Biological heterogeneity in idiopathic pulmonary arterial hypertension identified through unsupervised transcriptomic profiling of whole blood

Sokratis Kariotis1,2, Emmanuel Jammeh1,2, Emilia M. Swietlik3,4, Josephine A. Pickworth5,2, Christopher J. Rhodes5, Pablo Otero5, John Wharton5, James Iremonger6,2, Mark J. Dunning1, Divya Pandya3, Thomas S. Mascarenhas1, Niamh Errington6,12, A. A. Roger Thompson2,6, Casey E. Romanoski7, Franz Rischard7, Joe G. N. Garcia7, Jason X.-J. Yuan8, Tae-Hwi Schwantes An9, Ankit A. Desai9, Gerry Coghlan10, Jim Lordan11, Paul A. Corris11, Luke S. Howard6,5, Robin Condliffe2,6, David G. Kiely2,6,12, Colin Church13, Joanna Pepke-Zaba4, Mark Toshner3,4, Stephen Wort5, Stefan Gräf3, Nicholas W. Morrell3, Martin R. Wilkins6,5, Allan Lawrie1,2,29,2, Dennis Wang1,14,15,29,2 & UK National PAH Cohort Study Consortium*

Idiopathic pulmonary arterial hypertension (IPAH) is a rare but fatal disease diagnosed by right heart catheterisation and the exclusion of other forms of pulmonary arterial hypertension, producing a heterogeneous population with varied treatment response. Here we show unsupervised machine learning identification of three major patient subgroups that account for 92% of the cohort, each with unique whole blood transcriptomic and clinical feature signatures. These subgroups are associated with poor, moderate, and good prognosis. The poor prognosis subgroup is associated with upregulation of the ALAS2 and down-regulation of several immunoglobulin genes, while the good prognosis subgroup is defined by upregulation of the bone morphogenetic protein signalling regulator NOG, and the C/C variant of HLA-DPA1/DPB1 (independently associated with survival). These findings independently validated provide evidence for the existence of 3 major subgroups (endophenotypes) within the IPAH classification, could improve risk stratification and provide molecular insights into the pathogenesis of IPAH.

1Department of Neuroscience, University of Sheffield, Sheffield, UK. 2Department of Infection, Immunity & Cardiovascular Disease, University of Sheffield, Sheffield, UK. 3Department of Medicine, University of Cambridge, Cambridge, UK. 4Royal Papworth Hospital, Cambridge, UK. 5National Heart and Lung Institute, Imperial College London, London, UK. 6Sheffield Pulmonary Vascular Disease Unit, Royal Hallamshire Hospital, Sheffield, UK. 7Department of Cellular and Molecular Medicine, University of Arizona, Tucson, AZ, USA. 8Department of Medicine, University of California, San Diego, La Jolla, CA, USA. 9Department of Medicine, Indiana University, Indianapolis, IN, USA. 10Royal Free Hospital, University College London, London, UK. 11Newcastle University, Newcastle, UK. 12Insigneo institute for In Silico Medicine, Sheffield, UK. 13University of Glasgow, Glasgow, UK. 14Department of Computer Science, University of Sheffield, Sheffield, UK. 15Singapore Institute for Clinical Sciences, Singapore, Singapore. These authors jointly supervised this work: Allan Lawrie, Dennis Wang. *A list of authors and their affiliations appears at the end of the paper. Email: a.lawrie@sheffield.ac.uk; dennis.wang@sheffield.ac.uk
Pulmonary arterial hypertension (PAH) is a rare but devastating disease characterised by sustained pulmonary vasconstriction and progressive pulmonary vascular remodelling. This leads to an increase in pulmonary vascular resistance and pulmonary artery pressure, resulting in right heart failure and death\(^1\). The cause of idiopathic PAH (IPAH) remains unknown and diagnosis is derived from the exclusion of other forms of PAH, resulting in a heterogeneous group of patients who have significant differences in survival and treatment response across clinical cohort and registry studies\(^2-5\).

The pathobiology of PAH involves the complex interaction of resident vascular cells, including endothelial cells, arterial smooth muscle cells and fibroblasts, with infiltrating inflammatory cells, and has been shown to be regulated by an ever growing number of molecular and genetic mechanisms\(^6-8\). We have identified both rare mutations\(^9\) and common variants\(^10\) in heritable and idiopathic PAH (H/IPAH) that have provided further insight into the genetic underpinning of PAH. Additional proteomic\(^1\), metabolomic\(^12\) and transcriptomic\(^13\) studies have described diagnostic and prognostic biomarkers that add to our increasing understanding of the molecular mechanisms that regulate disease in this cohort. In Rhodes et al. we compared clinically defined H/IPAH cases to healthy controls and defined an imperfect diagnostic signature for H/IPAH; however, we have not previously examined the molecular heterogeneity that exists within H/IPAH cases. Deep RNA profiling of blood samples has provided accessible biomarkers to detect rare diseases\(^14\) and defined molecular mechanisms behind myocardial infarction\(^15\). We therefore investigated whether transcriptomic profiling of whole blood can provide more granular molecular ‘endophenotypes’ of H/IPAH to stratify patients better than is currently permissible with the standard clinical classification. Furthermore, we hypothesised that these transcriptome-defined subgroups would provide additional insights into biological pathways driving disease, and potential drug targets offering a route to precision medicine approaches for H/IPAH.

In this study, assessment of transcriptome patterns in whole blood was conducted using unsupervised machine learning agnostic to the clinical definitions and descriptors of H/IPAH. We describe the unbiased partitioning of patients into multiple distinct transcriptomic subgroups that associate with different survival properties, each with predictive clinical and genetic features. Specifically, we highlight the potential role of immunity and immune genes in discrimination of PAH endophenotypes associated with differential patient outcomes. These data further highlight the concept that inflammation is an important mediator of PAH pathogenesis\(^16-22\) and the discovery of distinct immune subgroups from blood cytokine profiles of patients with PAH\(^16-18\). Finally, we identify a specific panel of clinical features that describe each transcriptomic subgroup and replicate these subgroups in a validation cohort who did not undergo full transcriptomic profiling using their clinical phenotype data. The gene expression profile of key cluster associated genes was subsequently confirmed, and the correlation with key clinical variables validated in both internal and external validation cohorts, thereby validating our approach, and providing an alternative method to define these endophenotypes without the need for transcriptomic data.

**Results**

Unsupervised cluster analysis of whole-blood transcriptomes reveals five distinct subgroups of H/IPAH. Whole blood samples from patients with H/IPAH (\(n = 359\)) were processed for RNA-sequencing as previously described\(^13\). Samples from 359 patients and 21 samples collected from a second time point underwent RNAseq data processing to reduce noise, and gene filtering to remove gender bias as sex chromosomes produced the highest variation in gene expression during clustering (Supplementary Fig. 1). Sample collection site did not produce any discernible effect on clustering (Supplementary Figs. 2 and 3).

Simultaneously, the 300 genes that produced the most stable expression dataset were utilised to identify unique subgroups of gene expression profiles and describe the biological and clinical descriptors of these subgroups (Fig. 1). A clustering algorithm for selection and majority voting of multiple internal validation indexes (Supplementary Data 1) allowed us to identify as statistically optimal five distinct and stable subgroups of patients’ profiles (Fig. 2a) while retaining the maximum heterogeneity information found in our dataset. The largest of the patient clusters identified was subgroup I (\(n = 129\)), which had poorer survival (53%, five-year median survival from sampling; Fig. 2b).

The second largest, subgroup II (\(n = 112\)), demonstrated the best survival (78%, 5 years from sampling; Fig. 2b). Subgroup V (\(n = 89\)) demonstrated a mixed gene expression pattern and average survival outcome compared to subgroups I and II (Fig. 2a, b). Subgroups III (\(n = 19\)) and IV (\(n = 10\)) also demonstrated distinct gene expression patterns, with subgroup III most similar to subgroup II, and subgroup IV similar to subgroup I both in terms of gene expression level and survival outcomes. Due to the small size of subgroups III and IV (making statistical significance unattainable), we focused further characterisation of

![Fig. 1 Overview of IPAH subgroup identification methodology.](https://doi.org/10.1038/s41467-021-27326-0)
I

II

III

IV

V

Fig. 2 Gene expression profiles, survival and risk categories that demonstrate five distinct subgroups. a The expression heatmap for the five discovered subgroups showing distinct expression profiles. b Kaplan–Meier survival curves for the three predominant subgroups demonstrating the difference in survival profiles (from RNA sampling) for a span of 5 years along with two-sided log-rank test p values. c The percentage of predominant subgroups I, II and V patients across REVEAL risk categories. High- and very-high-risk populations mostly consist of subgroup I patients (45.5% and 73.3%, respectively), while the low-risk population is mostly composed of subgroup II (38.3%) and V (29.5%) patients. Fisher’s exact test showed a statistically significant difference (two-sided p value = 0.024) between subgroups I and II for low- and very-high-risk categories.

Changes observed with almost all samples belonging to functional class III. When including transcriptomes from healthy volunteers in our cluster analysis, the highest proportion of healthy volunteers (39.1%) grouped with subgroup II patients (better prognosis) (Supplementary Fig. 6b). To further investigate the defining characteristics of the three largest subgroups, we interrogated both their gene expression profiles and clinical features to define their endophenotype.

Relative expression of immunoglobulins define RNA-based subgroups of IPAH. We next interrogated the three largest RNA-based subgroups using a multivariate penalised regression to identify the relationship between gene expression profiles and each of the three subgroups. The most parsimonious model revealed 57 genes with measurable association to the subgroups. ALAS2 (erythroid ALA-synthase), a catalysing haeme biosynthesis enzyme, appeared in the signatures for both subgroups I and II, and was the most differentially expressed gene (~2-fold) between the two subgroups. Several immunoglobulin light chain genes (IGKV and IGLV) were key markers for the subgroups, and these were found to be either downregulated in subgroup I (poor prognosis) or upregulated in subgroup II (good prognosis; Fig. 3a). Other than immunoglobulins, Noggin, a bone morphogenetic protein 4 antagonist, and inhibitor of hyoxia-induced proliferation23, was the gene with the highest positive regression coefficient for subgroup II, underlining its association with good prognosis. BMP antagonist Noggin and immunoglobulin genes associated with the good prognostic subgroup II were all downregulated by more than twofold in subgroup I (Fig. 3b), fitting with contemporary understanding of perturbed BMP and inflammatory signalling in PAH pathogenesis16, 21, 24. Across the three major subgroups, the relative expression level of immunoglobulins ranged from low, intermediate and high for subgroups I, V and II, respectively (Fig. 3a, c), while Noggin showed significantly higher expression in subgroup II (Supplementary Fig. 7).

Differential immune cell composition between IPAH subgroups. To ascertain whether the large expression differences in immunoglobulin genes associated with subgroups I and II also corresponded to different levels of immune activity, we deconvoluted the RNA profiles to estimate the proportions of immune genetic and clinical correlates for subgroups I, II and V. The 33 HPAH patients in our PAH cohort showed an equal distribution (~10%) among the subgroups of our initial clustering (Supplementary Table 1), indicating that the inclusion of HPAH, or the small number of mis-classified patients, did not drive the partitioning procedure. An additional clustering pipeline exclusively utilising 313 samples from patients with IPAH (i.e. excluding those with HPAH, or re-classified PH) also showed five subgroups (Supplementary Fig. 4), where there were also a group of patients with poorer survival (clusters B and E, n = 149), a group with good survival (A and C, n = 109) and a group with moderate survival (D, n = 55).

In order to determine whether the survival differences between the three main (largest) transcriptomic subgroups were also associated with disease severity in the surviving patients, we calculated the REVEAL 2.0 risk score4 across all risk levels: low (n = 146), moderate (n = 41), high (n = 44) and very high (n = 15). Subgroup I which had the worst survival also had both the highest percentage of patients in high-risk categories (medium 43.9%, high 45.5% and very high 73.3%) and a lower percentage (32.2%) in the low-risk category (Fig. 2c). In contrast, subgroup II which had the best survival was composed mostly of low-risk patients (38.3%), a proportion significantly different to subgroup I (z-test p = 0.01422, Fig. 2c). The distribution of subgroup V was uniform across the risk groups, except for a small proportion of very-high-risk patients (6.6%). Age and sex were also included as covariates with the subgroups in a Cox regression model. Age above 52 years (median) was significantly associated with poor survival (HR = 2.29) while gender showed no relationship with overall survival. Even with these covariates, subgroup I was still significantly associated with survival and was the biggest risk factor (HR = 3.83) for poor outcome (Supplementary Fig. 5). Within each subgroup, a small number of patients had a second time-point sample collected on average after 463 days. Patients with these longitudinal samples (n = 19) were found to either remain within their subgroup or transition from either subgroup I (poor prognosis) or II (good prognosis) to the moderate prognosis subgroup V (Supplementary Fig. 6a).

Interestingly, no patient transitioned from subgroup II (best survival) directly to subgroup I (worst survival) or vice versa over time, 9 patients changed through the moderate subgroup, while 12 stayed in the same subgroup. Additionally, no functional class
Common clinical characteristics across RNA subgroups. Patients in this cohort were diagnosed at a median age of 45 years (IQR = 35–59 years) and sampled at a median age of 52 years (64–76) with an average of 5.3 years’ time between diagnosis and sampling. As shown in Table 1, patients in subgroup I were significantly older (p value < 0.01) at 57 [45–70] years than the other subgroups. Consistent with the incidence rate of IPAH in the UK population, patients in the cohort were predominantly females (70%). Patients in the subgroups were also predominantly females with 62%, 73% and 70% in subgroups I, II and V, respectively. Across the whole cohort, 16.4% of patients presented positive pulmonary vasodilator response, 44.4% were in Functional Class (FC) III at sampling date with 6-minute walk distance (6MWD) of 387 m and a mean N-terminal (NT)-proBNP of 222.5 [78.9–1162.8] ng/ml. When the cohort was stratified, subgroup I had the highest proportion of FC III (50.4%), whereas subgroup II had the highest proportion of patients for FC I and II (16.5% and 41.3%, respectively, p value = 0.013). The lowest 6MWD (median = 327 m, p < 0.01) and the highest N-terminal (NT)-proBNP was (median = 345.0 ng/ml, p = 0.055) were observed in patients from subgroup I (poorest survival group). Diagnostic RHC across the cohort showed mean pulmonary arterial pressure (mPAP) was 54 (46–61) mmHg, pulmonary arterial wedge pressure (PAWP) was 10 (7–12) mmHg and CO was 3.8 (3.0–4.9) l/min at diagnosis. The cohort at the time of sampling, 143 (40.2%) of the patients were FC II and 158 (44.4) FC III with a median 6MWD of 387 m, pulmonary vascular resistance (PVR) was 8.9. Wood units and an NT-proBNP 222.5 ng/ml suggestive of a slight improvement of disease phenotype in response to vasodilatory therapy. The full demographics table can be found in Supplementary Data 2.

Clinical signatures describe RNA-based subgroups. Identification of specific clinical characteristics associated with each transcriptome-derived subgroup could explain how the gene expression patterns manifest into differences in patient outcome. We therefore used supervised machine learning with feature selection to identify the most important clinical features to

---

**Fig. 3 Genes associated with subgroups I (low survival), II (high survival) and V (intermediate survival).**

**a** Genes with the highest 5% of LASSO coefficients across subgroups I, II and V. **b** Average expression fold change (log2 scaled) of the signature genes between subgroup I and II, with significance notations. Genes over-expressed in subgroup I are denoted by light blue bars while genes primarily expressed in subgroup II are represented by dark blue bars. **c** Expression level of immunoglobulin genes selected by LASSO across the three predominant subgroups with medians shown. Subgroups I (n = 134), II (n = 98) and V (n = 119) can be defined as having low, intermediate and high immunoglobulin characteristics. Vertical centre line represents the median, top and bottom bounds of the box represent the first and third quartile, while the tips of the whiskers represent min and max values.
describe the subgroups. The full list of clinical features used by the multivariate classifiers are described in Supplementary section 'Clinical features identification: Supervised learning' and in table values: pI-II (Dendritic cells activated) = 0.011, pI-III (Neutrophils) = 4.4 × 10^{-11}, pI-IV (Neutrophils) = 2.0 × 10^{-5}, pI-V (Neutrophils) = 1.7 × 10^{-3}, pI-III (T cells CD8) = 4.8 × 10^{-5}, pI-III (T cells CD4 naive) = 1.9 × 10^{-8}, pI-IV (T cells CD4 naive) = 3.8 × 10^{-3}, pI-V (T cells CD4 memory resting) = 2.3 × 10^{-5}, pI-IV (B cells naive) = 2 × 10^{-5}, pI-IV (B cells memory) = 2.5 × 10^{-15}, pI-V (B cells memory) = 3.9 × 10^{-3}, pI-IV (Plasma cells) = 6.4 × 10^{-4}, pI-V (Plasma cells) = 6.5 × 10^{-5} and pI-V (Monocytes) = 0.0053. Vertical centre line represents the median, top and bottom bounds of the box represent the first and third quartile, while the tips of the whiskers represent min and max values. b Whole-blood cell counts across subgroups I (n = 129), II (n = 112) and V (n = 89) using two-sided test and Bonferroni adjusted mean difference significance notation. pI-II (Neutrophils) = 7.2 × 10^{-12}, pI-III (Neutrophils) = 8.0 × 10^{-4}, pI-IV (Neutrophils) = 4.4 × 10^{-4}, pI-V (Neutrophils) = 7.2 × 10^{-4}, pI-II (Neutrophils/Lymphocytes) = 0.0061 and pI-II (Monocytes) = 0.0076. Vertical centre line represents the median, top and bottom bounds of the box represent the first and third quartile, while the tips of the whiskers represent min and max values. c Proportion of patients in each subgroup with DNA variants in HLA-DPA1/DPB1 (rs2856830), SOX17 (rs10106467) and rs13266183, homozygous and heterozygous, BMPR2 (rare pathogenic variant). Notably, pI-II (HLA-DPA1/DPB1) = 0.009. Generated using a two-sample test for equality of proportions with continuity correction. *P value ≤ 0.05, **P value ≤ 0.01, ***P value ≤ 0.001.

Fig. 4 Immunity cell composition across PAH transcriptomic subgroups. a CIBERSORT estimation of relative cell abundance in patients of subgroups I (n = 129), II (n = 112) and V (n = 89) using two-sided test and Bonferroni adjusted mean difference significance notation with p values: pI-II (Dendritic cells activated) = 0.011, pI-III (Neutrophils) = 4.4 × 10^{-11}, pI-IV (Neutrophils) = 2.0 × 10^{-3}, pI-V (Neutrophils) = 1.7 × 10^{-3}, pI-III (T cells CD8) = 4.8 × 10^{-5}, pI-III (T cells CD4 naive) = 1.9 × 10^{-8}, pI-IV (T cells CD4 naive) = 3.8 × 10^{-3}, pI-V (T cells CD4 memory resting) = 2.3 × 10^{-5}, pI-IV (B cells naive) = 2 × 10^{-5}, pI-IV (B cells memory) = 2.5 × 10^{-15}, pI-V (B cells memory) = 3.9 × 10^{-3}, pI-IV (Plasma cells) = 6.4 × 10^{-4}, pI-V (Plasma cells) = 6.5 × 10^{-5} and pI-V (Monocytes) = 0.0053. Vertical centre line represents the median, top and bottom bounds of the box represent the first and third quartile, while the tips of the whiskers represent min and max values. b Whole-blood cell counts across subgroups I (n = 129), II (n = 112) and V (n = 89) using two-sided test and Bonferroni adjusted mean difference significance notation. pI-II (Neutrophils) = 7.2 × 10^{-12}, pI-III (Neutrophils) = 8.0 × 10^{-4}, pI-IV (Neutrophils) = 4.4 × 10^{-4}, pI-V (Neutrophils) = 7.2 × 10^{-4}, pI-II (Neutrophils/Lymphocytes) = 0.0061 and pI-II (Monocytes) = 0.0076. Vertical centre line represents the median, top and bottom bounds of the box represent the first and third quartile, while the tips of the whiskers represent min and max values. c Proportion of patients in each subgroup with DNA variants in HLA-DPA1/DPB1 (rs2856830), SOX17 (rs10106467) and rs13266183, homozygous and heterozygous, BMPR2 (rare pathogenic variant). Notably, pI-II (HLA-DPA1/DPB1) = 0.009. Generated using a two-sample test for equality of proportions with continuity correction. *P value ≤ 0.05, **P value ≤ 0.01, ***P value ≤ 0.001.

CRP and 6MWD were the only clinical features present in signatures for subgroup I and II and V. Higher CRP was a marker for subgroup I, whereas lower levels indicated subgroups II and V. In contrast, 6MWD was negatively associated with subgroup I and positively with subgroups II and V. CRP showed a 37.19% increase in subgroup I compared to the average for subgroups II and V, and 20.75% reduction in subgroup V compared to the average for subgroups I and V. 6MWD was 29.05% lower in subgroup I compared to the average for II and V, and increased by 7.63% in subgroup V compared to the average for II and I and 16.97% increase in subgroup II compared to the
average for I and V. Five clinical features were present in signatures for subgroup I and II but had opposite coefficients (Fig. 5b). Higher age of diagnosis, BMI, RAA and creatinine are associated in subgroup I, whereas lower levels of those three features are associated with subgroup II. Subgroup I has 17.8% higher average age compared to the average for II and V and 21.2% lower in subgroup II compared to the average for I and V. BMI was 13.1% higher in subgroup I compared to the average for II and V, and 12.9% lower in subgroup II compared to the average.

Additionally, creatinine was higher by 12.8% in subgroup I compared to the average for II and V and lower by 26.7% in subgroup II compared to the average for I and V.

Table 1 Major clinical characteristics of the three main RNA subgroups in the discovery cohort (n = 359) at the time of sampling.

|                  | Low-risk subgroup II (high immunoglobulin) | Intermediate-risk subgroup V (intermediate immunoglobulin) | High-risk subgroup I (low immunoglobulin) | All patients |
|------------------|-------------------------------------------|------------------------------------------------------------|------------------------------------------|--------------|
| n                | 112                                       | 89                                                        | 129                                      | 359          |
| Age, years       | 46 [37–56]                                | 52 [41–62]                                                | 57 [45–70]                               | 52 [42–64]   |
| Age at diagnosis, years | 41 [31–51]                                | 46 [37–55]                                                | 52 [42–67]                               | 47 [35–59]   |
| Gender:Female    | 82 (73%)                                   | 69 (78%)                                                  | 80 (62%)                                 | 253 (70%)    |
| Vasoresponse     | 10 (21.7%)                                 | 6 (13.6%)                                                 | 6 (16.2%)                                | 23 (16.4%)   |
| Treatments       |                                           |                                                           |                                          |              |
| Phosphodiesterase 5 Inhibitors (PDE5i) | 12 (15.4%)                                  | 16 (21.9%)                                                | 22 (21.8%)                              | 53 (19.4%)   |
| Endothelin receptor antagonist (ERA)     | 6 (7.6%)                                   | 13 (17.8%)                                                | 8 (7.92%)                                | 33 (12.1%)   |
| PDE5i & ERA combination | 42 (53.8%)                                 | 30 (41.1%)                                                | 53 (52.5%)                               | 134 (49.1%)  |
| Prostacyclin      | 3 (3.85%)                                  | 1 (1.37%)                                                 | 3 (2.97%)                                | 7 (2.56%)    |
| Calcium channel blockers | 15 (19.2%)                                 | 13 (17.8%)                                                | 14 (13.9%)                               | 45 (16.5%)   |
| WHO functional class |                                         |                                                           |                                          |              |
| I                | 18 (16.5%)                                 | 10 (11.2%)                                                | 6 (4.7%)                                 | 35 (9.8%)    |
| II               | 45 (41.3%)                                 | 36 (40.4%)                                                | 44 (34.1%)                               | 143 (40.2%)  |
| III              | 43 (39.4%)                                 | 40 (44.9%)                                                | 65 (50.4%)                               | 158 (44.4%)  |
| IV               | 3 (2.8%)                                   | 3 (3.4%)                                                 | 14 (10.9%)                                | 20 (5.6%)    |
| 6-minute walking distance, m            | 397 [338–500]                              | 420 [367–464]                                            | 327 [183–390]                            | 387 [300–449] |
| NT-proBNP, ng/l | 131.7 [54.5–362.0]                         | 185.5 [76.3–463.5]                                        | 345.0 [91.0–1556.1]                      | 222.5 [78.9–1162.8] |
| Forced expiratory volume [% predicted] | 92 [82–101]                                 | 84 [72–98]                                               | 78 [66–98]                               | 85 [68–100]  |
| Forced vital capacity [% predicted]     | 101 (20)                                   | 99 (24)                                                  | 93 (29)                                  | 97 (24)      |
| Transfer factor of lung for carbon monoxide [% predicted] | 93 [87–106]                                 | 97 [92–101]                                              | 88 [67–96]                               | 94 [87–103]  |
| Diagnostic Right Heart Catheter Study   |                                           |                                                           |                                          |              |
| Mean pulmonary arterial pressure, mmHg  | 47 [39–60]                                 | 52 [37–65]                                                | 56 [41–65]                               | 51 [39–63]   |
| Mean right atrial pressure, mmHg        | 8 [4–10]                                   | 8 [4–11]                                                 | 11 [6–14]                                | 9 [4–12]     |
| Mean pulmonary arterial wedge pressure, mmHg | 10 [7–12]                                 | 10 [8–13]                                                | 12 [10–14]                               | 11 [8–13]    |
| Cardiac Index, l/min/m²                 | 2.3 [1.6–2.8]                              | 2.2 [1.7–2.4]                                            | 1.9 [1.5–2.5]                            | 2.2 [1.6–2.5] |
| Pulmonary vascular resistance, Wood Units | 8.1 [5.7–14.1]                             | 15.0 [5.9–16.1]                                          | 8.4 [5.9–13.2]                           | 8.9 [5.7–15.0] |

Intervals describe first and third quartiles. Parentheses describe standard deviation (SD).

Validation of clinical signatures on an independent cohort. To validate the relationship between clinical and gene features in the RNA subgroups, we used the clinical feature signatures of the subgroups to classify patients in an independent cohort of I/HPAH patients (n = 197) where whole-blood RNA profiling was not performed (Fig. 5c). Similar to the discovery cohort, patients were diagnosed at a median age of 52 years (IQR = 39–67) and 67% were female. In all, 17.7% of the patients showed positive pulmonary vasodilator response and the majority were categorised in Functional Class III (66%) with a 6MWD of 295 m (170–396) and NT-proBNP of 796 ng/pl (128–1092). Their mPAP was 51 mmHg (42–57) and PAWP was 9 mmHg (6–11). The clinical features associated with RNA subgroups from the discovery cohort subgroups I, II and V (Table 2). These subgroups also displayed differences in their 10-year survival outcome from diagnosis (Fig. 5d). Those characterised as subgroup I based on their clinical features (corresponding to the low Noggin and immunoglobulin expression subgroups from RNAseq) (n = 96) demonstrated the lowest survival of 71% from the time of diagnosis. Subgroup V (corresponding to the immune neutral, intermediate RNAseq subgroup) (n = 31) also had an intermediate survival of 86%, while patients in Subgroup II (corresponding to the best surviving subgroup with upregulated Noggin and immunoglobulin genes) showed a very high survival rate of 97.2% (n = 96). These results provide key validation of the existence of endophenotypes for the three major subgroups of patients within the H/IPAH clinical classification group, and that these new subgroups can be identified using routinely collected clinical features associated with RNA dysregulation.
Fig. 5 Clinical variables descriptive of RNA subgroups and used for classification of new patients. a Comparison of clinical variables deemed most important from our univariate feature selection model across subgroups I (n = 129), II (n = 112) and V (n = 89). Vertical centre line represents the median, top and bottom bounds of the box represent the first and third quartile, while the tips of the whiskers represent min and max values. b Clinical variables selected by ensemble feature selection from models predictive of each subgroup. Coefficients shown for each variable are from the most predictive support vector machine classifiers. c Selected clinical features are used to classify 197 IPAH patients from an independent validation cohort. d Kaplan–Meier survival curves per predicted subgroup in the validation cohort confirming the difference in survival outcomes between subgroups along with log-rank test p values. e Gene and clinical variable correlation network. Diamond nodes represent clinical variables drawn from the clinical signatures. Round nodes represent genes drawn from the gene signature generated by our LASSO model. Edges denoted Spearman rank correlation and have been thresholded to 0.25 and two-tailed test p value < 1.11 × 10^{-5}. Specifically, corrBMI-ALAS2 = 1.27 × 10^{-11}, corrBMI-Pi3 = 3.17 × 10^{-6}, corrBMI-IGHG2 = 4.13 × 10^{-6}, corrBMI-RP11.678G14.3 = 8.22 × 10^{-6}, corrBMI-IGKV2.24 = 3.09 × 10^{-6}, corrBMI-IGKV4.4 = 9.55 × 10^{-7}, corr6MWD-IGKV2.24 = 2.83 × 10^{-6}, corr6MWD-IGKV4.4 = 2.08 × 10^{-6}, corr6MWD-ALAS2 = 7.52 × 10^{-10}, corrAaO-IGHV2.5 = 3.72 × 10^{-10}, corrAaO-IGKJ4 = 1.06 × 10^{-9}, corrAaD-IGHM = 6.2 × 10^{-8}, corrAaD-NOG = 3.18 × 10^{-17}, corrAaD-IGKV3.48 = 7.7 × 10^{-7}, corrAaD-IGLV7.43 = 1.04 × 10^{-6}, corrAaD-IGKV4.4 = 6.35 × 10^{-10}, corrAaD-IGKV2.24 = 4.19 × 10^{-6}, corrAaD-IGKV2.27 = 3.93 × 10^{-7}, corrOxygenSat-NOG = 1.11 × 10^{-6}.
Clinical signatures are associated with subgroup-specific genes. We assessed the relationship between gene and clinical features of the subgroups by measuring the correlation between the most predictive features in both signatures. Immunoglobulins IGHV2.5, IGKV2.24, and IGKV1.27 are negatively correlated with age of diagnosis (Fig. 5e). Interestingly, ALAS2 correlated most strongly with age of diagnosis ($r = 0.443$) but positively correlated with oxygen saturation ($r = 0.275$). Interestingly, ALAS2 correlated most strongly with BMI ($r = 0.382$) but showed an inverse correlation with 6MWD ($r = -0.323$). This is consistent with our observations in the poor prognosis subgroup I where patients with higher expression of ALAS2 also had higher BMI and shorter walk distances. Genes negatively correlated with BMI included immunoglobulin light chains ($r < 0.01$), and NOG suggesting that a molecular classification (Supplementary Table 5).

Gene expression in clinical-feature defined subgroups. Although the RNAseq whole transcriptome was not measured in this internal validation cohort, we compared gene expression differences between subgroups in this cohort using TaqMan PCR for 17 of the 27 genes (GAPDH used as the endogenous control gene) previously associated with the subgroups and/or clinical variable correlations. Nine of the 11 genes we measured in our validation cohort were also examined in our discovery cohort (Supplementary Table 4). Differences in gene expression of key genes (IGHM, IGKV2.24, IGLV6.57, and NOG) were significant ($p < 0.01$) between subgroups I and II (Fig. 6a and Supplementary Table 4).

External validation of gene and clinical feature correlations. The correlations between gene and clinical features observed in the discovery cohort were also examined in our validation cohort of 91 subjects, and also in an external cohort of 32 subjects with RNA collected from PBMCs. We found that 64 of the 90 (71%) correlations measured in these two independent cohorts were consistent with our discovery cohort (Supplementary Table 5).

**Table 2 Major clinical characteristics of the three subgroups within the validation cohort ($n = 197$) at time of diagnosis.**

| WHO functional class | Subgroup I | Subgroup II | Subgroup V | All patients |
|----------------------|------------|-------------|------------|--------------|
| I                    | 4 (4.2%)   | 11 (15.7%)  | 2 (6.5%)   | 17 (8.5%)    |
| II                   | 22 (22.9%) | 26 (37.1%)  | 15 (48.4%) | 63 (32.0%)   |
| III                  | 60 (62.5%) | 32 (45.7%)  | 12 (38.7%) | 104 (52.8%)  |
| IV                   | 10 (10.4%) | 1 (1.4%)    | 2 (6.5%)   | 13 (6.6%)    |
| 6-minute walking distance, m | 306 (152)  | 419 (123)   | 409 (120)  | 360 (148)   |
| NT-proBNP, ng/l       | 492 [196; 1327] | 188 [90.0; 400] | 266 [128; 499] | 303 [128; 1092] |
| Forced expiratory volume [% predicted] | 84 (21)    | 90 (19)     | 90 (17)    | 87 (20)     |
| Forced vital capacity [% predicted] | 94 (21)    | 98 (20)     | 99 (15)    | 96 (20)     |
| Transfer factor of lung for carbon monoxide [% predicted] | 92 (15)    | 98 (17)     | 96 (11)    | 95 (15)     |
| Diagnostic Right Heart Catheter Study |          |             |            |              |
| Mean pulmonary artery pressure, mmHg | 50 [43; 57] | 48 [42; 58] | 49 [41; 57] | 49 [42; 57] |
| Mean right atrial pressure, mmHg | 9 [7; 12]  | 7 [5; 10]   | 6 [3; 7]   | 8 [5; 12]   |
| Mean pulmonary Arterial wedge pressure, mmHg | 10 (3)     | 8 (4)       | 8 (4)      | 9 (4)       |
| Cardiac Index, l/min/m | 2.0 [1.7; 2.5] | 2.1 [1.7; 2.5] | 2.0 [1.8; 2.4] | 2.0 [1.7; 2.5] |
| Pulmonary vascular resistance, Wood Units | 11 [7; 14] | 11 [9; 15]  | 12 [10; 14] | 11 [8; 15]  |

Intervals describe first and third quartiles. Parentheses describe standard deviation (SD).
The whiskers represent min and max values.

with BMI in PAH37 has been proposed as a potential therapeutic validating that our unbiased approach has identi

studies across multiple forms of PH, including IPAH, showed correlated with greater disease severity. Previous gene expression

ALAS2 biosynthesis through

sion patterns that de

role of adaptive immunity on PAH progression. 

receptor and B cell receptors may be needed to understand the subgroup I patients, deeper genomic characterisation of T cell

important subgroups. While we detected signi

expression of

partially de

remodelling and PH. Subgroup II with better prognosis can be

associated with autoimmune diseases, including systemic lupus erythe-

matosis (SLE), type 1 diabetes, and myasthenia gravis28, 29. The

several autoimmune diseases, including systemic lupus erythe-

reduction in the diversity of light chains has been associated with

induced proliferation of PASMC 23, and previously associated

role for

These data, and our own observations (Fig. 3), are suggestive of a

strong correlation with right atrial pressure, pulmonary vascular

levels also demonstrated

donation of ALAS2 in both systemic sclerosis-associated PAH (SSc-PAH) and IPAH32. In that study,
in IPAH patients increased

in IPAH patients increased

ALAS2

NPRL3

IGKV2.24

IGHM

CRISP3

MT-RNR1

IGHV3.75

LTF

IGLV6.57

may control self-reactivity of human antibodies27, and the reduction in the diversity of light chains has been associated with several autoimmune diseases, including systemic lupus erythematosis (SLE), type 1 diabetes, and myasthenia gravis28, 29. The association between autoimmunity and PAH has long been dis-
cussed. There are known associations with autoimmune diseases in other forms of PAH such as systemic sclerosis, SLE, Sjogren’s, etc., and the dysregulation of immune cells including T cells, B cells30 and natural killer cells31 are well described in IPAH, fur-

there. While we detected significant differences in lymphocyte, neutrophil and CRP levels in the blood samples of subgroup I patients, deeper genomic characterisation of T cell receptor and B cell receptors may be needed to understand the role of adaptive immunity on PAH progression.

Beyond the differences in immunoglobulin genes, the expres-
sion patterns that defined each subgroup also highlighted haeme biosynthesis through ALAS2 was a marker for subgroup I and correlated with greater disease severity. Previous gene expression studies across multiple forms of PH, including IPAH, showed significantly increased expression of ALAS2 in both systemic sclerosis-associated PAH (SSc-PAH) and IPAH32. In that study, in IPAH patients increased ALAS2 levels also demonstrated strong correlation with right atrial pressure, pulmonary vascular resistance, pulmonary artery saturation and cardiac index32. These data, and our own observations (Fig. 3), are suggestive of a role for ALAS2, iron33 and hepcidin34, 35 in pulmonary vascular remodelling and PH. Subgroup II with better prognosis can be partially defined by the downregulation of ALAS2 and increased expression of NOG, a BMP antagonist with high-affinity binding to BMP4 (ref. 36) which has been shown to inhibit hypoxia-induced proliferation of PASMC25, and previously associated with BMI in PAH35 has been proposed as a potential therapeutic target38. The role of Noggin in the low-risk group is particularly interesting given the proposed role of both Gremlin and Noggin in the mechanism of action for Sotatercept in the treatment of PAH39.

Previous studies have identified clinical features collected during the diagnosis of PAH that also have prognostic utility. The clinical features identified here share many commonalities with those previously included in widely used risk scores (e.g. REVEAL, ERS) assessment for PAH, including, for example, 6-MWD, WHO functional class, and NT-proBNP4, 40, 41. This provided further validation that the transcriptomic profile asso-
ciated with these subgroups provide insight into the biology of disease, and perhaps future drug targets. In addition to bio-

markers such as CRP which is known to be elevated in PAH and CTEPH and shown to be predictive of outcome and sensitive to therapies42 and NT-proBNP with high levels highly prognostic of right ventricular failure43, age of diagnosis, BMI and renal function were also identified. Renal function has previously been associated with outcome in PAH, although likely because of cardiac function44. The age of diagnosis is often discussed as a consequence of genetics45, or occurrence of co-morbidities; however, in our study the age of diagnosis was most strongly associated with the immunoglobulin light chain genes and Noggin. Carriers of BMPR2 mutations often present with PAH at a younger age and have a worse survival46 so the association with Noggin is interesting in the context of perturbed BMP signalling. However, the patients with BMPR2 mutation did not cluster within one subgroup perhaps fitting with the concept that it is dysfunctional TGFβ/BMP signalling rather than the precise mutation that is important.

There is a well-described sex-paradox in PAH47 with a 4:1 female to male prevalence but the worse survival in male patients48, 49. During our initial analyses of the RNAseq data, we identified subclusters exclusively defined by sex genes. To mitigate against any gender bias, we excluded sex-chromosome-associated genes in our preprocessing steps of the analysis pipeline. Although we cannot reject the possibilities of the
aforementioned genes contributing towards PAH or resilience, we believe that their removal ensures that the clustering algorithm captures heterogeneity independent of sex-associated expression variation. However, the interactions between gender and other autosomal genes in the context of PAH require further study.

The application of unsupervised learning from molecular profiles of IPAH is a powerful approach for revealing subgroups within a heterogeneous population that has not been defined clinically. Most studies employ widely used clustering algorithms without exploring their data suitability. By contrast, in this study we determine spectral clustering as the most consistent method in detecting differences and subsequently partitioning RNA-sequencing samples using robust performance criteria. Furthermore, previous studies have focused on clustering all PAH cases using a small set of immune markers, and captured immune phenotypes overlaid by the broad clinical classifications50. We used a much larger set of features, i.e., the whole transcriptome and clustered cases lacking causal pathologies, and also found immune phenotypes that differentiated the subgroups. While we controlled for confounding factors that affect clustering, such as gene–gender–disease interactions (Supplementary Fig. 1), there may yet be other hidden factors, such as viral infections related to age and gender that could influence patterns observed from whole blood51. The large degree of validation of the subgroups using both transcriptomic and clinical features to define them provides strong evidence that these endophenotypes are reproducible and may be useful to risk stratify or biologically classify subgroups of IPAH patients. However, further transcriptomic studies profiling patients at multiple timepoints are required to fully understand the dynamics of the immune components we identified, the frequency of acute infections, and the impact on PAH phenotype.

Transcriptomic profiling of the blood samples coupled with clinical data from IPAH patients provides an insight into endophenotypes that may describe this heterogeneous disease based on RNA expression. The use of additional ‘omic’ biomarkers to provide further molecular profiles (e.g. DNA, protein, metabolites) as stable biomarkers for stratifying patients could further improve our algorithmic predictions of patient outcomes and reveal endophenotypes to be targeted therapeutically. Furthermore, these data hold promise that these molecular endophenotypes may be tractable to existing therapies, may offer an alternative approach to tailor, and assess individual treatment response, in PAH as well as offering insights into disease pathogenesis that can be targeted by therapies as a precision medicine approach52 in PAH and potentially other diseases to improve our algorithmic predictions of patient outcomes and following informed consent.

Participants. Patients diagnosed with I/HPAH, PVOD or PCH, relatives of index cases and unrelated healthy controls were recruited at nine UK centres and followed up by a median of 7.9 years. In total, 358 patients (Supplementary Fig. 11) of which 96.7% were further verified to be I/HPAH, 13 relatives, and 21 healthy controls recruited to the I/HPAH Cohort study were analysed. Both prevalent and incident cases were allowed. Prevalent cases were defined as diagnosed earlier than 6 months before the study initiation. Patients in Cohort 1 were followed longitudinally as part of their clinical PAH care. All cases were diagnosed between March 1994 and November 2016, and diagnostic classification was made according to international guidelines53. Patients with PAH associated with anorexigen exposure were considered as IPAH, whereas HPAH was defined by the presence of a positive family history. Clinical, functional and demographic characteristics at the time of PAH diagnosis were prospectively entered into the database. The date of diagnosis corresponded to that of confirmatory right heart catheterisation.

Following diagnosis, subsequent treatments and follow-ups were at the discretion of the treating physician, according to the contemporary guidelines. In most centres, patients were seen every 3–6 months with an assessment of functional status and exercise capacity. Right heart catheterisation was repeated when considered necessary by the responsible clinician. Study visits were performed every 6 months. Healthy controls had been sampled only once and had clinical information recorded from the time of sampling.

Clinical data capture, processing and quality control. Pseudonymised results of routinely performed clinical tests reported in either clinical case notes or electronic medical records (EMR) were stored in web-based OpenClinica (OC) data capture system (Community edition). Twenty electronic Clinical Case Report Forms (eCRFs) distributed across seven events (Diagnostic, Continuous data, Follow-up, Epidemiology questionnaire, Suspension, Relatives, Unrelated healthy control) were constructed to accommodate routinely available clinical information. Details regarding data verification procedures were previously described in detail34. Information about participants’ status was collected every 6 months (via National Health System Digital Spine portal or an equivalent local system). Current analysis was performed on the census performed on 31 January 2020. Two risk assessment strategies were applied to the data. Reveal risk score4 and abbreviated ERS risk scores35 were calculated in all patients who had the necessary minimum phenotypic information available. Patients who died or were transplanted were suspended on the day of the event, patients who withdrew from the study were censored on the date of the last visit, the reason for withdrawal was recorded.

Missingness assessment and imputation. Missingness rates, patterns and causes were assessed per individual, variable and centre and visualised with vim package v5.1.1R (Supplementary Fig. 12). Multiple imputation by the chain equations method was used to impute missing data (mice v3.8.0 package R)36. The imputation model included all variables that were necessary in the analysis model, including cumulative baseline hazard function and variables that predicted both the imputed variable and if the imputed variable was observed and whether the case was incident or prevalent. Quality of predictors was assessed using outl–influ plot. Numerical data were imputed with predictive mean matching (pmm), factors with two levels were imputed using logistic regression, factors with more than two levels with multivariate logit model and ordered factors with more than two levels with the ordered logit model. Transformed variables (BMI, ratios, score sums) were imputed as just another variable as well as passively with good concordance. The visiting sequence was set to ‘monotone’ to speed up convergence. The number of iterations was set to 20. Following the rule of thumb proposed by White et al.37 that the number of imputations should be at least equal to the percentage of incomplete cases, the procedure was performed at m = 50. The convergence of the algorithm was checked, and the means and standard deviations of imputed values were plotted over 20 iterations. The streams of numerical and factor variables intermingled and showed no trends at later iterations. Factors influencing the accuracy of the imputation include the variability in time between diagnosis and sampling, higher missingness in clinical data for prevalent cases (diagnosed sometimes many years ago), and differences in measurement error between centres which followed different protocols for clinical data collection.

RNA data preprocessing. A number of preprocessing steps were required to prepare the raw sequencing data for unsupervised machine learning. High-throughput sequencing generated raw pair-end counts of 205,259 transcripts across 508 samples that belong to GenCode Release 28 (GRCh38.p12). Count-Seq. Salmon (https://combine-lab.github.io/salmon/) was used to estimate the relative abundance of the transcripts (TPM, units of transcripts per million) which were then mapped to genes (n = 60,144) using the tximport R package. Only genes with more than two reads (in a transcript level) in at least 95% of control and patient samples (considered gene level functional, functional enrichment scores >3.955). Hyperbolic arc sine transformation (package base v3.6.0) was applied to the final RNAseq TPM matrix. Further information on quality control of samples and genes

Methods

Study design. The Cohort study of idiopathic and heritable PAH is an observational, prospective and longitudinal study of patients with idiopathic and heritable PAH (clinicaltrials.gov NCT019972950). The Sheffield Teaching Hospitals Observational Study of Pulmonary Hypertension, Cardiovascular and other Respiratory Disease (UK REC Ref 18/YH/0441) is a longitudinal study of patients with suspected pulmonary hypertension or an associated cardiovascular or respiratory condition. Follow-up information is collected as a part of routine clinical care every 6 months. The study allows recruitment of both incident and prevalent cases. Patients consented to the study agreed to have blood taken for next-generation sequencing and other omics studies. Healthy adult controls were recruited for comparison studies. The subsequent whole-blood sample collection process is described in ref. 15.

Ethics. All UK samples were obtained following informed consent into the UK National Cohort Study of Idiopathic and Heritable Pulmonary Arterial Hypertension (clinicaltrials.gov NCT01907295; UK REC Ref. 13/EE/0203) and/or the Sheffield Teaching Hospitals Observational Study of Pulmonary Hypertension, Cardiovascular and other Respiratory Disease (UK REC Ref 18/YH/0441). Data were obtained from samples collected at the University of Arizona Pulmonary Hypertension clinic between 2012 and 2015 following institutional guidelines and following informed consent.
can be found in the Supplementary Methods. The RNA-sequencing and clinical data of healthy controls were not used in the main pipeline of this study. A second independent patient and healthy control cohort was implemented to demonstrate the lack of pure patient and healthy subgroups within our cohort (Supplementary Fig. 6b). Principal component analysis of expression profiles from samples with a second replicate clustered together according to the first four principal components (Supplementary Fig. 13).

**Spectral clustering: gene expression subgroup identification.** We performed cluster analysis to partition IPAH patients to distinct RNA-based groups. The spectral clustering model (package kernlab v0.9-29) was selected as the most suitable unsupervised learning algorithm based on the highest partial concordance when comparing multiple dissimilar algorithms (Supplementary Table 6). For the spectral clustering method, data points (i.e. patients) are embedded and partitioned in a low-dimensional space in the form of a similarity graph, rather than being characterised by more than 25,000 gene dimensions. High partial concordance was defined as the high adjusted Rand Index (package v0.3-7) and low standard deviation calculated between different variations of each clustering algorithm (k-means, spectral, hierarchical clustering), as described in Clustering algorithm selection. For the selection of the most appropriate clustering algorithm we utilised 25,955 genes across 359 IPAH patient samples (discovery cohort) after further filtering for repeated same-visit samples and non-HIPAH diagnosis. We compared three fundamentally different methods (hierarchical, k-means and spectral) and use partitioning consistency to determine which method picks up an unseen validation dataset. The model was trained on the discovery dataset to discriminate between subgroups I, II and V, and V used to predict subgroup membership of an unseen validation dataset. The predicted subgroup membership was then used to calculate survival of predicted subgroups. Survival of the predicted subgroups was compared to known survival of subgroups in the discovery dataset for validation purposes.

**qPCR on validation cohort.** Frozen Tempus tubes collected from patients in the validation cohort, collected under the UK National Cohort study, were obtained; RNA was extracted using Maxwell 16 LEV simplyRNA Blood Kit (Cat. # AS1310) as described in the manufacturer's instructions on the Maxwell® 16 Instrument (Cat.# AS2000). Extracted RNA was transcribed using the High-Capacity RNA-to-cDNA kit (Thermo Fisher Cat.# 437406) following the manufacturer's instructions. Resultant cDNA was analysed using custom TaqMan array cards (THERMO FISHER®) with Fast Advance II (Thermo Fisher Cat.# 4449864); duplicates were run 8 to 24 cards across 24 primers probes (Thermo Fisher) per sample (185-Hs99999901_s1, ALAS2-Hs01083701_m1, BMPR2-Hs00176148_m1, C4BPA-Hs00426339_m1, CRISP3-Hs00119598_m1, CTSG-Hs00175195_m1, GAPDH-Hs02786624_g1, HPRT1-Hs00280095_m1, IFI27-Hs01086373_g1, IGHM-Hs00941538_g1, IGHV3-75-Hs01382008_s1, IGKV2-24-Hs06671467_g1, IGLV6-57-Hs01696367_s1, LINC00221-Hs13082601_m1, LTTL-Hs00914334_m1, MT-RNR1-Hs02566859_g1, NEBL-Hs01067284_m1, NOG-Hs00271352_s1, NOS2-Hs01075529_m1, NPR1L3-Hs00429221_m1, PI3-Hs00106666_m1, SMIM11A;SMIM11B-Hs00938773_m1, XIST-Hs01079824_m1). These assays were performed in duplicate using the Applied Biosystems 7900HT Fast Real-time PCR system with the TaqMan Low Density Array card block following calibration using the TaqMan Low Density Array Calibration Kit (Thermo Fisher Cat.# 1034165). Ct values were determined with Automatic thresholding in the SDS2.2 software. GAPDH- Hs02786624_g1 was used as a control. Relative quantity was calculated using the ΔΔct method.

**External cohort validation.** An external validation cohort of patients with Group 1 PAH prospectively recruited at the University of Arizona Pulmonary Hypertension clinic between 2012 and 2015 following institutional guidelines and informed consent was used. The cohort comprised 84 subjects with Group 1 PAH of whom 32 were diagnosed with idiopathic PAH. For each subject, demographics and clinical variables were collected. 48 PMCs were stored in RNAlater as previously described. In total, approximately 3600 million clusters with paired-end 75 bp reads (~35M cluster per sample) were generated from PMCs-deriv RNA.

**Clinical variable and gene correlations.** We calculated correlations between the clinical and gene signatures we generated in previous steps of this study. For discovery and validation cohorts we used the rcorr function of R package Hmisc (version 4.5-0). For the external validation we used the values found in ref. 36.

**Study approval.** Study approval for the use of sample and data was obtained from the UK National PAH Cohort Study Data Access Committee (clinicaltrials.gov NCT019072795; UK REC Ref 13/EE/0203), and the Sheffield Teaching Hospitals Observational Study of Pulmonary Hypertension, Cardiovascular and other Respiratory Diseases Scientific Advisory Board (UK REC Ref 18/YH/0441).

### Reporting summary
Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**
The transcriptomic and clinical data used in this study have been deposited in the EGA (the European Genome-phenome Archive) database under accession code EGAS00001055326. In compliance with the Ethics with which these data and samples have been collected, the transcriptomic data are available through restricted access for approved researchers who agree to the conditions of use, i.e. keeping it secure and only using it for approved purposes. To apply for access please contact

All ML tasks were carried out using Scikit-learn® ML framework version 0.23.2 in a Python 3 environment. As machine learning classifiers, we used Logistic Regression (LR), support vector machines (SVM), Random Forests (RF) and k-nearest neighbor (KNN). RF is a powerful ensemble learning technique especially for high-dimensional classification tasks. Further details about classifier training and feature selection can be found in the Supplementary Methods.

**Classification of new patients using signatures.** Each clinical signature was used to develop a classification model trained on the discovery cohort to classify new patients into the RNA-based subgroups. Classification models were built using SVM, RF, kNN and KNNS. The candidate signature that obtained the best performance was selected. This process was repeated for all signature sizes, s = 1 to s = 20, for subgroups I and V. A final signature for each subgroup was selected based on a compromise between the fewest number of features (s = 1 to s = 20) and classification performance. Final selected signatures for each of the subgroups were pooled to create a composite signature, which was then used in a multi-class classification model. The model was trained on the discovery dataset to discriminate between subgroups I, II and V, to predict subgroup membership of an unseen validation dataset. The predicted subgroup membership was then used to calculate survival of predicted subgroups. Survival of the predicted subgroups was compared to known survival of subgroups in the discovery dataset for validation purposes.
References

1. Galé, N. et al. Risk stratification and medical therapy of pulmonary arterial hypertension. Eur. Respir. J. 53, 1801889 (2019).
2. Hurmdjian, J. et al. ASPIRE registry: assessing the spectrum of pulmonary hypertension identified at a RÉFÉRÉ honour. Eur. Respir. J. 39, 945–955 (2012).
3. Ling, Y. et al. Changing demographics, epidemiology, and survival of incident pulmonary arterial hypertension: results from the pulmonary hypertension registry of the United Kingdom and Ireland. Am. J. Respir. Crit. Care Med. 186, 790–796 (2012).
4. Benza, R. L. et al. Predicting survival in patients with pulmonary arterial hypertension: the REVEAL Risk Score Calculator 2.0 and comparison with ESC/ERS-based risk assessment strategies. Chest 156, 323–337 (2019).
5. Benza, R. L. et al. High levels of healthcare utilization prior to diagnosis in idiopathic pulmonary arterial hypertension support the feasibility of an early diagnosis algorithm: the SPHINX project. Pulm. Circ. 8, 2045894018796613 (2018).
6. Thompson, A. A. R. & Lawrie, A. Targeting vascular remodeling to treat pulmonary hypertension. J. Allergy Clin. Immunol. 138, 239–249 (2016).
7. Southgate, L., Machado, R. D., Gräf, S. & Morrell, N. W. Molecular genetic framework underlying pulmonary arterial hypertension. Nat. Rev. Cardiol. 8, 443–455 (2011).
8. Southgate, L., Machado, R. D., Gräf, S. & Morrell, N. W. Molecular genetic framework underlying pulmonary arterial hypertension. Nat. Rev. Cardiol. 17, 85–95 (2020).
9. Gräf, S. et al. Identification of rare sequence variation underlying heritable pulmonary arterial hypertension. Nat. Commun. 9, 1416 (2018).
10. Rhodes, C. J. et al. Genetic determinants of risk in pulmonary arterial hypertension: international genome-wide association studies and meta-analysis. Lancet Respir. Med. 7, 227–238 (2019).
11. Rhodes, C. J. et al. Platelet profile analysis in patients with pulmonary arterial hypertension: an observational cohort study. Lancet Respir. Med. 5, 717–726 (2017).
12. Rhodes, C. J. et al. Whole blood RNA profiles associated with pulmonary arterial hypertension and clinical outcome. Am. J. Respir. Crit. Care Med. https://doi.org/10.1164/rcrm.202003-0510OC (2020).
13. Frésard, L. et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. Nat. Med. 25, 911–919 (2019).
14. Vanhaverbeke, M., Veltman, D., Janssens, S. & Sinnave, P. Peripheral blood RNAs and left ventricular dysfunction after myocardial infarction: towards translation into clinical practice. J. Cardiovasc. Transl. Res. https://doi.org/10.1007/s12265-020-10048-x (2020).
15. Pickworth, J. et al. Differential IL-1 signaling induced by BMPR2 deficiency drives pulmonary vascular remodeling. Pulm. Circ. 7, 768–776 (2017).
16. Farkas, D. et al. Toll-like receptor 3 is a therapeutic target for pulmonary hypertension. Am. J. Respir. Crit. Care Med. 199, 199–210 (2019).
17. Hameed, A. G. et al. Inhibition of tumor necrosis factor-related apoptosis-inducing ligand (TRAIL) reverses experimental pulmonary hypertension. J. Exp. Med. 209, 1919–1935 (2012).
18. Frid, M. G. et al. Immunoglobulin-driven complement activation regulates proinflammatory remodeling in pulmonary hypertension. Am. J. Respir. Crit. Care Med. 201, 224–239 (2020).
19. Mamazhakypov, A., Viswanathan, G., Lawrie, A., Schermuly, R. T. & Rajagopal, S. The role of chemokines and chemokine receptors in pulmonary arterial hypertension. Br. J. Pharmacol. https://doi.org/10.1111/bph.14826 (2019).
20. Hurst, L. A. et al. TNFα drives pulmonary arterial hypertension by suppressing the BMP type-II receptor and altering NOTCH signalling. Nat. Immunol. 18, 14079 (2017).
21. Steiner, M. K. et al. Interleukin-6 overexpression induces pulmonary hypertension. Circ. Res. 104, 236–244, 28p following 244 (2009).
22. Yang, K. et al. Noggin inhibits hypoxia-induced proliferation by targeting store-operated calcium entry and transient receptor potential cation channels. Am. J. Physiol. Cell Physiol. 308, C869–C878 (2015).
23. Bell, R. D. et al. TNF induces obliterative pulmonary vascular disease in a novel model of connective tissue disease associated pulmonary arterial hypertension (CTD-PAH). Arthritis Rheumatol. https://doi.org/10.1002/art.41309 (2020).
24. Foris, V. et al. Neutrophil-to-lymphocyte ratio as a prognostic parameter in pulmonary arterial hypertension. 4, Pulmonary Circulation and Pulmonary Vascular Diseases. https://doi.org/10.1183/13993003.congress-2016.2pa2479 (2016).
25. Romanoski, C. E. et al. Transcriptomic profiles in pulmonary arterial hypertension associate with disease severity and identify novel candidate genes. Pulm. Circ. 10, 2045894020696331 (2020).
26. Collins, A. M. & Watson, T. Immunoglobulin light chain gene rearrangements, receptor editing and the development of a self-tolerant antibody repertoire. Front. Immunol. 9, 2249 (2018).
27. Panigrahi, A. K. et al. R5 rearrangement frequency as a marker for receptor editing in lupus and type 1 diabetes. J. Exp. Med. 205, 2985–2994 (2008).
28. Rander Heiden, J. A. et al. Dysregulation of B cell repertoire formation in myasthenia gravis patients revealed through deep sequencing. J. Immunol. 198, 1460–1473 (2017).
29. Nicoll, M. R., Taraseviciene-Stewart, L., Rai, P. R., Badesch, D. B. & Voelkel, N. F. Autoimmunity and pulmonary hypertension: a perspective. Eur. Respir. J. 26, 1110–1118 (2005).
30. Ormiston, M. L. et al. Impaired natural killer cell phenotype and function in idiopathic and heritable pulmonary hypertension. Circulation 126, 1099–1109 (2012).
31. Cheadle, C. et al. Erythroid-specific transcriptional changes in PMBCs from pulmonary hypertension patients. PLoS ONE 7, e34951 (2012).
32. Rhodes, C. J. et al. Iron deficiency in pulmonary arterial hypertension: a potential therapeutic target. Eur. Respir. J. 38, 1453–1460 (2011).
33. Rhodes, C. J. et al. Iron deficiency and raised hepcidin in idiopathic pulmonary arterial hypertension: clinical prevalence, outcomes, and mechanistic insights. J. Am. Coll. Cardiol. 58, 300–309 (2011).
34. Ramakrishnan, L. et al. The Hepcidin/Ferroportin axis modulates proliferation of pulmonary artery smooth muscle cells. Sci. Rep. 8, 12972 (2018).
35. Mehler, M. F., Mabie, P. C., Zhang, D. & Kessler, J. A. Bone morphogenetic proteins in the nervous system. Trends Neurosci. 20, 309–317 (1997).
36. Al-Khafaji, K. H. A., Al-Dujaili, M. N. & Al-Dujaili, A. N. G. Assessment of nogo-n protein in pulmonary arterial hypertension patients. Curr. Iss. Pharm. Med. Sci. 31, 122–130 (2018).
37. Boucherat, O. & Bonnet, N. NOGGIN: a new therapeutic target for PH Focus on "Noggin inhibits hypoxia-induced proliferation by targeting store-operated calcium entry and transient receptor potential cation channels. Am. J. Physiol. Cell Physiol. 308, C867–C868 (2015).
38. Humbert, M. et al. Solotaccept for the treatment of pulmonary arterial hypertension. N. Engl. J. Med. 384, 1204–1215 (2021).
39. Benza, R. L. et al. The REVEAL registry risk score calculator in patients newly diagnosed with pulmonary arterial hypertension. Chest 141, 354–362 (2012).
40. Benza, R. L. et al. Prognostic implications of serial risk score assessments in patients with pulmonary arterial hypertension: A Registry to Evaluate Early and Long-Term Pulmonary Arterial Hypertension Disease Management (REVEAL) analysis. J. Heart Lung Transpl. 34, 356–361 (2015).
41. Quarc, R., Navrot, T., Meyns, B. & Delcroix, M. C-Reactive protein: a new predictor of adverse outcome in pulmonary arterial hypertension. J. Am. Coll. Cardiol. 53, 1211–1218 (2009).
42. Nickel, N. et al. The prognostic impact of follow-up assessments in patients with idiopathic pulmonary arterial hypertension. Eur. Respir. J. 39, 589–596 (2012).
43. Kaiser, R., Seiler, S., Held, M., Bals, R. & Wilkens, H. Prognostic impact of renal function in precapillary pulmonary hypertension. J. Intern. Med. 275, 116–126 (2014).
44. Phillips, J. A. et al. Synergistic heterozygosity for TGFB1 SNPs and BMPR2 mutations modulates the age at diagnosis and penetrance of familial pulmonary arterial hypertension. Genet. Med. 10, 359–365 (2008).
45. Evans, I. D. W. et al. BMPR2 mutations and survival in pulmonary arterial hypertension: an individual participant data meta-analysis. Lancet Respir Med. 4, 129–137 (2016).
46. Austin, E. D. Gender, sex hormones, and pulmonary arterial hypertension. Adv. Exp. Med. Biol. 85, 9–12 (2012).
47. Fodorová, A. & Ventúlova, C. E. Pulmonary arterial hypertension and the sex hormone paradox. Curr. Hypertens. Rep. 18, 84 (2016).
49. Hester, J., Ventetoulo, C. & Lahm, T. Sex, gender, and sex hormones in pulmonary hypertension and right ventricular failure. *Compr. Physiol.* 10, 125–170 (2019).
50. Sweat, A. J. et al. Discovery of distinct immune phenotypes using machine learning in pulmonary arterial hypertension. *Circ. Res.* 124, 904–919 (2019).
51. Bongen, E. et al. Sex differences in the blood transcriptome identify robust changes in immune cell proportions with aging and influenza infection. *Cell Rep.* 29, 1963–1973.e4 (2019).
52. Morrell, N. W. et al. Genetics and genomics of pulmonary arterial hypertension. *Eur. Respir. J.* 53, 1801899 (2019).
53. Galé, N. et al. 2015 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension. The Joint Task Force for the Diagnosis and Treatment of Pulmonary Hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS). *Eur. Respir. J.* 46, 903–975, 1855–1856 (2015).
54. Swietlik, E. M. et al. Bayesian inference associates rare KDR variants with specific phenotypes in pulmonary arterial hypertension. *Circ Genom Precis Med.* 14, e003155 https://doi.org/10.1161/CIRCGEN.120.003155 (2021).
55. Hooper, M. M. et al. Mortality in pulmonary arterial hypertension: prediction by the 2015 European pulmonary hypertension guidelines risk stratification model. *Eur. Respir. J.* 50, 1700740 (2017).
56. van Buuren, S. & Groothuis-Oudshoorn, K. mice: multivariate imputation by chained equations in R. *J. Statist. Softw.* 45, 1–67 (2011).
57. White, I. R., Royston, P. & Wood, A. M. Multiple imputation using chained equations: issues and guidance for practice. *Stat. Med.* 30, 377–399 (2011).
58. Polkar, R. Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.* 6, 21–45 (2006).
59. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550 (2014).
60. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830 (2011).
61. Choi, E., Schuetz, A., Stewart, W. F. & Sun, J. Using recurrent neural network models for early detection of heart failure onset. *J. Am. Med. Inform. Assoc.* 24, 361–370 (2017).
62. Breiman, L. Random Forests. *Mach. Learn.* 45, 5–32 (2001).
63. Hosmer, D. W., Jr., Lemeshow, S. & Cook, E. D. *Applied Logistic Regression, Second Edition: Book and Solutions Manual Set* (Wiley-Interscience, 2001).
64. Peterson, L. K-nearest neighbor. *Scholarpedia* 4, 1883 (2009).
65. Kariotis, S., Wang, D. & Lawrie, A. PAH sequencing study—EGA European Genome-Phenome Archive. https://ega-archive.org/studies/EGAS00001005532 (2021).
66. Kariotis, S. & Jammeh, E. BioSok/spectral_clustering_of_IPAH: v1.0.1. https://doi.org/10.5281/zenodo.5549872 (2021).

**Acknowledgements**

The UK National Cohort of Idiopathic and Heritable PAH is supported by grants from the British Heart Foundation (SP/12/12/29836 & SP/18/10/33975) and the UK Medical Research Council (MR/K020919/1). Additional samples from the Sheffield Teaching Hospitals Observational Study of Pulmonary Hypertension, Cardiovascular and other Respiratory Diseases were supported by British Heart Foundation (PG/11/116/29288). We gratefully acknowledge financial support from the UK Department of Health via the NIHR comprehensive Biomedical Research Centre award to Imperial College Healthcare supported by a British Heart Foundation Intermediate Clinical Research fellowship (FS/18/33808). E.J. is supported by the Academy of Medical Sciences Springboard (ref: SBF004/1052). M.W. is in receipt of a British Heart Foundation Centre for Research Excellence award (RE/18/4/34215). M.J.D. and D.W. are supported by the NIHR Sheffield Biomedical Research Centre. We thank and thank all the patients and their families who contributed to this research, the UK Pulmonary Hypertension Association for their support, NIHR BioResource volunteers for their participation, and gratefully acknowledge NIHR BioResource centres, NHS Trusts and staff for their contribution. We thank the National Institute for Health Research, NHS Blood and Transplant, and Health Data Research UK as part of the Digital Innovation Hub Programme. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

**Author contributions**

SK, EF and EMS contributed equally. All authors made substantial contributions to the conception or design and data acquisition of the work. S.K., E.I., E.M.S., J.A.P., C.I.R., P.O., J.W., J.L., M.I.D., D.P., T.S.M., N.E., A.A.R.T., C.E.R., F.R., J.G.N.G., J.X.-J.Y., T.-H.S., A.A.D., G.C., J.L., P.A.C., L.S.H., R.C., D.G.K., C.C., J.P.-Z., M.T., S.W., S.G., N.W.M., M.R.W., A.L. and D.W. performed the analysis and/or interpretation of data. S.K., E.I., E.M.S., A.L. and D.W. drafted the work and all authors revised it critically for important intellectual content; and gave final approval of the version submitted for publication; and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

**Competing interests**

The authors declare no competing interests.

**Additional information**

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-27326-0.

**Correspondence** and requests for materials should be addressed to Allan Lawrie or Dennis Wang.

**Peer review information** Nature Communications thanks Polina Mamoshina and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2021, corrected publication 2022

---

**UK National PAH Cohort Study Consortium**

Marta Bleda3, Charaka Hadinnapola3, Matthias Haimel3, Kate Auckland3, Tobias Tilly3, Jennifer M. Martin3, Katherine Yates3, Carmen M. Treacy3, Margaret Day16, Alan Greenhalgh16, Debbie Shipley16, Andrew J. Peacock17, Val Irvine17, Fiona Kennedy17, Shahin Moledina18, Lynsay MacDonald18, Eleni Tamvak18,
Anabelle Barnes, Victoria Cookson, Latifa Chentouf, Souad Al, Shokri Othman, Lavanya Ranganathan, J. Simon R. Gibbs, Rosa DaCosta, Joy Pinguel, Natalie Dormand, Alice Parker, Della Stokes, Dipa Ghedia, Yvonne Tan, Tanaka Ngcozana, Ivy Wanjiku, Gary Polwarth, Rob V. Mackenzie Ross, Jay Suntharalingam, Mark Grover, Ali Kirby, Ali Grove, Katie White, Annette Seatter, Amanda Creaser-Myers, Sara Walker, Stephen Roney, Charles A. Elliot, Athanasios Charalampopoulos, Ian Sabroe, Abdul Hameed, Iain Armstrong, Neil Hamilton, Alex M. K. Rothman, Andrew J. Swift, James M. Wild, Florent Soubrier, Mélanie Eyries, Marc Humbert, David Montani, Barbara Girerd, Laura Scelsi, Stefano Ghio, Henning Gall, Ardi Ghofrani, Harm J. Bogaard, Anton Vonk Noordegraaf, Arjan C. Houweling, Anna Huis in’t Veld & Gwen Schotte

16Freeman Hospital, Newcastle, UK. 17Golden Jubilee National Hospital, Glasgow, UK. 18Great Ormond Street Hospital, London, UK. 19Hammersmith Hospital, London, UK. 20Royal Brompton Hospital, London, UK. 21Royal Free Hospital, London, UK. 22Royal United Bath Hospitals, Bath, UK. 23Sheffield NIHR Clinical Research Facility, Royal Hallamshire Hospital, Sheffield, UK. 24Département de génétique, hôpital Pitié-Salpêtrière, Assistance Publique-Hôpitaux de Paris, and UMR_S 1166-ICAN, INSERM, UPMC Sorbonne Universités, Paris, France. 25Université Paris-Sud, Faculté de Médecine, Université Paris-Saclay, AP-HP, Centre de référence de l’hypertension pulmonaire sévère, INSERM UMR_S 999, Hôpital Bicêtre, Le Kremlin-Bicêtre, France. 26San Matteo, Pavia, Italy. 27University of Giessen, Giessen, Germany. 28VU University Medical Center, Amsterdam, The Netherlands.