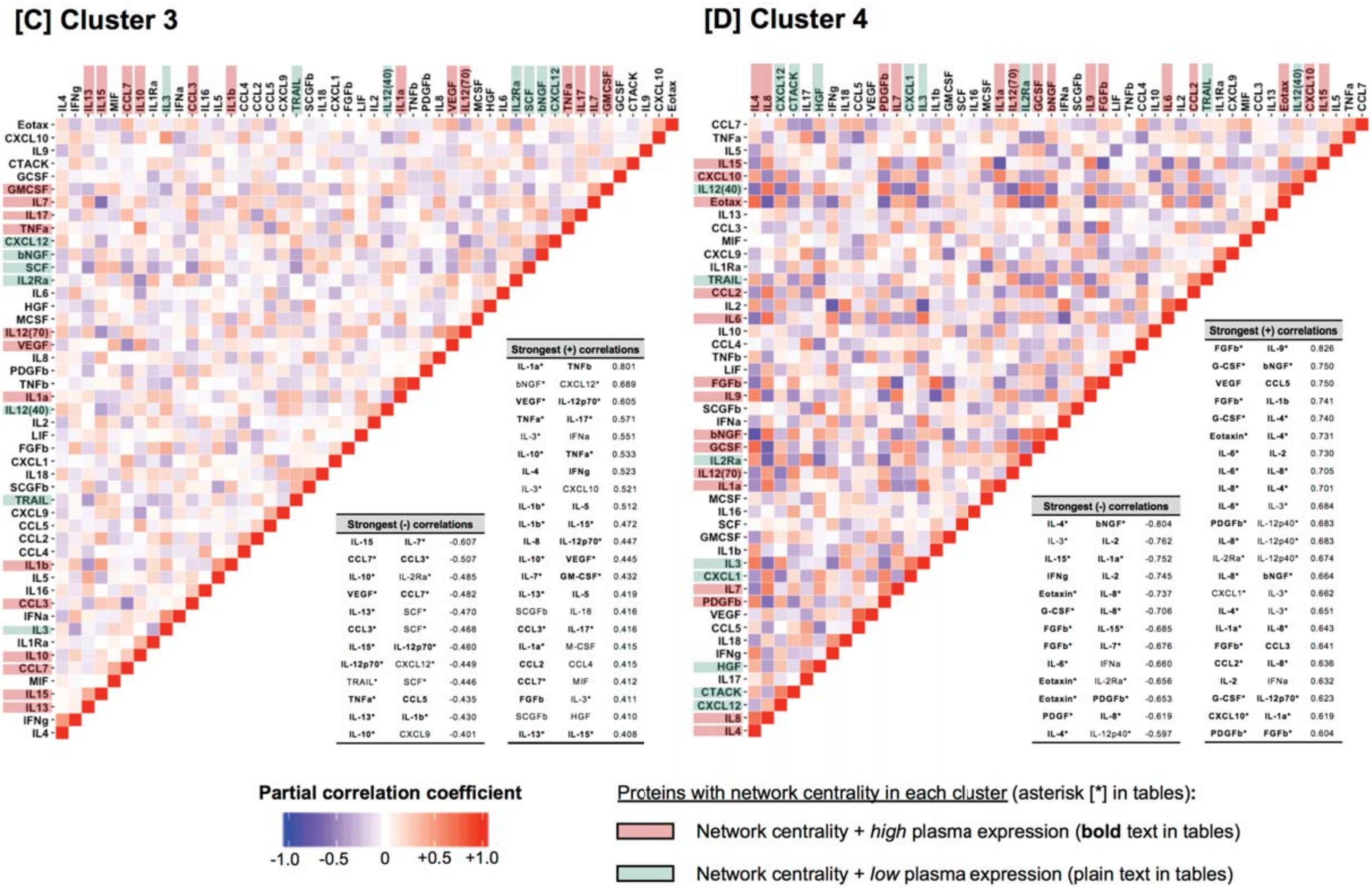
Strongest (-) correlations		
IL-17*	IL-5*	-0.463
CXCL1*	IL-9*	-0.416
IL1a	IL-13	-0.360
TNFa*	IL-6	-0.360
SCF	CXCL12	-0.359
M-CSF	IL-3	-0.354
IFNg*	CXLC12	-0.330
IL-12p70*	Eotaxin*	-0.314
CCL3*	CXCL12	-0.314
CXCL10*	G-CSF	-0.311
CCL5	IL3	-0.298
VEGF*	CXCL10*	-0.295

Proteins with network centrality in each cluster (asterisk [*] in tables):

🟥 Network centrality + *high* plasma expression (**bold** text in tables)

🟩 Network centrality + *low* plasma expression (plain text in tables)

Figure V (continued). Discovery cohort: pairwise protein-protein relationships within machine learned PAH clusters.



[C] Cluster 3

Strongest (+) correlations		
IL-1a*	**TNFb**	0.801
bNGF*	CXCL12*	0.689
VEGF*	**IL-12p70***	0.605
TNFa*	**IL-17***	0.571
IL-3*	IFNa	0.551
IL-10*	**TNFa***	0.533
IL-4	**IFNg**	0.523
IL-3*	CXCL10	0.521
IL-1b*	**IL-5**	0.512
IL-1b*	**IL-15***	0.472
IL-8	**IL-12p70***	0.447
IL-10*	**VEGF***	0.445
IL-7*	**GM-CSF***	0.432
IL-13*	**IL-5**	0.419
SCGFb	IL-18	0.416
CCL3*	**IL-17***	0.416
IL-1a*	M-CSF	0.415
CCL2	CCL4	0.415
CCL7*	MIF	0.412
FGFb	IL-3*	0.411
SCGFb	HGF	0.410
IL-13*	**IL-15***	0.408

Strongest (-) correlations		
IL-15	**IL-7***	-0.607
CCL7*	**CCL3***	-0.507
IL-10*	IL-2Ra*	-0.485
VEGF*	**CCL7***	-0.482
IL-13*	SCF*	-0.470
CCL3*	SCF*	-0.468
IL-15*	**IL-12p70***	-0.460
IL-12p70*	CXCL12*	-0.449
TRAIL*	SCF*	-0.446
TNFa*	**CCL5**	-0.435
IL-13*	**IL-1b***	-0.430
IL-10*	CXCL9	-0.401

[D] Cluster 4

Strongest (+) correlations		
FGFb*	**IL-9***	0.826
G-CSF*	**bNGF***	0.750
VEGF	**CCL5**	0.750
FGFb*	**IL-1b**	0.741
G-CSF*	**IL-4***	0.740
Eotaxin*	**IL-4***	0.731
IL-6*	**IL-2**	0.730
IL-6*	**IL-8***	0.705
IL-8*	**IL-4***	0.701
IL-6*	IL-3*	0.684
PDGFb*	IL-12p40*	0.683
IL-8*	IL-12p40*	0.683
IL-2Ra*	IL-12p40*	0.674
IL-8*	**bNGF***	0.664
CXCL1*	IL-3*	0.662
IL-4*	IL-3*	0.651
IL-1a*	**IL-8***	0.643
FGFb*	**CCL3**	0.641
CCL2*	**IL-8***	0.636
IL-2	IFNa	0.632
G-CSF*	**IL-12p70***	0.623
CXCL10*	**IL-1a***	0.619
PDGFb*	**FGFb***	0.604

Strongest (-) correlations		
IL-4*	**bNGF***	-0.804
IL-3*	**IL-2**	-0.762
IL-15*	**IL-1a***	-0.752
IFNg	**IL-2**	-0.745
Eotaxin*	**IL-8***	-0.737
G-CSF*	**IL-8***	-0.706
FGFb*	**IL-15***	-0.685
FGFb*	**IL-7***	-0.676
IL-6*	IFNa	-0.660
Eotaxin*	IL-2Ra*	-0.656
Eotaxin*	**PDGFb***	-0.653
PDGF*	**IL-8***	-0.619
IL-4*	IL-12p40*	-0.597

Proteins with network centrality in each cluster (asterisk [*] in tables):

🟥 Network centrality + *high* plasma expression (**bold** text in tables)

🟩 Network centrality + *low* plasma expression (plain text in tables)

Figure VI. Discovery cohort: comparison of circulating total white blood cell (WBC) count and differential cell subsets by PAH immune cluster. Boxplots compare circulating total white blood cell count and both the percent and absolute count of each cell subset. Patient-level measurements are shown by dots, boxes are colored according to immune cluster and represent the 25-75% interquartile range (IQR), bold horizontal lines indicate cluster medians, and whiskers delineate data within the 1.5 x IQR range. The Kruskal-Wallis test was used to test for differences across clusters and p-values are shown. Analysis included patients who had WBC measurements within one month of proteomic sampling [n=270/281 (96.1%), where WBC and proteomic data sampling occurred on the same day for 219 (81.1%) and within 3 days for 234 (86.3%)]. WBC/differential data was missing for n=4, 2, 4, and 1 subject(s) in clusters 1-4, respectively.

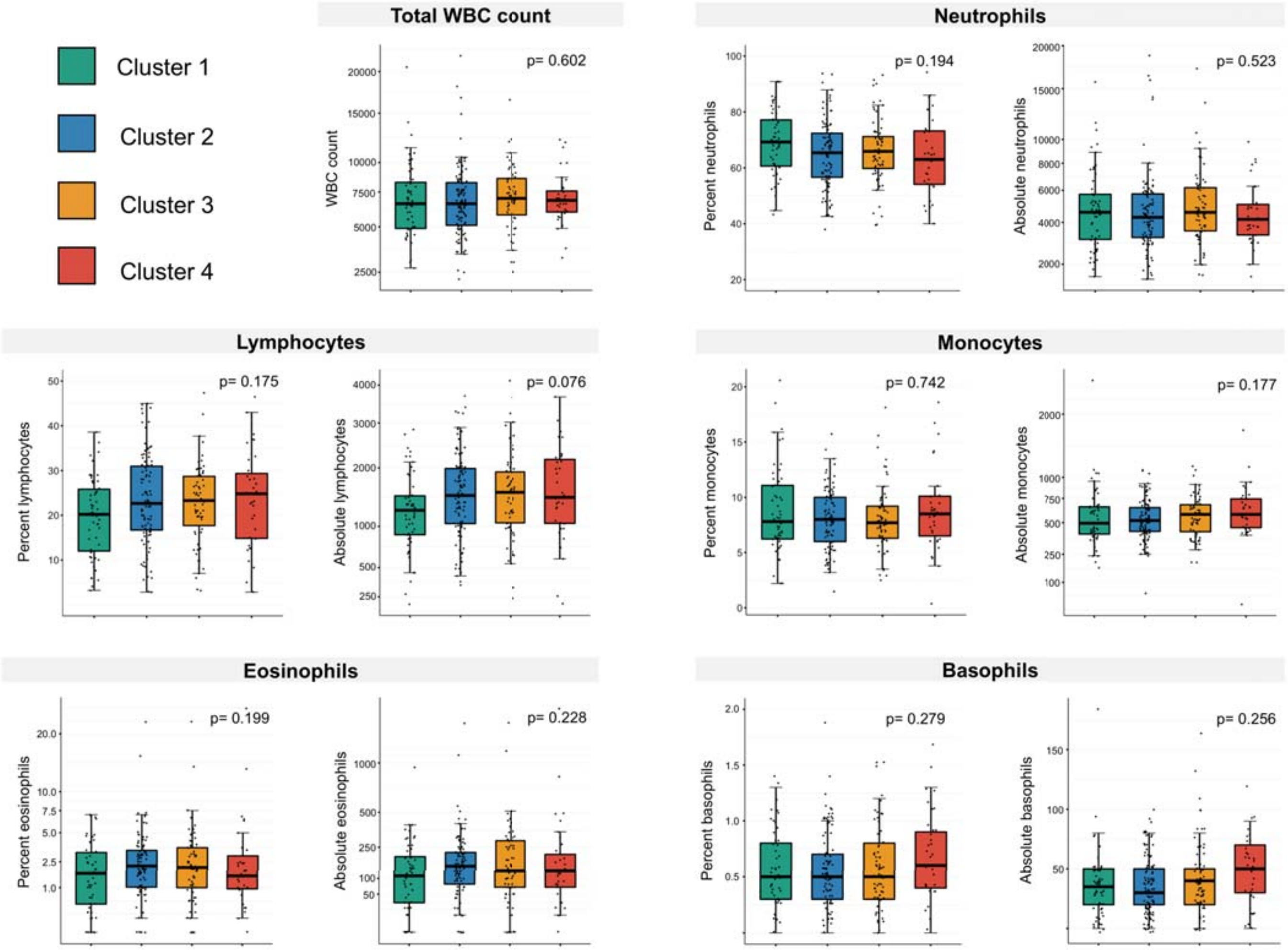


Figure VII. Discovery cohort: standardized patient-level proteomic measurements by PAH etiology. Heatmap panels show patients grouped by PAH etiology. Each column is a patient, and rows are individual measured proteins (displayed in the same order as that specified in Figure 2A). Standardized protein MFI measurements are color-coded according to z-score. Within each etiology panel, patients are sorted by assigned proteomic-based cluster (from cluster 1 at left to cluster 4 at right). Abbreviations: APAH= associated pulmonary arterial hypertension, CHD= congenital heart disease, CTD= connective tissue disease, D&T= drugs and toxins, HPAH= hereditary PAH, IPAH= idiopathic PAH, PoPH= portopulmonary hypertension.

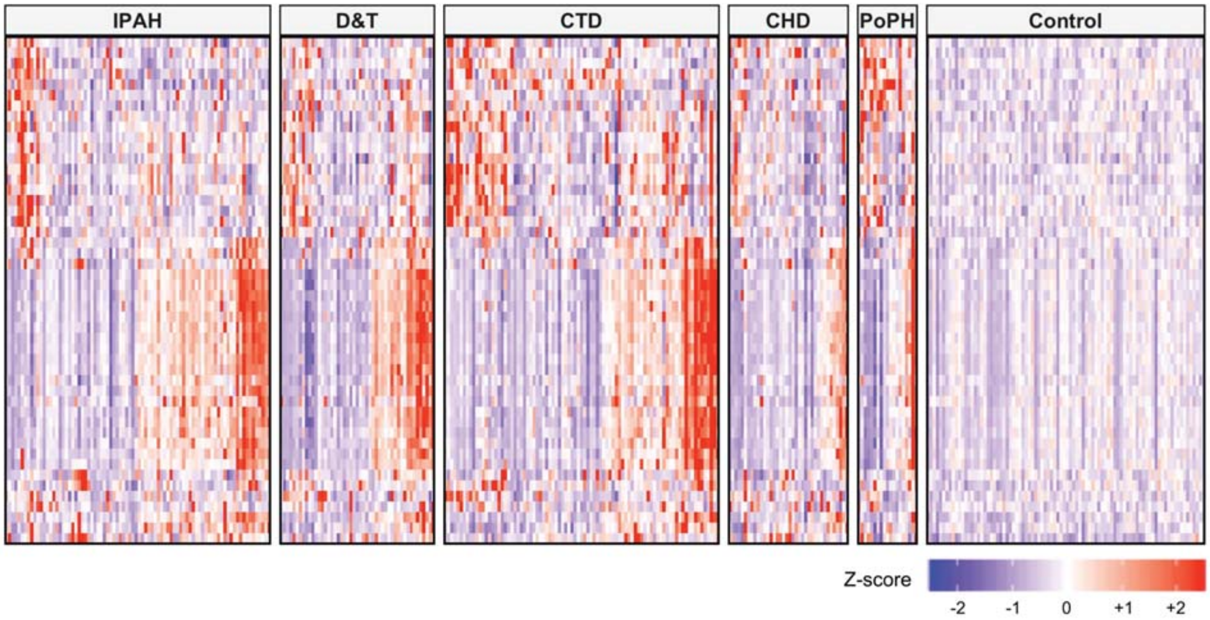


Figure VIII. Discovery cohort: timing of proteomic sample collection in proteomic clusters relative to PAH diagnosis and symptom onset. Boxplots display the patient distribution within each cluster for **[A]** time from PAH diagnosis to plasma sample collection and **[B]** time from patient-reported symptom onset to plasma sample collection. Patient-level data is shown by dots, boxes represent the 25-75% interquartile range (IQR), bold vertical lines indicate medians, and whiskers delineate data within 1.5 x IQR bounds.

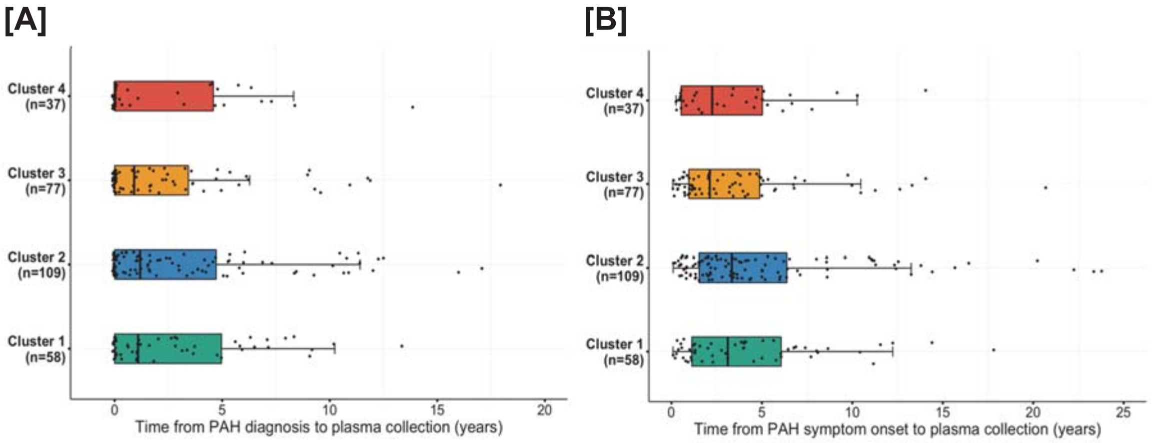


Figure IX. Discovery cohort: principal component analysis of proteomic immune profiles according to background PAH-guided therapy. In scatter plots of the first and second principal components (PC1 vs. PC2), the multi-dimensional proteomic profile of each PAH patient is reduced to a single point and colored according to: **[A]** PAH therapy status (treatment naïve vs. any background pulmonary vasodilator agent), and class of therapies including **[B]** phosphodiesterase-5 inhibitors, **[C]** endothelin receptor antagonists, and **[D]** prostanoids. Patient proteomic profiles do not appear to separate on the basis of treatment status or any class of PAH therapy.

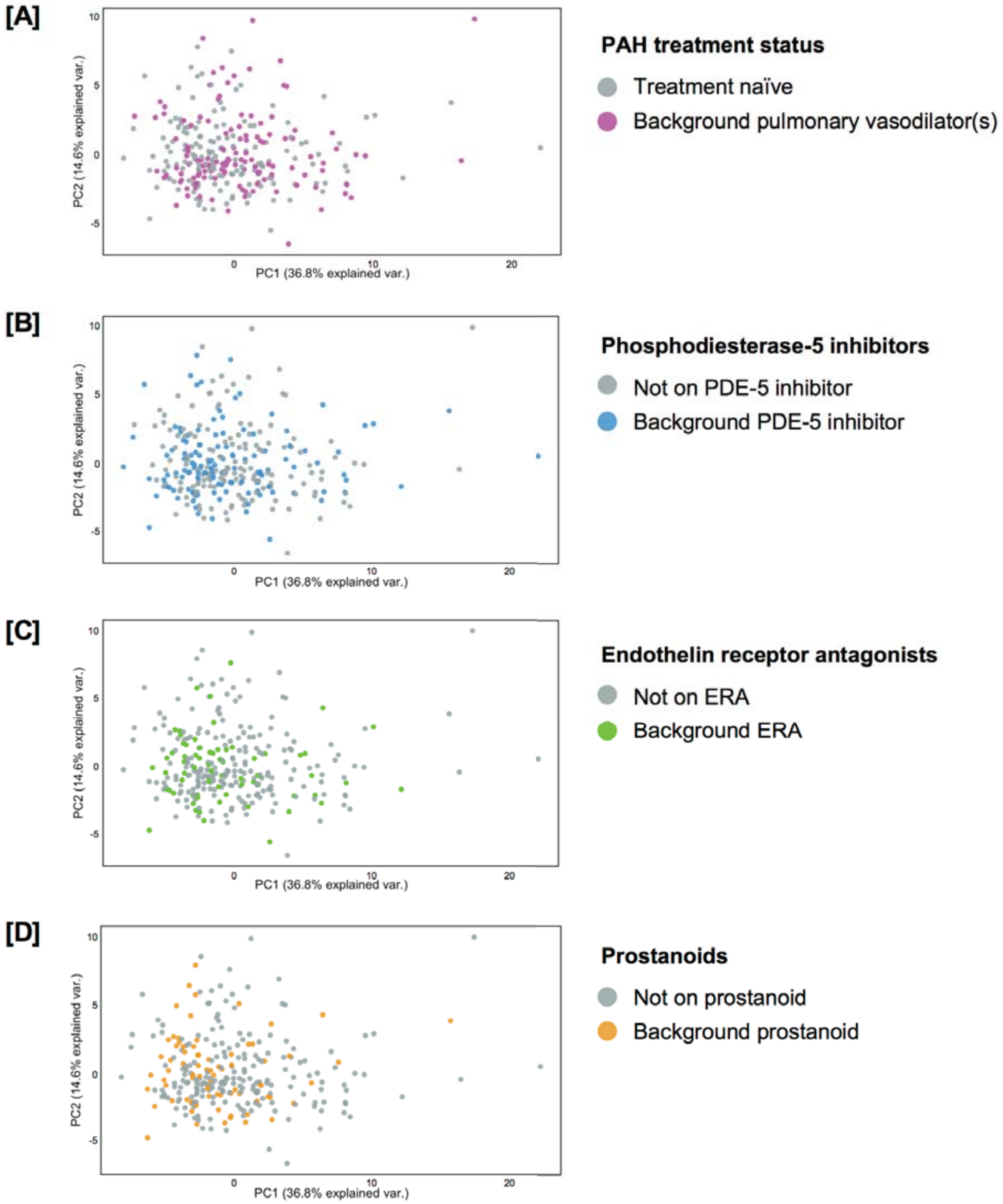


Figure X. Discovery cohort: subgroup analysis according to immune modulating therapy (IMT) status.

[A] Principal component analysis of proteomic immune profiles by IMT status. In a scatter plot of the first two principal components (PC1 vs PC2), the proteomic profile of each PAH patient is reduced to a single point (purple= background IMT, gray= no IMT). Patient proteomic profiles do not appear to separate on the basis of IMT status, as subjects on IMT are homogeneously distributed amidst those not on IMT. **[B]** Re-application of consensus clustering to partition the subgroup of patients not on IMT (n=219) into 4 clusters (*k*=4) using the k-medoids algorithm, Euclidean distance metric, and 1000 resampling iterations (95% of subgroup per resampling run). All 219 patients partitioned into the same clusters as they were assigned when analysis was conducted for the entire cohort irrespective of IMT status. The heatmap shows standardized patient-level proteomic measurements by PAH cluster for this IMT-free subgroup analysis. Heatmap columns represent individual patients (grouped according to consensus clusters), and each row an assayed protein. Measured protein MFI is displayed as a color-coded z-score. Healthy control samples are shown for reference. **[C]** Principal component analysis of PAH cluster proteomic profiles (PC1 vs PC2 scatter plot) that derived from consensus clustering in the no IMT subgroup. Subfigures B and C demonstrate that IMT-free cluster profiles appear similar to those that were observed in total cohort analysis.

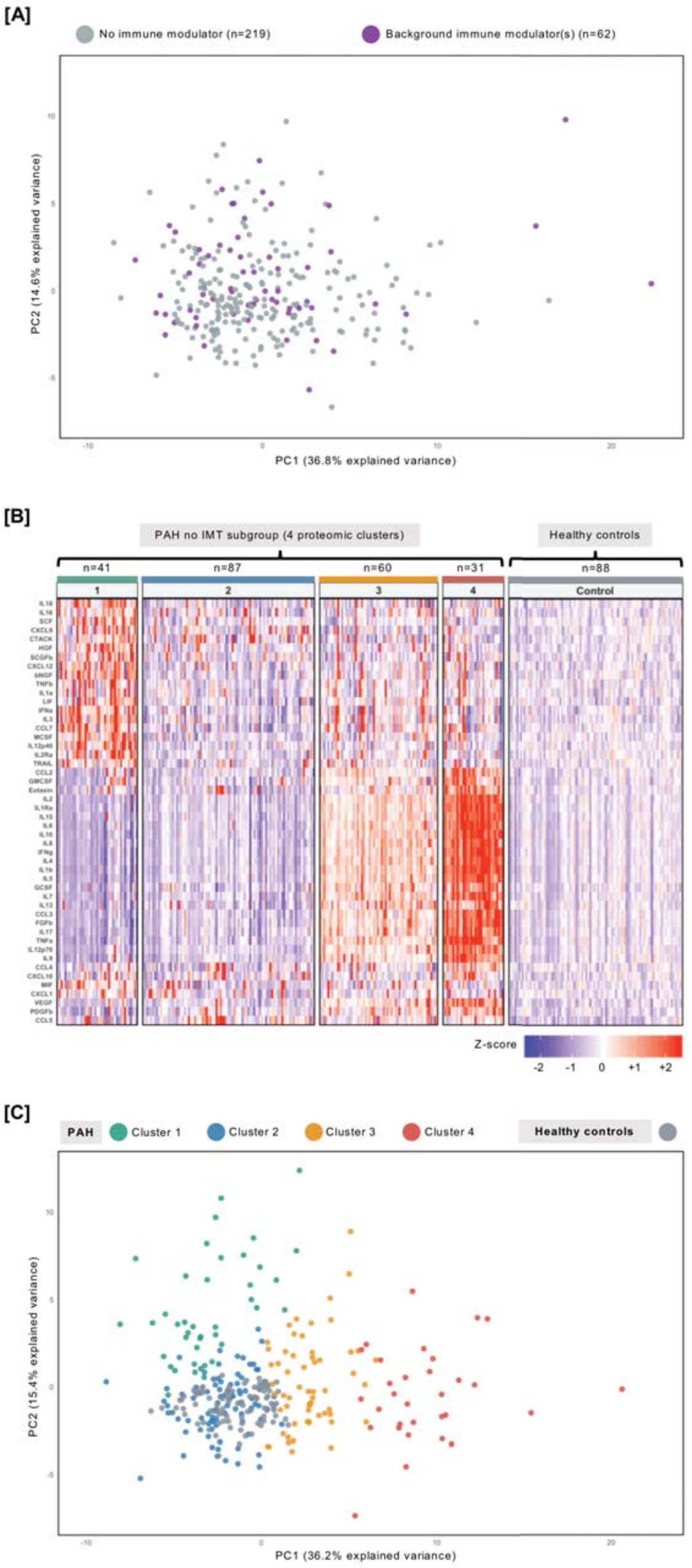


Figure XI. Discovery cohort: right heart catheterization hemodynamic profiles by proteomic immune cluster. Boxplots compare hemodynamic measurements across PAH clusters: **[A]** mean pulmonary arterial pressure, **[B]** pulmonary vascular resistance, **[C]** cardiac index, and **[D]** right atrial pressure. Patient-level data is shown by dots, boxes represent the 25-75% interquartile range (IQR), a bold horizontal line indicates the median, and whiskers delineate data within 1.5 x IQR bounds. The Kruskal-Wallis test was used to compare each variable across clusters, and the Dunn's test was applied for post-hoc pairwise comparisons.

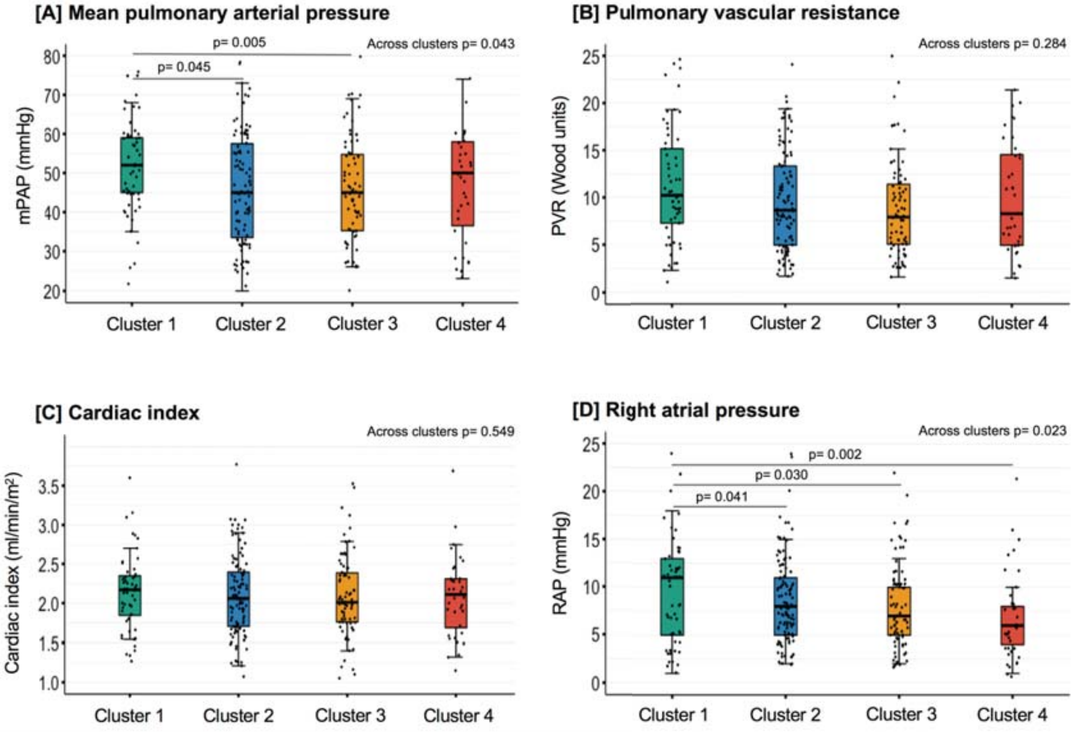


Figure XII. Discovery cohort: sensitivity analysis of cluster survival accounting for treatment status. Kaplan-Meier estimates of transplant-free survival are displayed by proteomic immune cluster for treatment naïve subjects (n=123). Cross-tags on survival curves indicate censoring, and the number of patients remaining at risk at each time point is shown. Transplant-free survival distributions are compared by log-rank test across clusters.

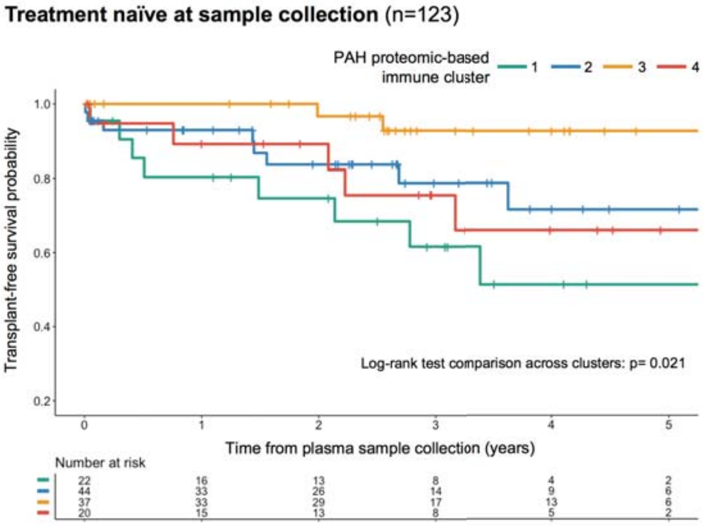
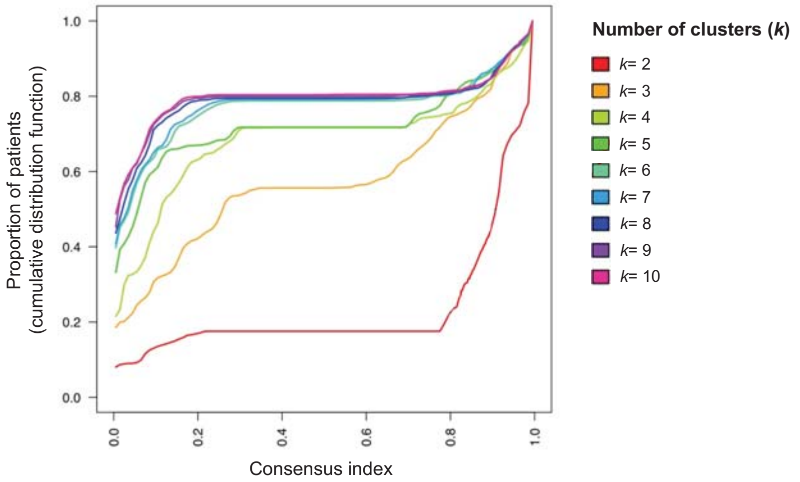


Figure XIII. Validation cohort consensus clustering: measuring consensus and determining the optimal number of clusters (*k* optima) (k-medoids algorithm, Euclidean distance). **[A]** Consensus cumulative distribution functions (CDFs) for each *k* (from *k*=2 to *k*=10). Refer to Online Figure IIB legend for general overview of CDF plots. Based on the CDF plot shown, a *k*=4 optima was identified. The consensus CDF at *k*=4 has a bimodal appearance, and from *k*=4 to *k*=5 the area under CDF curves does not increase significantly. **[B]** Heatmap of the consensus matrix for *k*=4. The heatmap represents the symmetric *consensus matrix*, with the same patients as rows and columns. Patients are grouped by consensus cluster membership, which is specified by the horizontal colored bar above the heatmap. Pairwise *consensus index* values are shown and color-coded over a spectrum from 0 (white= never clustered together across resampling runs) to 1 (dark blue= always clustered together). Perfect cluster assignment consensus across runs would appear as solid dark blue blocks along the diagonal on a white-only background.

[A]



[B]

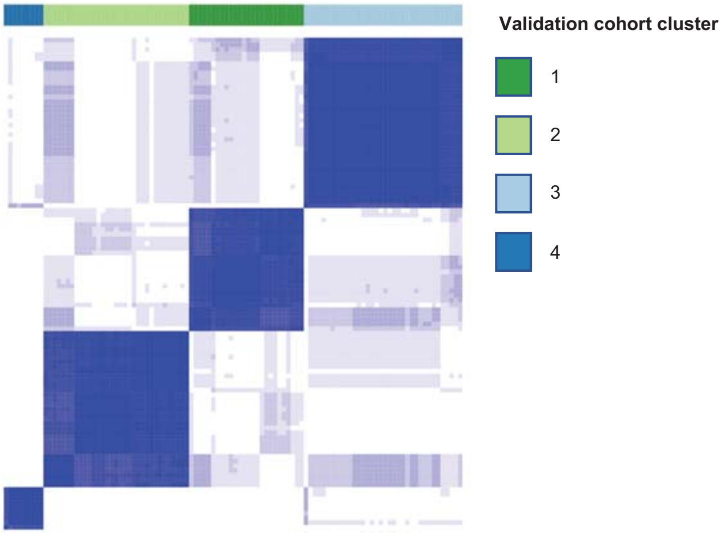
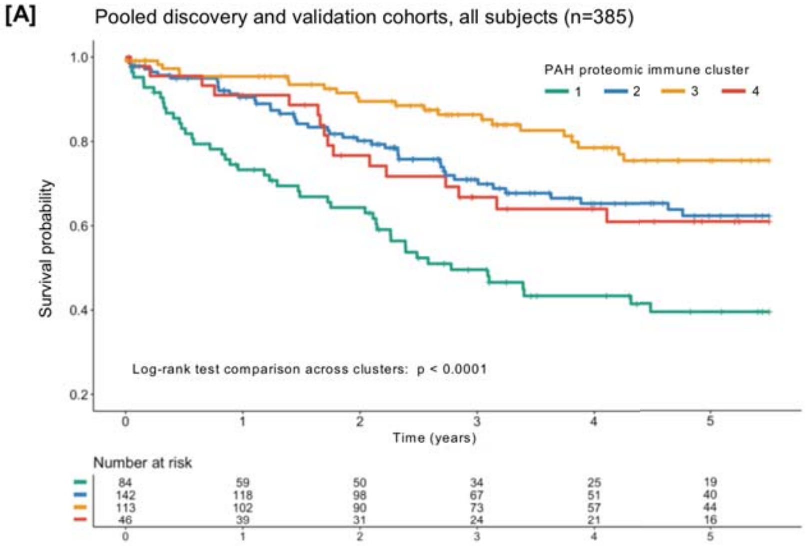


Figure XIV. Pooled discovery and validation cohort survival analysis. [A] Kaplan-Meier estimates of five-year survival are displayed by proteomic immune clusters for the combined pool of PAH patients (n=385). **[B]** Kaplan-Meier sensitivity analysis according to treatment status in the pooled cohort, where cluster survival estimates are limited to treatment naïve subjects (n=227). Cross-tags on survival curves indicate censoring, and the number of patients remaining at risk at each time point is shown. Survival distributions are compared across clusters by log-rank test.


[A] Pooled discovery and validation cohorts, all subjects (n=385)

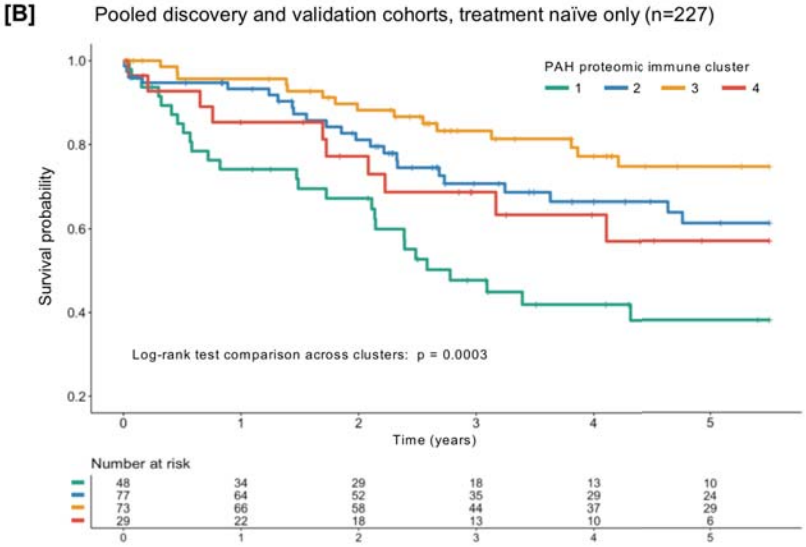

[B] Pooled discovery and validation cohorts, treatment naïve only (n=227)

Figure XV. Discovery cohort: plasma expression of individual cytokines, chemokines, and growth factors for PAH patients versus healthy controls. Boxplots show the log-transformed batch-adjusted median fluorescence intensity (MFI) detected for each protein, among PAH patients (red, n=281) and healthy controls (blue, n=88). Horizontal dashed lines indicate theoretical lower detection limits (two standard deviations above background MFI means). Patient-level data is shown by dots, boxes represent the 25-75% interquartile range (IQR), bold hortizontal lines indicate medians, and whiskers delineate data within 1.5 x IQR bounds. The Wilcoxon-ranked sum test was applied to compare PAH vs controls (displayed p-values are Bonferroni-adjusted to account for multiple testing).

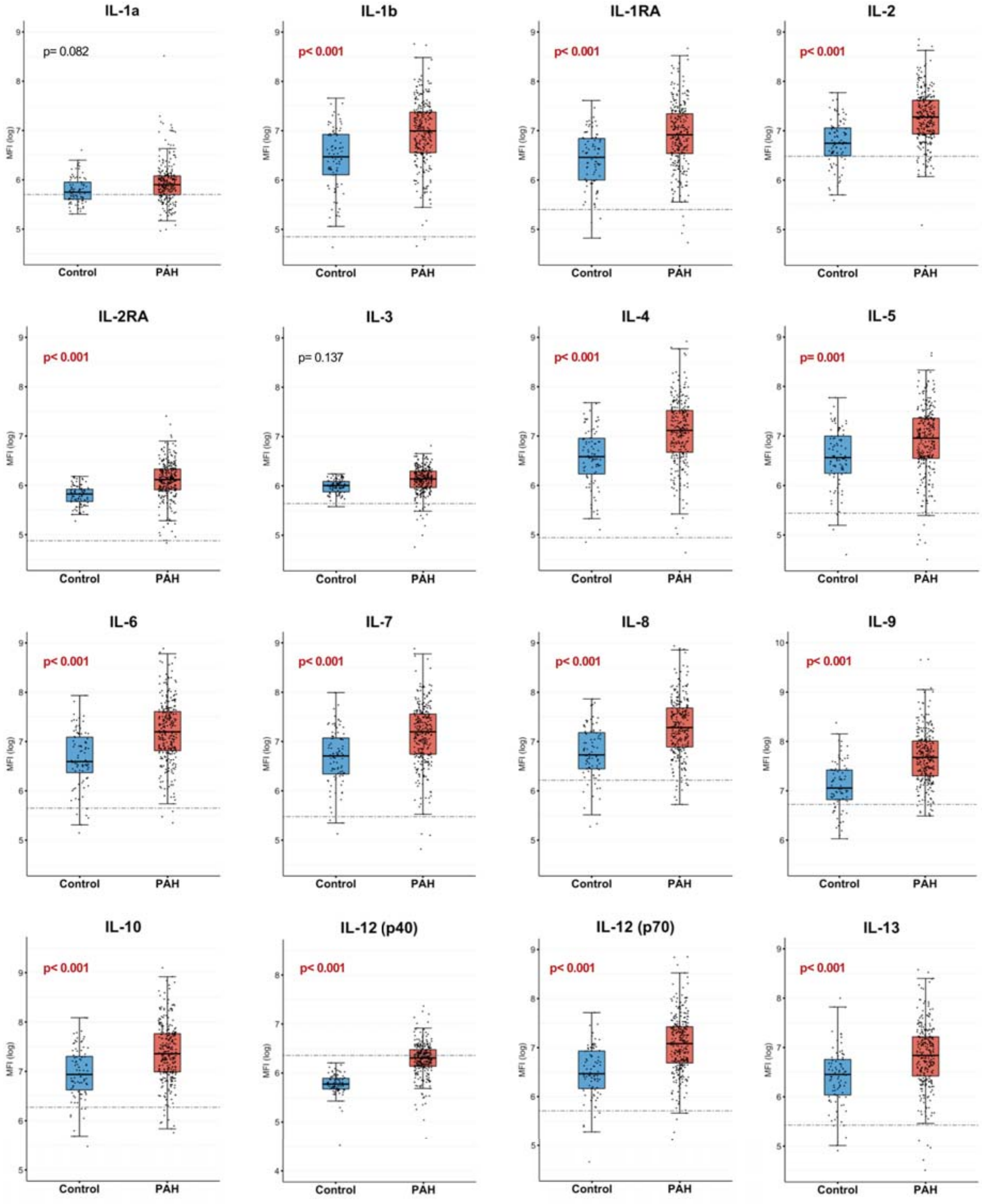


Figure XV (continued). Discovery cohort: plasma expression of individual cytokines, chemokines, and growth factors for PAH patients versus healthy controls.

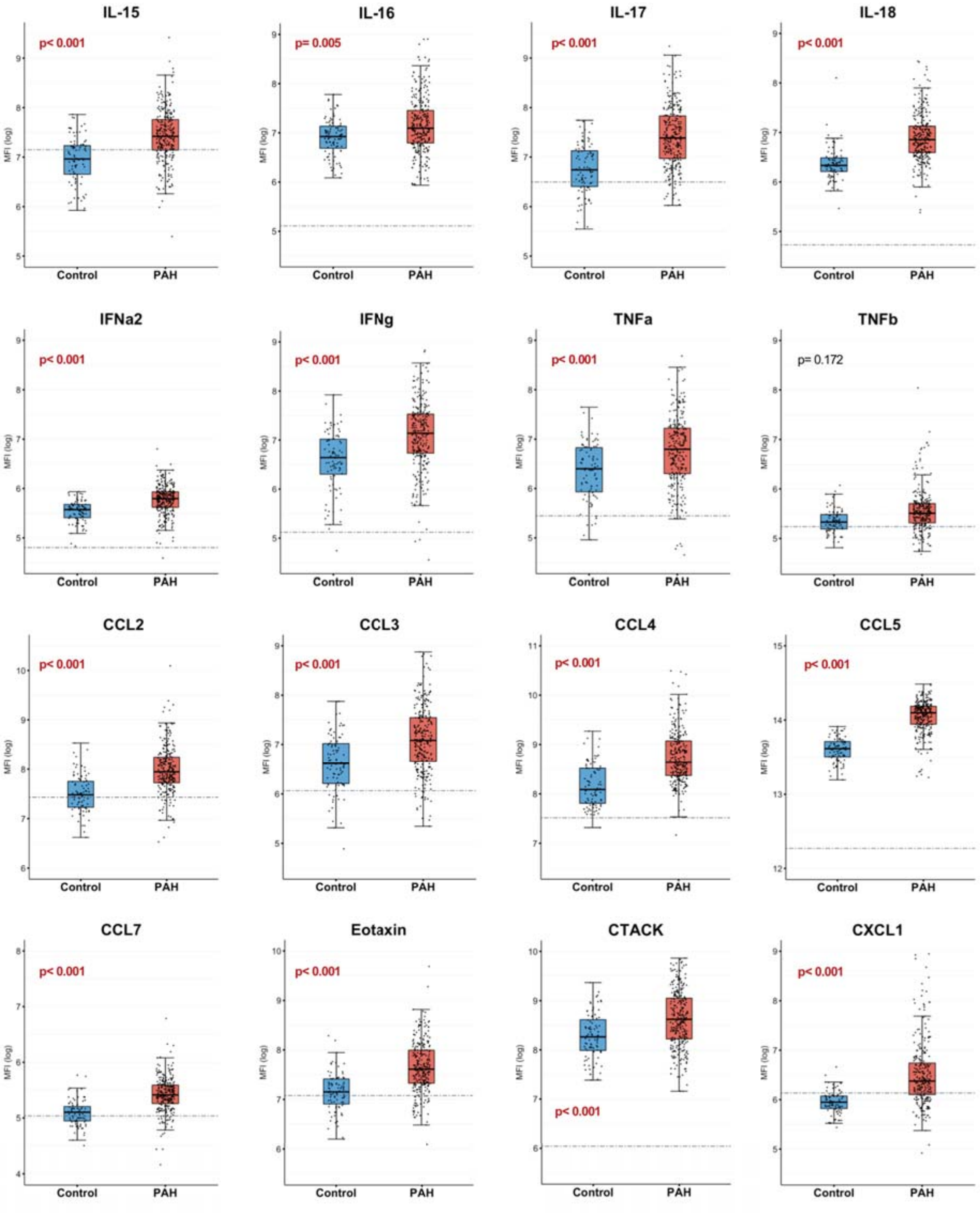
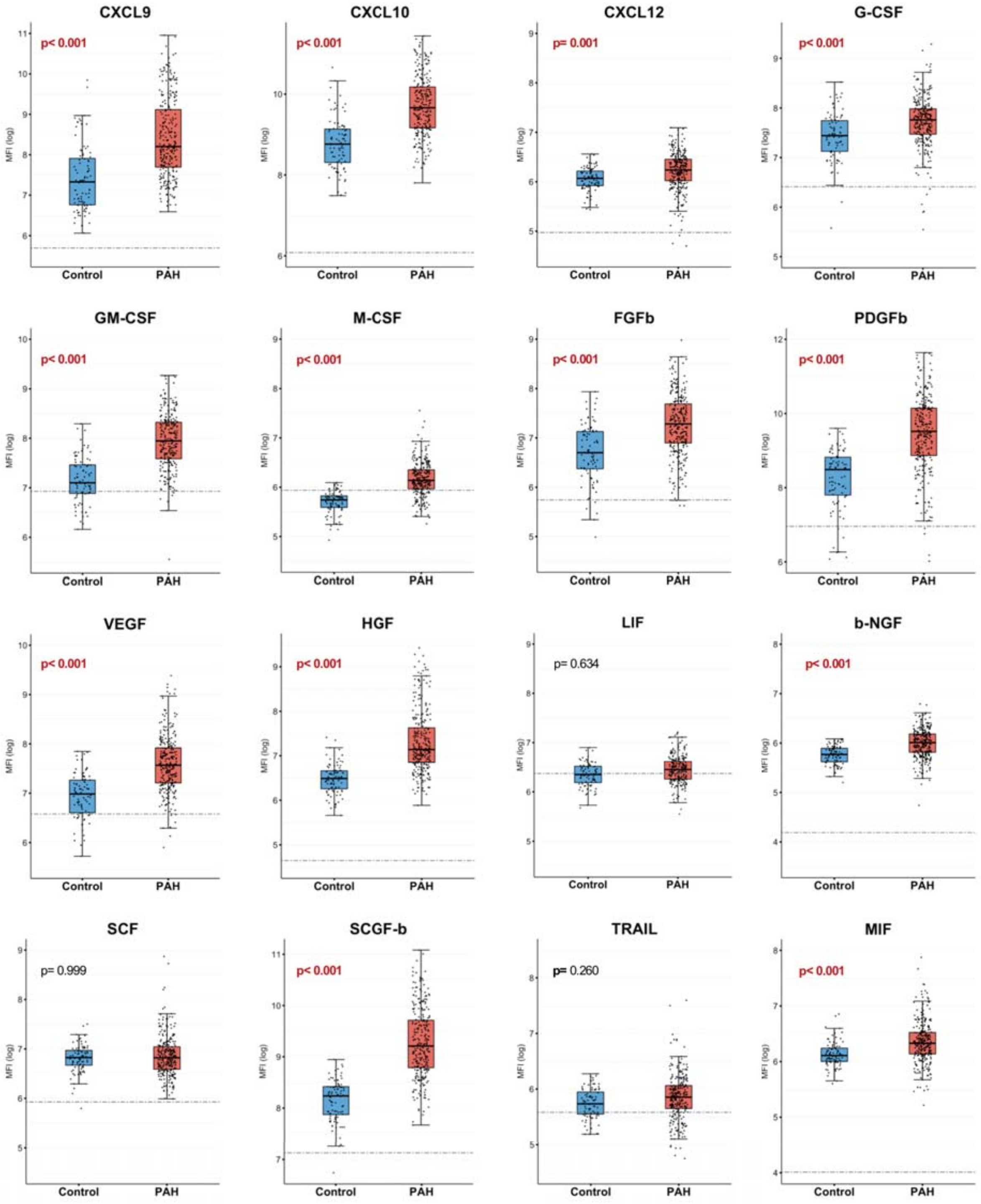


Figure XV (continued). Discovery cohort: plasma expression of individual cytokines, chemokines, and growth factors for PAH patients versus healthy controls.



ONLINE SUPPLEMENT REFERENCES

1. Simonneau G, Gatzoulis MA, Adatia I, Celermajer D, Denton C, Ghofrani A, Gomez Sanchez MA, Krishna Kumar R, Landzberg M, Machado RF, Olschewski H, Robbins IM and Souza R. Updated clinical classification of pulmonary hypertension. *J Am Coll Cardiol*. 2013;62:D34-41.

2. Benza RL, Miller DP, Gomberg-Maitland M, Frantz RP, Foreman AJ, Coffey CS, Frost A, Barst RJ, Badesch DB, Elliott CG, Liou TG and McGoon MD. Predicting survival in pulmonary arterial hypertension: insights from the Registry to Evaluate Early and Long-Term Pulmonary Arterial Hypertension Disease Management (REVEAL). *Circulation*. 2010;122:164-72.

3. Johnson WE, Li C and Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8:118-27.

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

5. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L and Liu C. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One*. 2011;6:e17238.

6. Monti S, Tamayo P, Mesirov J and Golub T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning*. 2003;52:91-118.

7. Wilkerson MD and Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*. 2010;26:1572-3.

8. 'fpc' package for R- Flexible Procedures for Clustering (http://cran.r-project.org/web/packages/fpc/) [computer program]. Version 2.1-10; 2015.

9. Gordon A. *Classification.* 2nd ed. Boca Raton, FL: Chapman & Hall/CRC; 1999.

10. Handl J, Knowles J and Kell DB. Computational cluster validation in post-genomic data analysis. *Bioinformatics*. 2005;21:3201-12.

11. Senbabaoglu Y, Michailidis G and Li JZ. Critical limitations of consensus clustering in class discovery. *Sci Rep*. 2014;4:6207.

12. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. 1987;20:53-65.

13. Milligan GW and Cooper MC. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*. 1985;50:159-179.

14. Halkidi M, Batistakis Y and Vazirgiannis M. On Clustering Validation Techniques. *Journal of Intelligent Information Systems*. 2001;17:107-145.

15. Caliński T and Harabasz J. A dendrite method for cluster analysis. *Communications in Statistics*. 1974;3:1-27.

16. Hennig C and Liao TF. How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2013;62:309-369.

17. Kaufman L and Rousseeuw PJ. Partitioning Around Medoids (Program PAM) *Finding Groups in Data*: John Wiley & Sons, Inc.; 2008: 68-125.

18. Lauritzen SL. *Graphical models*. Oxford New York: Clarendon Press ; Oxford University Press; 1996.

19. Koller D and Friedman N. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*: The MIT Press; 2009.

20. Steuer R. Review: analysis and interpretation of correlations in metabolomic data. *Brief Bioinform*. 2006;7:151-8.

21. Krumsiek J, Suhre K, Illig T, Adamski J and Theis FJ. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst Biol*. 2011;5:21.

22. de la Fuente A, Bing N, Hoeschele I and Mendes P. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*. 2004;20:3565-74.

23. Yuan Y, Li CT and Windram O. Directed partial correlation: inferring large-scale gene regulatory network through induced topology disruptions. *PLoS One*. 2011;6:e16835.

24. Freudenberg J, Wang M, Yang Y and Li W. Partial correlation analysis indicates causal relationships between GC-content, exon density and recombination rate in the human genome. *BMC Bioinformatics*. 2009;10 Suppl 1:S66.

25. Telesca D, Muller P, Kornblau SM, Suchard MA and Ji Y. Modeling Protein Expression and Protein Signaling Pathways. *J Am Stat Assoc*. 2011;107:1372-1384.

26. Meinshausen N and Buhlmann P. High-dimensional graphs and variable selection with the Lasso. *Ann Statist*. 2006;34:1436-1462.

27. Pourahmadi M. Covariance Estimation: The GLM and Regularization Perspectives. *Statist Sci*. 2011;26:369-387.

28. Fruchterman TMJ and Reingold EM. Graph drawing by force-directed placement. *Software: Practice and Experience*. 1991;21:1129-1164.

29. Epskamp S, Cramer AOJ, Waldorp LJ, Schmittmann VD and Borsboom D. qgraph: Network Visualizations of Relationships in Psychometric Data. *2012*. 2012;48:18.

30. Pavlopoulos GA, Secrier M, Moschopoulos CN, Soldatos TG, Kossida S, Aerts J, Schneider R and Bagos PG. Using graph theory to analyze biological networks. *BioData Mining*. 2011;4:10.

31. Barrat A, Barthélemy M, Pastor-Satorras R and Vespignani A. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*. 2004;101:3747-3752.

32. Opsahl T, Agneessens F and Skvoretz J. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*. 2010;32:245-251.

33. Dijkstra EW. A Note on Two Two Problems in Connexion with Graphs. *Numerische Mathematik*. 1959;1:269-71.

34. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2011;73:273-282.

35. Friedman J, Hastie T and Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. 2008;9:432-41.

36. 'glasso' R software package- Graphical LASSO estimation of Gaussian graphical models (http://cran.r-project.org/web/packages/glasso/) [computer program]. 2015.

37. Chen J and Chen Z. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*. 2008;95:759-771.

38. Foygel R and Drton M. *Extended Bayesian Information Criteria for Gaussian Graphical Models*; 2010.

39. Efron B. The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association*. 1977;72:557-565.

40. Breen EJ, Tan W and Khan A. The Statistical Value of Raw Fluorescence Signal in Luminex xMAP Based Multiplex Immunoassays. *Sci Rep*. 2016;6:26996.