Whole blood RNA profiles associated with pulmonary arterial hypertension and clinical outcome

Short title: Whole blood RNAseq in PAH

Christopher J Rhodes 1* , Pablo Otero-Núñez 1*, John Wharton 1 , Emilia M Swietlik 2 , Sokratis Kariotis 3 , Lars Harbaum 1 , Mark J Dunning 4 , Jason M Elinoff 5 , Niamh Errington 3 , A. A. Roger Thomson 6 , James Iremonger 6 , J. Gerry Coghlan 7 , Paul Corris 8 , Luke S Howard 1 , David Kiely 3, Colin Church 9 , Joanna Pepke-Zaba 10 , Mark Toshner 2 , Stephen Wort 1 , Ankit A. Desai 11 , Marc Humbert 12 , Prof William C. Nichols 13 , Laura Southgate 14 , David-Alexandre Trégouët 15 , Richard C. Trembath 16 , Inga Prokopenko 17, 18, Stefan Gräf 2,19 , Nicholas W Morrell 2 , Dennis Wang 3,4 , Allan Lawrie 6 , Martin R Wilkins 1

On behalf of the NIHR BioResource – Rare Diseases PAH Consortium and the UK National PAH Cohort Study Consortium; *these authors contributed equally

	Affiliation(s)	Country
1	National Heart and Lung Institute, Imperial College London	United Kingdom
2	Department of Medicine, University of Cambridge	United Kingdom
3	Sheffield Institute for Translational Neuroscience, University of Sheffield	United Kingdom
4	Sheffield Bioinformatics Core, The University of Sheffield	United Kingdom
5	Critical Care Medicine Department, National Institutes of Health Clinical Center	United States
6	Department of Infection, Immunity & Cardiovascular Disease, University of Sheffield	United Kingdom
7	University College London	United Kingdom
8	University of Newcastle	United Kingdom
9	University of Glasgow	United Kingdom
10	Papworth Hospital, Papworth	United Kingdom
11	Indiana University, Indianapolis IN	United States
12	Université Paris-Sud, Faculté de Médecine, Université Paris-Saclay; AP-HP, Service de Pneumologie, Centre de référence de l'hypertension pulmonaire, Hôpital Bicêtre, Le Kremlin-Bicêtre; INSERM UMR_S 999, Hôpital Marie Lannelongue, Le Plessis Robinson	France
13	Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Department of Pediatrics, University of Cincinnati College of Medicine	United States
14	Molecular and Clinical Sciences Research Institute, St George's University of London	United Kingdom
15	INSERM UMR_S 1219, Bordeaux Population Health research center, University of Bordeaux, Bordeaux	France
16	Division of Genetics and Molecular Medicine, King's College London	United Kingdom
17	Department of Clinical and Experimental Medicine, University of Surrey	United Kingdom
18	Dept of Metabolism, Digestion and Reproduction, Imperial College London	United Kingdom
19	NIHR BioResource for Translational Research, Cambridge Biomedical Campus	United Kingdom

Correspondence to: Prof Martin R Wilkins, m.wilkins@imperial.ac.uk, Imperial College London, 2S4 Commonwealth Building, Hammersmith Campus, Du Cane Road, LONDON, W12 0NN, United Kingdom

9.35 Pulmonary Hypertension: Clinical-Diagnosis/Pathogenesis/Outcome

Author contributions: All authors made substantial contributions to the conception or design and data acquisition of the work. CJR, PON, JW, EMS, SK, LH, MJD, JME, NE, AART, JI, DW AL and MRW performed the analysis and/or interpretation of data. CJR, PON, JW and MRW drafted the work and

all authors revised it critically for important intellectual content; and give 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.

Funding statement: We thank 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 Imperial Clinical Research Facility and NHS Blood and Transplant. 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. The UK National Cohort of Idiopathic and Heritable PAH is supported by the NIHRBR; the BHF (SP/12/12/29836) and the UK Medical Research Council (MR/K020919/1). We also gratefully acknowledge the participation of patients recruited to the US National Institutes of Health/National Heart, Lung, and Blood Institute (NIH/NHLBI)-sponsored National Biological Sample and Data Repository for PAH (also known as PAH Biobank). We also acknowledge funding from NHLBI R01HL136603 (AAD). This work was supported in part by the Assistance Publique-Hopitaux de Paris, INSERM, University Paris-Sud, and Agence Nationale de la Recherche (Departement Hospitalo-Universitaire Thorax Innovation; LabEx LERMIT, ANR-10-LABX-0033; and RHU BIO-ART LUNG 2020, ANR-15-RHUS-0002); and British Heart Foundation Centre award RE/18/4/34215. CJR is supported by a BHF Intermediate Basic Science Research fellowship (FS/15/59/31839) and Academy of Medical Sciences Springboard fellowship (SBF004\1095). LH is a recipient of ERS Fellowship (LTRF 2016–6884). AART is supported by a BHF Intermediate Clinical Fellowship (FS/18/13/3328). LS is supported by the Wellcome Trust Institutional Strategic Support Fund (204809/Z/16/Z) awarded to St. George's, University of London. DW is supported by an Academy of Medical Sciences Springboard fellowship (SBF004\1052). NWM is a British Heart Foundation Professor and NIHR Senior Investigator. AL is supported by a BHF Senior Basic Science Research fellowship (FS/13/48/30453 & FS/18/52/33808).

Abstract (246 words)

Rationale

Idiopathic and hereditary pulmonary arterial hypertension (PAH) are rare but comprise a genetically heterogeneous patient group. RNA-sequencing linked to the underlying genetic architecture can be used to better understand the underlying pathology by identifying key signalling pathways and stratify patients more robustly according to clinical risk.

Objectives

To use a three-stage design of RNA discovery, RNA validation/model construction and model validation to define a set of PAH-associated RNAs and a single summarising RNA model score. To define genes most likely to be involved in disease development, we performed Mendelian randomisation (MR) analysis.

Methods

RNA-sequencing was performed on whole blood samples from 359 patients with idiopathic, heritable and drug-induced PAH and 72 age- and sex-matched healthy volunteers. The score was evaluated against disease severity markers including survival analysis using all-cause mortality from diagnosis. MR used known eQTL and summary statistics from a PAH GWAS.

Measurements and Main Results

We identified 507 genes with differential RNA expression in PAH patients compared to controls. A model of 25 RNAs distinguished PAH with 87% accuracy (AUC 95% CI: 0.791-0.945) in model validation. The RNA model score was associated with disease severity and long-term survival ($p=4.66x10^{-6}$) in PAH. MR detected an association between SMAD5 levels and PAH disease susceptibility (OR:0.317, 95%CI:0.129-0.776, p=0.012).

Conclusions

A whole blood RNA signature of PAH, which includes RNAs relevant to disease pathogenesis, associates with disease severity and identifies patients with poor clinical outcomes. Genetic variants associated with lower SMAD5 expression may increase susceptibility to PAH.

Current word count: 3375 (main text)

Background

Pulmonary arterial hypertension (PAH) is associated with vasoconstriction and occlusion of distal pulmonary arteries, characterised by endothelial damage, smooth muscle and fibroblast proliferation, and inflammation. Increased pulmonary vascular resistance leads to right heart failure, with survival rates estimated at 52-75% at 5-years, even with modern day therapy(1). The rate of deterioration and response to therapy varies between patients, driving a search for better predictors of clinical outcome and tools to inform drug selection. Molecular profiling using multiple omics technologies offers greater granularity than standard clinical phenotypes for characterising PAH patients and could improve initial risk stratification, treatment selection and monitoring as well as providing insights into biological pathways not yet targeted by current therapies(2-4).

Transcriptome profiling through RNA sequencing permits a comprehensive analysis of gene expression in tissue samples. Whole blood RNA analysis offers an alternative "liquid biopsy" to lung biopsy, which carries a high risk in PAH, and can be performed sequentially. This approach can also investigate immune mechanisms in PAH that have recently been highlighted (5). Previous studies of blood RNA in PAH have been limited by patient numbers and the use of microarrays, which are less sensitive than high quality RNAseq and limited by the probe set of each specific array. A recent meta-analysis of these microarray studies identified some consistent differentially expressed genes not appreciated in individual studies(6). Explanted lung tissues from late-stage PAH patients have also shown differences in RNA profiles(7). The results from both studies remain to be validated in independent cohorts.

The aim of this study was to characterise gene pathways associated with PAH and to assess their association with disease heterogeneity, specifically in terms of disease severity and outcomes including response to vasodilators and mortality, and genetic background. We compare gene expression in whole blood samples from 359 patients with idiopathic, heritable or drug induced PAH from the UK PAH Cohort study with 72 age- and sex-matched healthy volunteers without any cardiac or respiratory disease as controls. Using equal distribution of samples into a three-stage design, we identified reproducible RNA expression differences by RNAseq. Two distinct computational approaches were used to estimate white blood cell (WBC) fractions and thus account for the potential effect of different cell numbers on gene transcript levels across samples. A predictive statistical model that combined gene expression differences performed well at identifying PAH patients in a separate case-control analysis. The RNA-based score was also associated with disease severity and clinical outcomes (all-cause mortality). Enrichment of genetic variants that determine the levels of PAH RNAs was detected, implicating these RNAs in the pathogenesis of the disease.

Methods

Comprehensive methods are included in the online supplement.

Study participants and sample analysis

Patients with idiopathic, heritable or drug-induced pulmonary arterial hypertension (referred to throughout as PAH) were recruited from expert centres across the UK as part of the PAH Cohort study (www.ipahcohort.com). In each case, diagnosis was confirmed by right heart catheterisation following established international guidelines(1), which remained unchanged for the duration of this study. Healthy volunteers were recruited at the same centres and samples processed using the same standard operating procedure at all sites. All individuals gave written, informed consent with local ethical committee approval. Whole blood (3 ml) was collected in Tempus™ Blood RNA Tubes, and RNAsequencing performed using established Illumina methodologies (see online supplement for further details). Genomics data were obtained from a published PAH genome-wide association study(8).

359 PAH patients were randomised into 3 data analysis groups for RNA discovery (n=120), RNA validation (n=120) and model validation (n=119). Each of these 3 groups were then compared to an independent set of age- and sex-matched healthy volunteers as controls (n=24 in each set; Table 1 and Supplementary Figure 1).

RNAseq Data analysis

Fastq files (raw reads from RNAseq) were analysed using Salmon v0.9.1(9) and GENCODE release 28 to produce transcript abundance estimates which were converted to gene expression data using tximport in R with Rstudio(10). Quality control is detailed in the supplement.

RNAseq analysis of tissues with mixed cell types such as blood can be affected significantly by the cell composition of each sample. Two different computational approaches, CIBERsort(11) and quanTIseq(12), were used to predict white blood cell (WBC) profiles associated with PAH, which were included as covariates in secondary differential gene expression analyses (see supplement). Differential expression analysis was performed using edgeR v3.22.5(13) correcting for the 3 principal components, which each explained more than 1% of the variance in the dataset. Differentially expressed genes were defined in analyses both with and without WBC fractions as covariates in distinct discovery and validation sample sets. These sets were then combined and only significant genes ($p<0.05$) and directionally consistent in the initial analyses and meeting false discovery rate multiple test corrections (based on all detected genes, $\alpha=0.1$) in the combined analysis were taken forward. 507 genes meeting these criteria were considered to generate a model to distinguish PAH from controls. Subset selection of RNAs which best distinguish PAH in combination was performed by least absolute shrinkage and selection operator (LASSO) regression analysis, using the glmnet v2.0-18 package from CRAN, with k-fold cross-validation (k=10) selecting the largest value of lambda such that error is within 1 standard error of the minimum. This produces an RNA score from a linear weighted combination of the mRNAs identified by the LASSO analysis. Receiver operating characteristic (ROC) analysis was performed using the pROC v1.14.0 package from Bioconductor(14).

Survival curves from date of diagnostic right heart catheterisation were constructed using Kaplan-Meier estimates with left-truncation for date of sampling for this study to correct for survival bias. Differences in survival estimates were assessed by log rank test. RNA scores were also compared across disease severity markers 6-minute walk test (Spearman's rank), WHO functional class (Kruskal-Wallis ANOVA) and by cardiac biomarkers (circulating BNP or NT-proBNP as available, using cut-offs from European guidelines for risk assessment(1)).

Functional annotation and enrichment of the genes associated with PAH was performed using DAVID (david.ncifcrf.gov) and Ingenuity Pathway Analysis (IPA®) using in-built false discovery rate corrections for multiple tests.

Mendelian randomisation analysis using all independent genome-wide significant whole blood expression quantitative trait loci (eQTLs) from two published studies(15, 16) and PAH association from our published GWAS(8) was performed using the TwoSampleMR package(17).

Results

Identification of RNAs differentially expressed in PAH patients compared to controls

The three-stage study design is depicted in Figure 1. 507 RNAs were significantly different ($p<0.05$) between PAH and controls with directional concordance in both discovery and validation analyses (Figure 2 and Supplementary Table 3) before and after accounting for WBC fractions (see Methods and Supplementary Tables 1-3 and Supplementary Figure 2). All 507 genes were included after accounting for multiple testing (FDR $\alpha<0.1$, Supplementary Table 3). None were associated with exposure to the main PAH therapies (see supplement for further details). These included RNAs in pathways relevant to PAH; for example, *SMAD5* (Mothers Against Decapentaplegic Homolog-5), encoding a downstream mediator of signalling of *BMPR2*(3), was reduced in PAH patients (Figure 2C), consistent with documented reduced BMPR2 signalling in this condition, and the transient receptor potential cation channel, *TRPC1* (Figure 2D), also associated with the development of PH(18) was also reduced in patients.

A set of differentially expressed RNAs relevant to PAH pathobiology were selected for validation of RNAseq quantification by qPCR, namely, *NRG1, TRPC1, FBLN2, SESN1, SMAD5* and *CCND3*. *SPAST* was also selected for its stability across samples as a potential control gene (see Methods for details). All qPCR measurements correlated significantly with RNAseq quantification ($p<0.05$, Spearman's rho = 0.3 - 0.89, Supplementary Table 4).

Patients and controls from the RNA discovery and validation analyses were then combined and LASSO analysis was used to determine the combination of RNAs that performed best in discriminating PAH patients from controls in a single model; this analysis yielded 25 RNAs (Supplementary Table 5). The model was then tested in an independent validation set (n=119 PAH patients and n=24 controls) and demonstrated an area-under-the-curve of 0.868, 95% CI: 0.791-0.945 (Figure 3). The optimum cut-off for identifying PAH with the LASSO model of 1.768 recognised 88.9% of patients with 72.2% specificity.

Survival association of RNA model in clinically diagnosed idiopathic and heritable PAH

We next examined whether the RNA model was also associated with patients with the poorest outcomes, using all 359 patients. The optimum cut-off (1.910) for identifying PAH non-survivors with the LASSO model separated PAH patients into low- and high-risk groups in survival analysis (Figure 4A-B and Supplementary Figure 3).

To determine which of the 25 transcripts in the model were responsible for the association with survival, we tested each transcript for associations with all-cause mortality during follow-up. After false discovery rate correction for multiple testing, three intronic long non-coding RNAs (RP4-751H13.6/ATP6V0E2-AS1, RP4-534N18.4/AL136115.3/Lnc-PTP4A2-13, RP11-701H24.5/lnc-SNRPN-1:6) were associated with survival in all the PAH patients analysed and cut-offs distinguished high- and low-risk patient subgroups (Supplementary Table 6 and Supplementary Figure 4).

To further characterise the RNA signature, we analysed its association with three clinical measures of disease severity – WHO functional class, exercise capacity (6-minute walk) and cardiac biomarkers (BNP or NT-proBNP). We found a significant difference in RNA scores between patients in different

WHO functional classes (p= 0.008; Figure 4C) and BNP/NT-proBNP levels (p= $5.10x10^{-4}$; Figure 4D) and a significant negative correlation with exercise capacity (Spearman's rho= -0.256, p= $8.7x10^{-5}$).

RNA profiles in responders to vasodilator therapy

Several RNAs have been proposed as biomarkers for identifying vasoresponders to calcium antagonists using cultured lymphocytes(19). We compared the transcriptome of the 17 vasoresponders and 223 non-vasoresponders in the RNA discovery and first validation groups but found no transcripts consistently distinguished responders and non-responders, and that none of the previously implicated RNAs were even nominally different (Supplementary Table 7).

Comparison with differentially expressed genes identified in previous studies

A meta-analysis of previous transcriptomic studies in blood samples from PAH patients(6) and a study of PAH lung tissue samples(7) were compared to results from the current RNAseq analysis.

416 out of the 507 top dysregulated genes from the current RNAseq study were present in the PAH blood transcriptomic meta-analysis(6). Out of those, 300/416 (72%) were directionally consistent with results from the current RNAseq study. 126/300 (42%) of these were nominally significant and 70 met FDR corrected significance. 37/70 genes also met FDR corrected significance in the IPAH subgroup analysis from that same study (Supplementary Table 8).

372 out of the 507 top dysregulated genes from the current RNAseq study were present in the lung tissue microarray study and 161/372 (43%) were also directionally consistent. 41/161 (25%) genes were nominally significant and 26 met FDR corrected significance (Supplementary Table 9). Only one gene was found to be dysregulated in PAH patients across all three studies; *AMD1* (encoding a polyamine biosynthesis intermediate enzyme, adenosylmethionine decarboxylase 1) was consistently lower in PAH.

Functional characterisation of RNAs related to PAH

Of the 507 RNAs found to be differentially expressed in PAH, 435 were present in the functional annotations in DAVID. Enrichment of DNA-binding transcription factors (TFs), such as hypoxia-inducible factor 1α (HIF1α) and Krüppel-like factor 10 (KLF10), and many zinc-finger containing TFs was observed compared to a background of the genes detected (Supplementary Table 10). The Ingenuity Knowledge Base mapped 505/507 transcripts. Double-stranded DNA repair, T cell receptor, PI3K signaling in B lymphocytes, the role of JAK family kinases in IL6-type cytokine signaling and hypoxia signaling were among the top canonical pathways identified by IPA® (Supplementary Figure 5; Supplementary Table 11). *AMD1*, the only gene in common between the current RNAseq study, the recent lung tissue microarray study and the gene expression meta-analysis, was part of the top IPA® gene network (Supplementary Figure 6).

Mendelian randomisation (MR) analysis for association of RNAs with PAH development

To determine which of the 507 RNAs associated with PAH are most likely to be causal in disease pathogenesis, we performed a two-sample MR analysis using whole blood expression quantitative trait loci (eQTL) from two population-based studies(15, 16) and summary statistics from a published PAH genome-wide association study of 2085 PAH patients and 9659 controls(8). MR analysis

determines whether genetic variation associated with a trait (e.g. high or low RNA expression) is itself associated with a phenotype, in this case the development of PAH. In this exploratory analysis where eQTL were available for 293/507 RNAs, two genes, *SESN1* (Sestrin-1) and *SMAD5*, reached nominal significance using eQTL in both data sets(15, 16); eleven more reached nominal significance using eQTL from one or other of the two studies (Supplementary Tables 12 & 13).

The *SMAD5* eQTL SNP rs4146187 was clearly associated with *SMAD5* RNA levels in PAH patients in this study ($p=3.56x10^{-6}$, Figure 5; gnomAD database allele frequency in non-Finnish European Population = 0.275). Patients with the A/A genotype had comparable *SMAD5* RNA levels to controls, whereas patients with the C/C genotype had a median 18% lower *SMAD5* RNA level. The C/C genotype was present in 49.4% of PAH patients, and in the PAH genome-wide association study(8), each copy of the A allele (associated with higher *SMAD5* levels) was associated with an 8.5% reduction in the risk of developing PAH (odds ratio:0.915, 95% confidence intervals:0.846-0.990, p=0.0266). *SMAD5* levels were similarly reduced in PAH patients with and without pathogenic *BMPR2* variants (Supplementary Figure 7A), supporting the observation that impaired signalling in the BMPR2 pathway is more common in PAH than rare mutations in *BMPR2* suggest. The PAH RNA model score was similarly elevated in PAH patients with and without pathogenic *BMPR2* variants (Supplementary Figure 7B).

Discussion

Here we report a RNA signature that separates idiopathic and heritable PAH from healthy individuals. The signature also stratifies patients according to disease severity and risk of early death, adding plausibility to its association with PAH. Several of the discriminating mRNAs encode transcription factors, including SMAD5, HIF-1α and KLF10. Mendelian randomisation analysis to integrate genomic data and identify underlying pathogenic signalling pathways revealed that genetic variants associated with lower expression levels of *SMAD5* are more common in PAH patients.

Mendelian randomisation is a powerful tool for separating cause from consequence, as an individual's genetic status pre-dates the development of PAH. We harnessed genetic data from a published international genome-wide association study(8) and information on previously identified eQTL, common genetic variants which alter gene expression levels, from large published studies in whole blood RNA samples(15, 16). The identification of SMAD5 by this unbiased strategy has biological plausibility as *SMAD5* encodes an intracellular transcriptional modulator which is activated by ligand binding of BMPR2, the most common genetic risk factor in heritable PAH(3). SMAD5 also has BMP-independent roles, including regulation of cellular metabolism, directly interacting with hexokinase to increase glycolysis and responding to changes in intracellular pH(20). SMAD5 also controls levels of the master iron regulator hepcidin(21), which may be elevated and drive iron deficiency associated with poor outcomes in PAH(22). Novel therapeutics under development target restoration of BMPR2 signalling in a variety of ways, including ligand or co-activator replacement and allowing read-through of premature stop codons(23-25). Whole blood SMAD5 profiling may represent an accessible measurement of BMPR2 pathway dysfunction in PAH patients.

We were able to externally validate our findings with a meta-analysis of published PAH blood transcriptome studies(6) and a PAH lung RNA profiling study(7). The concordance of our whole blood RNAseq findings is better with blood transcriptomics than with lung tissue data, as might be expected. *AMD1* was the only gene consistently dysregulated in both these studies and our dataset, with lower levels in PAH samples in all three studies. *AMD1* encodes a key enzyme controlling the supply of decarboxylated S-adenosylmethionine for polyamine biosynthesis and is regulated by several mechanisms, including increased protein degradation in the presence of elevated polyamines and the inhibition of mRNA translation by spermidine and spermine(26). We have previously demonstrated that elevated circulating levels of polyamine metabolites, such as acisoga, 4-acetamidobutanoate and N-acetylputrescine, are associated with poor outcomes in PAH(4). Reduced *AMD1* expression in PAH patients may be in part due to negative feedback from elevated polyamines. These data contrast with observations in hypoxic rodents. *Amd1* expression is increased in hypoxic animals and both $AMD1^{+/-}$ mice and mice treated with the AMD1 inhibitor, SAM486a, were partially protected from the development of hypoxic PH(27). Interestingly, TRPC1 transcript levels were also reduced in PAH, in contrast to findings in hypoxic mice(18). This might reflect the limitations of the hypoxic rodent as a model of human PAH. Support for further investigation of polyamine metabolism in PAH comes from the recent association of rare deleterious mutations in *ATP13A3* with PAH(3). ATP13A3 is linked with polyamine biosynthesis and is a potential, target for drug development(28).

Many of the transcripts that passed robust statistical evaluation in this study are novel. We identified 3 specific lncRNAs that are reduced in PAH and associated with poor outcomes. These

transcripts are not well studied and their role in regulating lysosomal proton pump protein ATP6V0E2, protein tyrosine phosphatase PTP4A2 or small nuclear ribonucleoprotein-associated protein N remain to be established. Their association with survival selects these out as worthy of further investigation. In addition to suggesting biological relevance, the association of the RNA model, developed to distinguish PAH from controls, with survival and disease severity in PAH patients suggests that the differences observed could be useful to identify patients who may require a more aggressive treatment strategy, such as upfront, triple therapy(29). This strategy could also be used to prioritise patients more likely to have events to power clinical trials. RNA profiles and in particular, RNAs with known eQTLs, could be of further use in clinical trial design to identify patients more likely to respond to specific therapeutics. It would be of great interest, for example, to observe whether patients with lower SMAD5 levels, or simply with the variant associated with lower SMAD5 levels, showed a differential response to novel therapeutics targeting relevant signalling, such as the TGF-beta ligand trap, Sotatercept.

There is considerable interest in developing a biochemical test to identify PAH patients who respond well to calcium antagonists that would supersede the current test; namely, an acute vasodilator challenge while undergoing cardiac catheterisation. In contrast to previous reports, we were unable to demonstrate the association of a peripheral transcript signature to vasoresponder status, including previously studied RNAs(19). A strength of our study is the number of patients studied. Previous reports have relied on much smaller patient numbers. That we were able to demonstrate consistency between our main analysis results and those of a recently-published meta-analysis of PAH blood transcriptome studies(6) suggests our methodological approaches are robust.

One of the limitations of sampling whole blood to derive transcripts is that samples comprise a mixed population of cells. We used established deconvolution methods to correct for potential confounding in RNA expression analysis. In support of the validity of this approach, we noted altered numbers of regulatory T cells and CD8-positive T cells in PAH patients, consistent with previous reports (30, 31), although there is debate on precise changes in cell subpopulations and this may be due in part to differing methodologies and definitions of cell types(32).

This study does not assess the role of post-translational modifications in the pathogenesis of PAH, which could add further information to the circulating transcriptome. It is important that further mechanistic studies consider the role of the genes highlighted in this study in the tissues of primary interest, namely, the lung and heart. The study design included more patient samples than controls to account for the higher heterogeneity typically observed in patient populations. This heterogeneity is observed in the overlap in boxplots of individual RNA levels or scores between subsets of PAH patients and controls and emphasises the importance of using any molecular markers in combination with best practice clinical assessments. We chose to use LASSO regression modelling not only as it is known to perform well in these kinds of datasets but also because it is widely applied and often easier to interpret than other methodologies. Another consideration is that the patients studied here were prevalent cases. Moreover, all the patients recruited had a clinical diagnosis of idiopathic, heritable or drug-induced PAH. It would be of interest to sample patients with other presentations of PAH and other cardiopulmonary diseases to better understand the clinical utility of our RNA signature.

In summary, we report a whole blood RNA profile that distinguishes PAH patients from healthy controls and reflects disease severity. Integration with genomic data suggests that SMAD5 is important in the development of PAH, prioritising restoration of normal SMAD5 function as a target for therapeutic intervention.

Acknowledgements

The authors thank Dr. Robert L. Danner and Dr. Junfeng Sun at the National Institutes of Health Clinical Center for their discussions and for sharing results from the blood transcriptome meta-analysis, and Prof Winston Hide at Beth Israel Deaconess Medical Centre for his insights into analysis of RNAseq datasets.

We thank 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 Imperial Clinical Research Facility and NHS Blood and Transplant. 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. The UK National Cohort of Idiopathic and Heritable PAH is supported by the NIHRBR; the BHF (SP/12/12/29836) and the UK Medical Research Council (MR/K020919/1). We also gratefully acknowledge the participation of patients recruited to the US National Institutes of Health/National Heart, Lung, and Blood Institute (NIH/NHLBI)-sponsored National Biological Sample and Data Repository for PAH (also known as PAH Biobank). We also acknowledge funding from NHLBI R01HL136603 (AAD). This work was supported in part by the Assistance Publique-Hopitaux de Paris, INSERM, University Paris-Sud, and Agence Nationale de la Recherche (Departement Hospitalo-Universitaire Thorax Innovation; LabEx LERMIT, ANR-10-LABX-0033; and RHU BIO-ART LUNG 2020, ANR-15-RHUS-0002); and British Heart Foundation Centre award RE/18/4/34215. CJR is supported by a BHF Intermediate Basic Science Research fellowship (FS/15/59/31839) and Academy of Medical Sciences Springboard fellowship (SBF004\1095). LH is a recipient of ERS Fellowship (LTRF 2016–6884). AART is supported by a BHF Intermediate Clinical Fellowship (FS/18/13/3328). LS is supported by the Wellcome Trust Institutional Strategic Support Fund (204809/Z/16/Z) awarded to St. George's, University of London. DW is supported by an Academy of Medical Sciences Springboard fellowship (SBF004\1052). NWM is a British Heart Foundation Professor and NIHR Senior Investigator. AL is supported by a BHF Senior Basic Science Research fellowship (FS/13/48/30453 & FS/18/52/33808).

Conflicts of interest

CJR declares personal consultancy fees from Actelion and United Therapeutics. JW declares personal consultancy fees from Actelion. JE declares the NIH Clinical Center PAH Program received support from Aadi Bioscience for a research coordinator but no personal financial relationship with Aadi or any other entity. DK reports grants, personal fees and non-financial support from Actelion, Bayer, GSK and MSD, outside the submitted work. JGC declares Consultancy & Speakers fees J&J; GSK; Bayer, Grants: J&J; Conference fees: Bayer. AL declares outside of scope and grant funding from BHF, MRC, Actelion, GSK; Conference support from Actelion; Consultancy from Actelion, GSK. AART declares non-financial support from Actelion to attend educational events. MH reports consultancy fees from Acceleron, Actelion, Bayer, GSK, Merck, and United Therapeutics, outside the submitted work. MRW reports consultancy fees from Actelion, GSK and MorphogenIX. All other authors report no conflicts of interest.

References

1. Galie N, Humbert M, Vachiery JL, Gibbs S, Lang I, Torbicki A, Simonneau G, Peacock A, Vonk Noordegraaf A, Beghetti M, Ghofrani A, Gomez Sanchez MA, Hansmann G, Klepetko W, Lancellotti P, Matucci M, McDonagh T, Pierard LA, Trindade PT, Zompatori M, Hoeper M. 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): Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC), International Society for Heart and Lung Transplantation (ISHLT). *Eur Respir J* 2015; 46: 903-975.
2. Rhodes CJ, Wharton J, Ghataorhe P, Watson G, Girerd B, Howard LS, Gibbs JSR, Condliffe R, Elliot CA, Kiely DG, Simonneau G, Montani D, Sitbon O, Gall H, Schermuly RT, Ghofrani HA, Lawrie A, Humbert M, Wilkins MR. Plasma proteome analysis in patients with pulmonary arterial hypertension: an observational cohort study. *Lancet Respir Med* 2017; 5: 717-726.
3. Graf S, Haimel M, Bleda M, Hadinnapola C, Southgate L, Li W, Hodgson J, Liu B, Salmon RM, Southwood M, Machado RD, Martin JM, Treacy CM, Yates K, Daugherty LC, Shamardina O, Whitehorn D, Holden S, Aldred M, Bogaard HJ, Church C, Coghlan G, Condliffe R, Corris PA, Danesino C, Eyries M, Gall H, Ghio S, Ghofrani HA, Gibbs JSR, Girerd B, Houweling AC, Howard L, Humbert M, Kiely DG, Kovacs G, MacKenzie Ross RV, Moledina S, Montani D, Newnham M, Olschewski A, Olschewski H, Peacock AJ, Pepke-Zaba J, Prokopenko I, Rhodes CJ, Scelsi L, Seeger W, Soubrier F, Stein DF, Suntharalingam J, Swietlik EM, Toshner MR, van Heel DA, Vonk Noordegraaf A, Waisfisz Q, Wharton J, Wort SJ, Ouwehand WH, Soranzo N, Lawrie A, Upton PD, Wilkins MR, Trembath RC, Morrell NW. Identification of rare sequence variation underlying heritable pulmonary arterial hypertension. *Nat Commun* 2018; 9: 1416.
4. Rhodes CJ, Ghataorhe P, Wharton J, Rue-Albrecht KC, Hadinnapola C, Watson G, Bleda M, Haimel M, Coghlan G, Corris PA, Howard LS, Kiely DG, Peacock AJ, Pepke-Zaba J, Toshner MR, Wort SJ, Gibbs JS, Lawrie A, Graf S, Morrell NW, Wilkins MR. Plasma Metabolomics Implicates Modified Transfer RNAs and Altered Bioenergetics in the Outcomes of Pulmonary Arterial Hypertension. *Circulation* 2017; 135: 460-475.
5. Rabinovitch M, Guignabert C, Humbert M, Nicolls MR. Inflammation and immunity in the pathogenesis of pulmonary arterial hypertension. *Circ Res* 2014; 115: 165-175.
6. Elinoff JM, Mazer AJ, Cai R, Lu M, Graninger G, Harper B, Ferreyra GA, Sun J, Solomon MA, Danner RL. Meta-analysis of Blood Genome-Wide Expression Profiling Studies in Pulmonary Arterial Hypertension. *Am J Physiol Lung Cell Mol Physiol* 2019.
7. Stearman RS, Bui QM, Speyer G, Handen A, Cornelius AR, Graham BB, Kim S, Mickler EA, Tuder RM, Chan SY, Geraci MW. Systems Analysis of the Human Pulmonary Arterial Hypertension Lung Transcriptome. *Am J Respir Cell Mol Biol* 2019; 60: 637-649.
8. Rhodes CJ, Batai K, Bleda M, Haimel M, Southgate L, Germain M, Pauciulo MW, Hadinnapola C, Aman J, Girerd B, Arora A, Knight J, Hanscombe KB, Karnes JH, Kaakinen M, Gall H, Ulrich A, Harbaum L, Cebola I, Ferrer J, Lutz K, Swietlik EM, Ahmad F, Amouyel P, Archer SL, Argula R, Austin ED, Badesch D, Bakshi S, Barnett C, Benza R, Bhatt N, Bogaard HJ, Burger CD, Chakinala M, Church C, Coghlan JG, Condliffe R, Corris PA, Danesino C, Debette S, Elliott CG, Elwing J, Eyries M, Fortin T, Franke A, Frantz RP, Frost A, Garcia JGN, Ghio S, Ghofrani HA, Gibbs JSR, Harley J, He H, Hill NS, Hirsch R, Houweling AC, Howard LS, Ivy D, Kiely DG, Klinger J, Kovacs G, Lahm T, Laudes M, Machado RD, MacKenzie Ross RV, Marsolo K, Martin LJ, Moledina S, Montani D, Nathan SD, Newnham M, Olschewski A, Olschewski H, Oudiz RJ, Ouwehand WH, Peacock AJ, Pepke-Zaba J, Rehman Z, Robbins I, Roden DM, Rosenzweig EB, Saydain G, Scelsi L, Schilz R, Seeger W, Shaffer CM, Simms RW, Simon M, Sitbon O, Suntharalingam J, Tang H, Tchourbanov AY, Thenappan T, Torres F, Toshner MR, Treacy CM, Vonk Noordegraaf A, Waisfisz Q, Walsworth AK, Walter RE, Wharton J, White RJ, Wilt J, Wort SJ, Yung D, Lawrie A, Humbert M, Soubrier F, Tregouet DA, Prokopenko I, Kittles R, Graf S, Nichols WC, Trembath RC, Desai AA, Morrell NW, Wilkins MR, Consortium UNBRD,

Consortium UPCS, Consortium UPB. Genetic determinants of risk in pulmonary arterial hypertension: international genome-wide association studies and meta-analysis. *Lancet Respir Med* 2019; 7: 227-238.
9. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017; 14: 417-419.
10. R core team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/. 2016.
11. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* 2015; 12: 453-457.
12. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, Krogsdam A, Loncova Z, Posch W, Wilflingseder D, Sopper S, Ijsselsteijn M, Brouwer TP, Johnson D, Xu Y, Wang Y, Sanders ME, Estrada MV, Ericsson-Gonzalez P, Charoentong P, Balko J, de Miranda N, Trajanoski Z. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Genome Med* 2019; 11: 34.
13. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010; 26: 139-140.
14. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011; 12: 77.
15. Joehanes R, Zhang X, Huan T, Yao C, Ying SX, Nguyen QT, Demirkale CY, Feolo ML, Sharopova NR, Sturcke A, Schaffer AA, Heard-Costa N, Chen H, Liu PC, Wang R, Woodhouse KA, Tanriverdi K, Freedman JE, Raghavachari N, Dupuis J, Johnson AD, O'Donnell CJ, Levy D, Munson PJ. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biol* 2017; 18: 16.
16. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, Christiansen MW, Fairfax BP, Schramm K, Powell JE, Zhernakova A, Zhernakova DV, Veldink JH, Van den Berg LH, Karjalainen J, Withoff S, Uitterlinden AG, Hofman A, Rivadeneira F, Hoen PAC, Reinmaa E, Fischer K, Nelis M, Milani L, Melzer D, Ferrucci L, Singleton AB, Hernandez DG, Nalls MA, Homuth G, Nauck M, Radke D, Volker U, Perola M, Salomaa V, Brody J, Suchy-Dicey A, Gharib SA, Enquobahrie DA, Lumley T, Montgomery GW, Makino S, Prokisch H, Herder C, Roden M, Grallert H, Meitinger T, Strauch K, Li Y, Jansen RC, Visscher PM, Knight JC, Psaty BM, Ripatti S, Teumer A, Frayling TM, Metspalu A, van Meurs JBJ, Franke L. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* 2013; 45: 1238-1243.
17. Burgess S, Dudbridge F, Thompson SG. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Stat Med* 2016; 35: 1880-1906.
18. Malczyk M, Veith C, Fuchs B, Hofmann K, Storch U, Schermuly RT, Witzenrath M, Ahlbrecht K, Fecher-Trost C, Flockerzi V, Ghofrani HA, Grimminger F, Seeger W, Gudermann T, Dietrich A, Weissmann N. Classical transient receptor potential channel 1 in hypoxia-induced pulmonary hypertension. *Am J Respir Crit Care Med* 2013; 188: 1451-1459.
19. Hemnes AR, Trammell AW, Archer SL, Rich S, Yu C, Nian H, Penner N, Funke M, Wheeler L, Robbins IM, Austin ED, Newman JH, West J. Peripheral blood signature of vasodilator-responsive pulmonary arterial hypertension. *Circulation* 2015; 131: 401-409; discussion 409.
20. Fang Y, Liu Z, Chen Z, Xu X, Xiao M, Yu Y, Zhang Y, Zhang X, Du Y, Jiang C, Zhao Y, Wang Y, Fan B, Terheyden-Keighley D, Liu Y, Shi L, Hui Y, Zhang X, Zhang B, Feng H, Ma L, Zhang Q, Jin G, Yang Y, Xiang B, Liu L, Zhang X. Smad5 acts as an intracellular pH messenger and maintains bioenergetic homeostasis. *Cell Res* 2017; 27: 1083-1099.
21. Wang CY, Core AB, Canali S, Zumbrennen-Bullough KB, Ozer S, Umans L, Zwijsen A, Babitt JL. Smad1/5 is required for erythropoietin-mediated suppression of hepcidin in mice. *Blood* 2017; 130: 73-83.

22. Rhodes CJ, Howard LS, Busbridge M, Ashby D, Kondili E, Gibbs JS, Wharton J, Wilkins MR. Iron deficiency and raised hepcidin in idiopathic pulmonary arterial hypertension: clinical prevalence, outcomes, and mechanistic insights. *J Am Coll Cardiol* 2011; 58: 300-309.
23. Drake KM, Dunmore BJ, McNelly LN, Morrell NW, Aldred MA. Correction of nonsense BMPR2 and SMAD9 mutations by ataluren in pulmonary arterial hypertension. *Am J Respir Cell Mol Biol* 2013; 49: 403-409.
24. Long L, Ormiston ML, Yang X, Southwood M, Graf S, Machado RD, Mueller M, Kinzel B, Yung LM, Wilkinson JM, Moore SD, Drake KM, Aldred MA, Yu PB, Upton PD, Morrell NW. Selective enhancement of endothelial BMPR-II with BMP9 reverses pulmonary arterial hypertension. *Nat Med* 2015; 21: 777-785.
25. Spiekerkoetter E, Sung YK, Sudheendra D, Scott V, Del Rosario P, Bill M, Haddad F, Long-Boyle J, Hedlin H, Zamanian RT. Randomised placebo-controlled safety and tolerability trial of FK506 (tacrolimus) for pulmonary arterial hypertension. *Eur Respir J* 2017; 50.
26. Casero RA, Jr., Murray Stewart T, Pegg AE. Polyamine metabolism and cancer: treatments, challenges and opportunities. *Nat Rev Cancer* 2018; 18: 681-695.
27. Weisel FC, Kloepping C, Pichl A, Sydykov A, Kojonazarov B, Wilhelm J, Roth M, Ridge KM, Igarashi K, Nishimura K, Maison W, Wackendorff C, Klepetko W, Jaksch P, Ghofrani HA, Grimminger F, Seeger W, Schermuly RT, Weissmann N, Kwapiszewska G. Impact of S-adenosylmethionine decarboxylase 1 on pulmonary vascular remodeling. *Circulation* 2014; 129: 1510-1523.
28. Madan M, Patel A, Skruber K, Geerts D, Altomare DA, Iv OP. ATP13A3 and caveolin-1 as potential biomarkers for difluoromethylornithine-based therapies in pancreatic cancers. *Am J Cancer Res* 2016; 6: 1231-1252.
29. Sitbon O, Jais X, Savale L, Cottin V, Bergot E, Macari EA, Bouvaist H, Dauphin C, Picard F, Bulifon S, Montani D, Humbert M, Simonneau G. Upfront triple combination therapy in pulmonary arterial hypertension: a pilot study. *Eur Respir J* 2014; 43: 1691-1697.
30. Ulrich S, Nicolls MR, Taraseviciene L, Speich R, Voelkel N. Increased regulatory and decreased CD8+ cytotoxic T cells in the blood of patients with idiopathic pulmonary arterial hypertension. *Respiration* 2008; 75: 272-280.
31. Huertas A, Phan C, Bordenave J, Tu L, Thuillet R, Le Hiress M, Avouac J, Tamura Y, Allanore Y, Jovan R, Sitbon O, Guignabert C, Humbert M. Regulatory T Cell Dysfunction in Idiopathic, Heritable and Connective Tissue-Associated Pulmonary Arterial Hypertension. *Chest* 2016; 149: 1482-1493.
32. Qiu H, He Y, Ouyang F, Jiang P, Guo S, Guo Y. The Role of Regulatory T Cells in Pulmonary Arterial Hypertension. *J Am Heart Assoc* 2019; 8: e014201.
33. Kosinova L, Cahova M, Fabryova E, Tycova I, Koblas T, Leontovyc I, Saudek F, Kriz J. Unstable Expression of Commonly Used Reference Genes in Rat Pancreatic Islets Early after Isolation Affects Results of Gene Expression Studies. *PLoS One* 2016; 11: e0152664.

	A - RNA discovery		B - RNA validation		C - Model validation	
	Controls	PAH	Controls	PAH	Controls	PAH
Female	17	89	16	80	17	86
Male	7	31	8	40	7	33
Age	43.9 (30.2 - 53.4)	45.2 (35.7 - 55)	45.1 (31 - 53.5)	43.2 (33.3 - 54.8)	44.4 (31.7 - 51.6)	47.1 (36.7 - 61.2)

PAH Patient characteristics in A, B & C	median (25% - 75%) or counts
Age at diagnosis	44.8 (34.3 - 58.1)
Female/Male	255 / 104
Ethnicity: white/other	314 / 45
WHO functional class: I/II/III/IV	36 / 138 / 155 / 18
Six-minute walk distance, m	330 (224.5 - 411)
Mean pulmonary artery pressure, mmHg	53 (46 - 61)
Mean right atrial pressure, mmHg	8 (6 - 12)
Pulmonary capillary wedge pressure, mmHg	10 (7 - 12)
Cardiac output, L/min	3.8 (3.05 - 4.9)
Cardiac index, L/min/m2	2.07 (1.68 - 2.57)
Pulmonary vascular resistance, Wood units	11.55 (7.93 - 15.78)
Years since diagnosis sampled	3.95 (1.38 - 7.7)
Years survived since sampling	3.14 (2.3 - 3.67)
Years survived since diagnosis	6.99 (4.19 - 10.88)
Vasoresponders	21

Table 1 - Basic demographics of controls and PAH patients in three analysis groups, and more detailed clinical characteristics including disease severity of PAH patients as a cohort. Controls are healthy volunteers without any cardiac or respiratory disease.

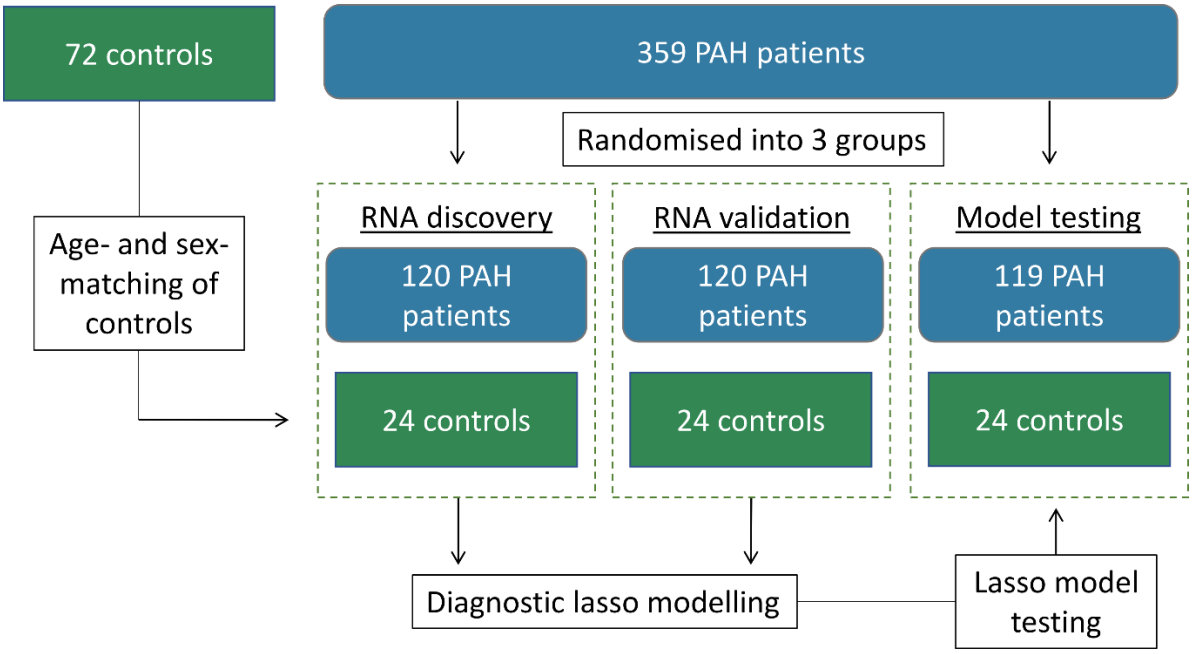

Figure 1 – Study design. 359 consecutively recruited PAH patients from the UK National Cohort study of PAH were randomised into 3 groups. Age and sex-matched controls for each group were then identified from the same study. The first two groups were analysed separately for RNA discovery and RNA validation and then combined for modelling of the best combination of RNAs to distinguish between controls and PAH. This model was then tested in the final group of individuals. All patients were combined for subsequent clinical analyses.

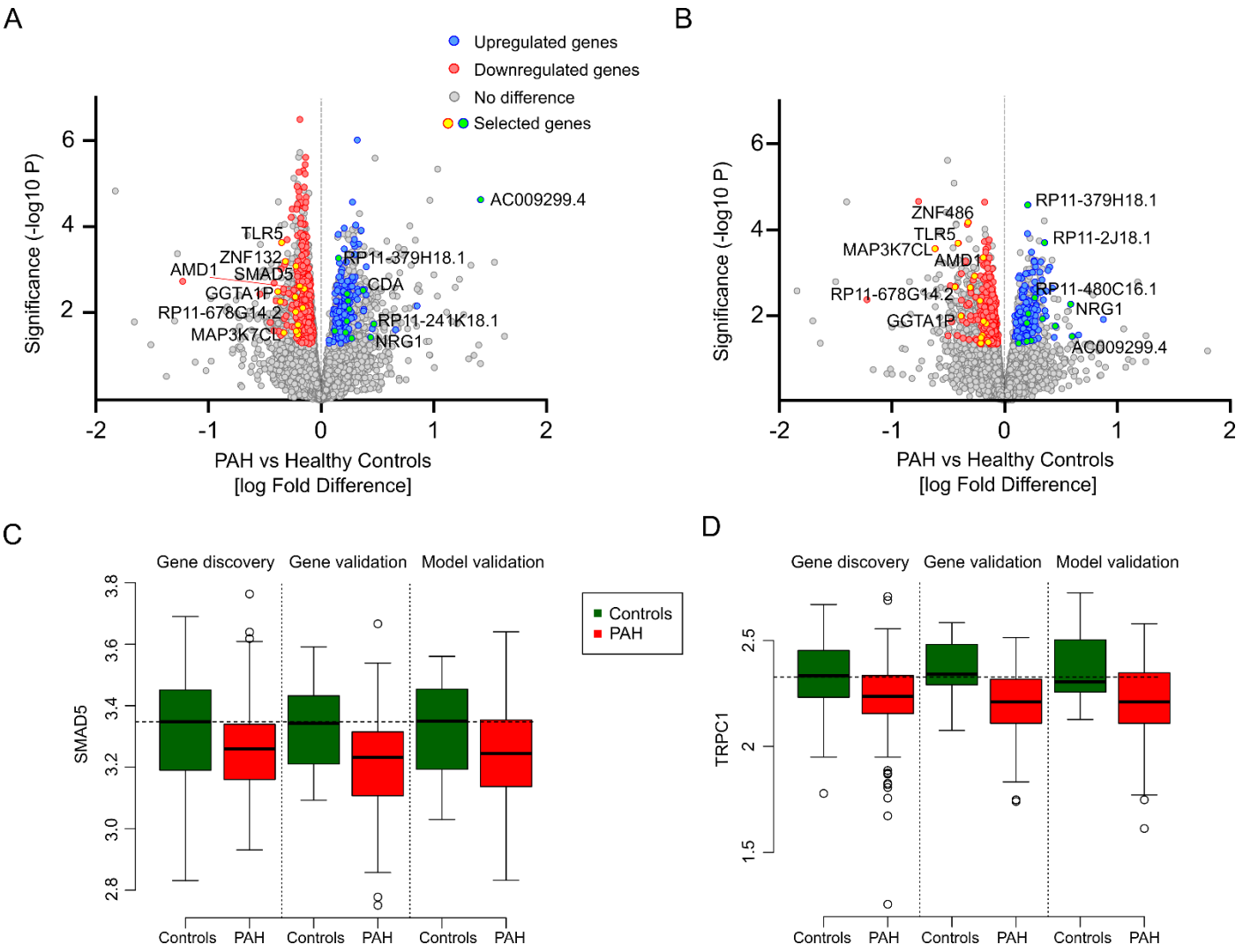

Figure 2: Identification of RNAs with different levels in PAH patients and controls. A&B. Volcano plot of RNA log fold differences between PAH and controls in A. discovery and B. validation analyses. Selected genes were identified in further analyses presented throughout manuscript including modelling, external validation and Mendelian randomisation. C-D. Boxplots of C. SMAD5 and D. TRPC1 levels (log10 reads) in controls and PAH patients in the RNA discovery, validation and model validation (validation 2) analysis groups.

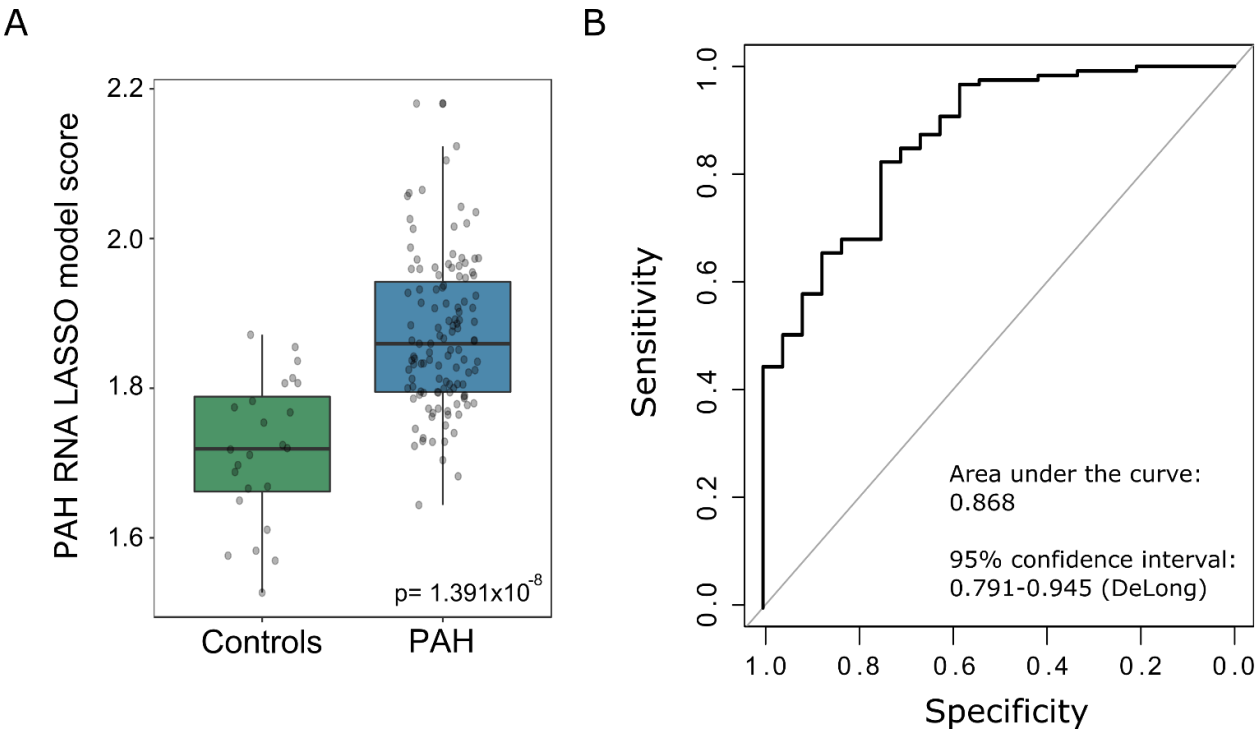

Figure 3 – LASSO model performance in an independent validation group of age- and sex-matched subjects. A. Boxplot showing LASSO model scores for controls (n=24) and PAH patients (n=119). B. Receiver operating curve showing the performance of LASSO model scores for determining PAH status in the model validation group.

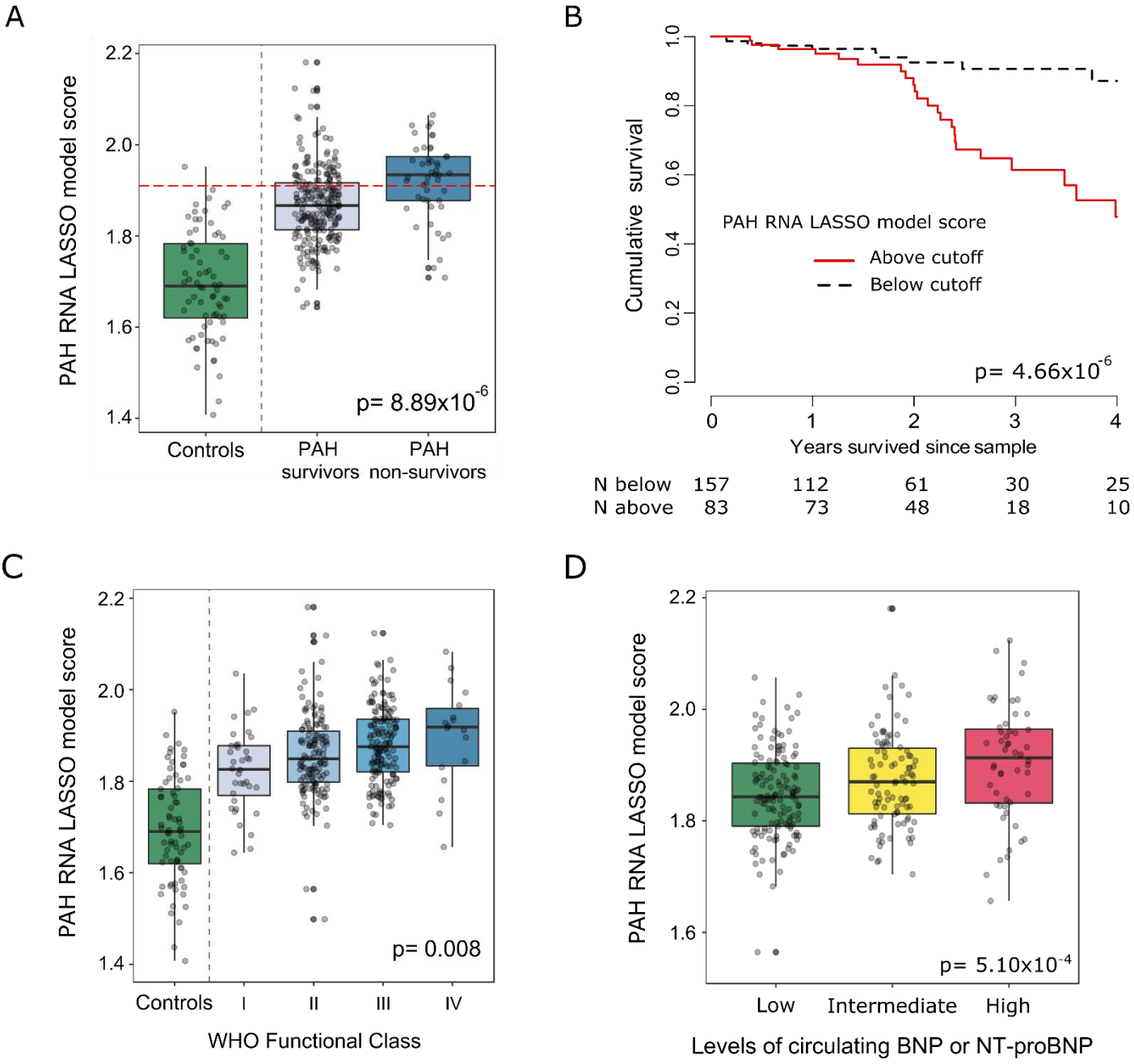

Figure 4 - Diagnostic RNA model and survival in PAH. A. Boxplot of LASSO model score in controls and PAH patients separated by survival status during follow-up. Dashed line shows the cut-off that identified 88.9% of non-surviving PAH patients. B. Kaplan-Meier survival plot separating patients on basis of the 88.9% sensitive cut-off. C. Boxplot of RNA model score in healthy volunteers and PAH patients divided by WHO functional class. D. Boxplot of RNA model score in PAH patients divided by presence of low, intermediate or high levels of cardiac biomarkers BNP (<50 pg/ml, 50-300 pg/ml or >300pg/ml) or NT-proBNP (<300 pg/ml, 300-1400 pg/ml or >1400 pg/ml), as per European guidelines for PAH assessment.

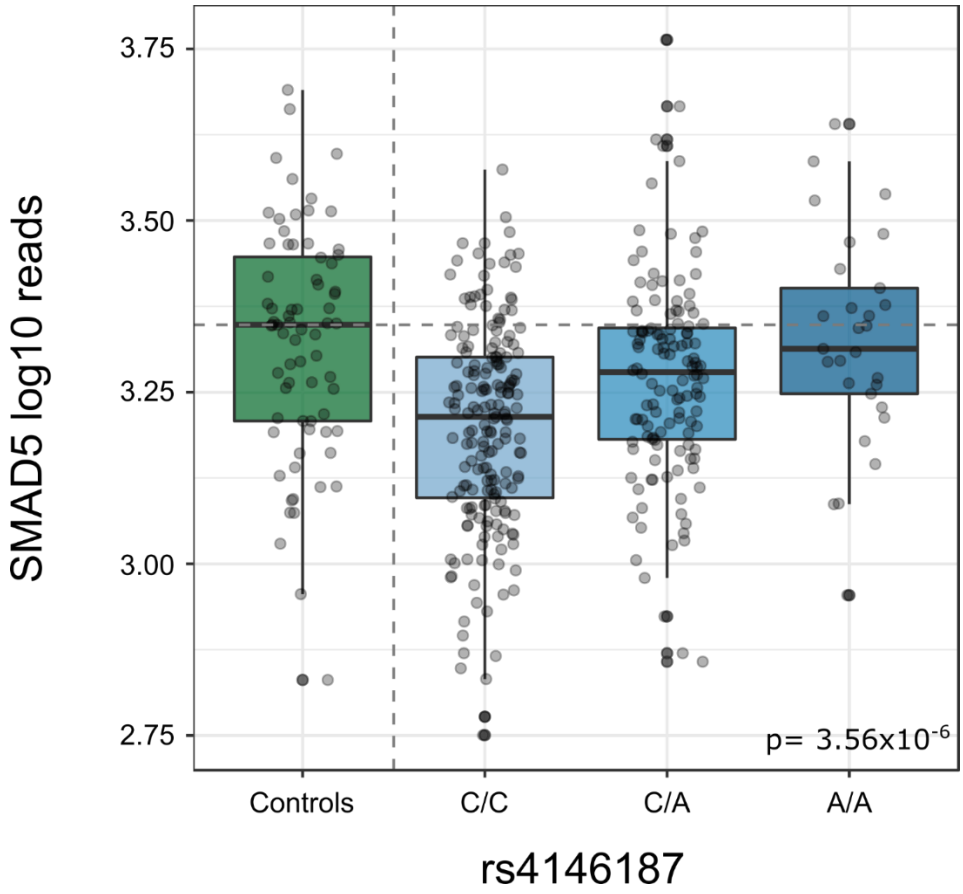

Figure 5 - Whole blood RNA levels of SMAD5 in controls, and PAH patients stratified by genotype at the SMAD5 eQTL rs4146187.