Reduced transfer coefficient of carbon monoxide in pulmonary arterial hypertension implicates rare protein-truncating variants in *KDR*

Genotype-phenotype inference reveals novel PAH risk genes

Emilia M. Swietlik, MD[1], Daniel Greene, PhD[2,3], Na Zhu, PhD[4,5], Karyn Megy, PhD[2,3], Marcella Cogliano, MSc[6], Smitha Rajaram, MD[7], Divya Pandya, MSc[1], Tobias Tilly, MSc[1], Katie A. Lutz, BS[8], Carrie C. L. Welch, PhD[4], Michael W. Pauciulo, BS, MBA[8,9], Laura Southgate, PhD[10], Jennifer M. Martin, MSt[3], Carmen M. Treacy, BSc[1], Harm J. Bogaard, MD, PhD[11], Colin Church, PhD[12], Gerry Coghlan, MD[13], Anna W. Coleman, MS[8], Robin Condliffe, MD[14], Mélanie Eyries, PhD[15], Henning Gall, MD, PhD[16], Stefano Ghio, MD[17], Barbara Girerd, PhD[18], Simon Holden, PhD[19], Luke Howard, MD, PhD[20], Marc Humbert, MD, PhD[18], David G. Kiely, MD[14], Gabor Kovacs, MD[21,22], Jim Lordan, PhD[23], Rajiv D. Machado, PhD[10], Robert V. MacKenzie Ross, MB, BChir[24], Shahin Moledina, MBChB[25], David Montani, MD, PhD[18], Horst Olschewski, MD[21,22], Joanna Pepke-Zaba, PhD[26], Christopher J. Rhodes, PhD[20], Werner Seeger, MD[16], Florent Soubrier, MD, PhD[15], Jay Suntharalingam, MD[24], Mark R. Toshner, MD[1,26], Anton Vonk Noordegraaf, MD[11], John Wharton, PhD[20], Jim Wild, PhD[6], Stephen John Wort, PhD[20,27], NIHR BioResource for Translational Research - Rare Diseases[28], National Cohort Study of Idiopathic and Heritable PAH[29], PAH Biobank Enrolling Centers' Investigators[30], Allan Lawrie, PhD[6], Martin R. Wilkins, MD[20], Richard C. Trembath, FRCP[31], Yufeng Shen, PhD[5,32], Wendy K. Chung, MD[33], Andrew J. Swift, PhD[6], William C. Nichols, PhD[8,9], Nicholas W. Morrell, MD[1,3,19,26,$], Stefan Gräf, PhD[1,2,3,$]

[$] these authors jointly supervised this work

Affiliations

[1]Department of Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, United Kingdom.

[2]Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, United Kingdom.

[3]NIHR BioResource for Translational Research, Cambridge Biomedical Campus, Cambridge, United Kingdom.

[4]Department of Pediatrics, Columbia University, New York, United States.

[5]Department of Systems Biology, Columbia University, New York, United States.

[6]Department of Infection, Immunity and Cardiovascular Disease, University of Sheffield, Sheffield, United Kingdom.

[7]Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, United Kingdom.

[8]Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, United States.

[9]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, United States.

[10]Molecular and Clinical Sciences Research Institute, St George's, University of London, London, United Kingdom.

[11]Department of Clinical Genetics, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, Amsterdam, The Netherlands.

[12]Golden Jubilee National Hospital, Glasgow, United Kingdom.

[13]Royal Free Hospital, London, United Kingdom.

[14]Sheffield Pulmonary Vascular Disease Unit, Royal Hallamshire Hospital, Sheffield, United Kingdom.

[15]Département de génétique, hôpital Pitié-Salpêtrière, Assistance Publique-Hôpitaux de Paris, and UMR_S 1166-ICAN, INSERM, UPMC Sorbonne Universités, Paris, France.

[16]University of Giessen and Marburg Lung Center (UGMLC), member of the German Center for Lung Research (DZL) and of the Excellence Cluster Cardio-Pulmonary Institute (CPI), Giessen, Germany.

[17]Fondazione IRCCS Policlinico San Matteo, Pavia, Italy.

[18]Université Paris-Sud, Faculté de Médecine, Université Paris-Saclay; AP-HP, Service de Pneumologie, Centre de référence de l'hypertension pulmonaire; INSERM UMR_S 999, Hôpital Bicêtre, Le Kremlin-Bicêtre, Paris, France.

[19]Addenbrooke's Hospital NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, United Kingdom.

[20]National Heart & Lung Institute, Imperial College London, London, United Kingdom.

[21]Ludwig Boltzmann Institute for Lung Vascular Research, Graz, Austria.

[22]Medical University of Graz, Graz, Austria.

[23]Freeman Hospital, Newcastle upon Tyne, United Kingdom.

[24]Royal United Hospitals Bath NHS Foundation Trust, Bath, United Kingdom.

[25]Great Ormond Street Hospital, London, United Kingdom.

[26]Royal Papworth Hospital NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, United Kingdom.

[27]Royal Brompton & Harefield NHS Foundation Trust, London, United Kingdom.

[28]University of Cambridge and Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, United Kingdom.

[29]www.ipahcohort.com, Cambridge, United Kingdom.

[30]www.pahbiobank.org, Cincinnati, United States.

[31]Department of Medical and Molecular Genetics, King's College London, London, United Kingdom.

[32]Department of Biomedical Informatics, Columbia University, New York, United States.

[33]Columbia University Medical Center, New York, United States.

Corresponding authors:

Dr Stefan Gräf, PhD & Professor Nicholas W. Morrell, M.D.

Department of Medicine, University of Cambridge, Level 5, Cambridge University Hospitals, Box 157, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, United Kingdom

Email: sg550@cam.ac.uk / nwm23@cam.ac.uk, tel: (+44) 1223 588036 / (+44) 1223 331666

Total word count: 4422

Abstract

Background

To date, approximately 25% of patients with pulmonary arterial hypertension (PAH) have been found to harbour rare mutations in disease-causing genes. Given the small number of patients affected by mutations in most PAH genes, the identification of the missing heritability in PAH is challenging. We hypothesised that integrating deep phenotyping data with whole-genome sequencing data will reveal additional disease variants that are extremely rare and/or have a unique phenotypic signature.

Methods

We analysed whole-genome sequencing data from 13,037 participants enrolled in the NIHR BioResource - Rare Diseases (NIHRBR-RD) study, of which 1148 were recruited to the PAH domain. To test for genetic associations between genes and selected phenotypes of pulmonary hypertension (PH), we used the Bayesian rare-variant association method BeviMed. We defined the groups for comparison by assigning labels ('tags') inferred from the current diagnostic classification of PAH, stratification by age at diagnosis and transfer coefficient of carbon monoxide (KCO).

Results

Protein truncating variants (PTV) in *KDR* were strongly associated with the lower KCO tertile (posterior probability (PP)=0.989) and the higher age tertile (PP=0.912) groups. On computed tomographic imaging of the lungs, a range of parenchymal abnormalities were observed in the

patients harbouring PTV in *KDR*. KCO stratification also highlighted an association between Isocitrate Dehydrogenase (NAD(+)) 3 Non-Catalytic Subunit Gamma (*IDH3G*) and moderately reduced KCO in patients with pulmonary hypertension (PP=0.920). The US PAH Biobank was used to independently validate these findings and identified four additional PAH cases with PTV in *KDR* and two in *IDH3G*. We confirmed associations between previously established genes and PAH.

Conclusions

PTVs in *KDR*, the gene encoding vascular endothelial growth factor receptor 2 (VEGFR2), are significantly associated with two specific phenotypes of PAH, reduced KCO and later age of onset, highlighting a role for VEGF signalling in the pathogenesis of human PAH. We also report *IDH3G* as a new PAH risk gene. Moreover, we demonstrate that the use of deep clinical phenotyping data advances the identification of novel causative rare variants.

Introduction

Pulmonary arterial hypertension is a rare condition characterised by pulmonary vascular narrowing and obliteration, causing elevation of pulmonary vascular resistance and ultimately, right ventricular failure. Multiple concepts have been proposed to explain the mechanisms leading to pulmonary vessel remodelling[1]. More recently, hallmarks of cancer, such as aberrant angiogenesis[2], metabolic reprogramming[3] and resistance to apoptosis[4], have been proposed. A breakthrough in our understanding of PAH pathobiology was the discovery of heterozygous germline mutations in the gene encoding bone morphogenetic protein type 2 receptor (*BMPR2*)[5,6]. It is now established that *BMPR2* mutations are responsible for over 70% of familial cases of PAH (HPAH) and 15-20% of idiopathic cases of PAH (IPAH). The penetrance of *BMPR2* mutations is incomplete, so not all carriers develop the disease[7]. A smaller proportion (up to 10%) of PAH is caused by mutations in activin-like kinase 1 (*ACVRL1*)[8], endoglin (*ENG*)[9], SMAD family member 9 (*SMAD9*)[10], caveolin-1 (*CAV1*), involved in colocalization of BMP receptors[11], and the potassium channel, *KCNK3*, responsible for membrane potential and vascular tone[12]. Using burden tests, we have recently identified rare pathogenic variants in growth differentiation factor 2 (*GDF2)*, which encodes BMP9, a major ligand for BMPR2, as well as rare variants in ATPase 13A3 (*ATP13A3*), aquaporin 1 (*AQP1*) and SRY-box 17 (*SOX17*), and reported a list of additional putative genes potentially contributing to the pathobiology of PAH[13]. Together, these established genes explain approximately 25% of cases with idiopathic/hereditary pulmonary arterial hypertension (I/HPAH). To identify additional genes harbouring potentially causal rare variants in PAH cases, we increased the cohort size and deployed a Bayesian framework incorporating refined phenotype data.

Methods

Study design, ethics, and subject recruitment

The National Institute for Health Research BioResource - Rare Diseases study (NIHRBR-RD), the Rare Disease pilot for Genomics England Ltd. 100,000 Genomes Project, was established to identify genetic causes, improve rates of molecular diagnosis and develop new treatments for rare diseases through whole-genome sequencing and deep phenotyping[14]. Of the 18 domains, 15 were defined either as a single rare disease or a group of rare disorders (Table S1). The PAH domain comprised 1148 subjects including individuals diagnosed with either idiopathic or heritable PAH, pulmonary veno-occlusive disease (PVOD) or pulmonary capillary haemangiomatosis (PCH) and a small number of healthy relatives. Adult and paediatric onset cases were eligible, as well as incident and prevalent cases. Recruitment was carried out across the nine PAH specialist centres in the UK and retrospectively by international collaborators at the Université Paris-Saclay and Sorbonne Université (France), University of Giessen and Marburg (Germany), and hospitals in Graz (Austria), Pavia (Italy) and Amsterdam (The Netherlands). Patients recruited to the NIHRBR-RD study provided written, informed consent for genetic analysis and clinical data capture (REC REF: 13/EE/0325); patients recruited by European collaborators consented to genetic testing and clinical data collection locally.

Patients with rare diseases recruited to domains other than PAH were used as non-PAH controls in the genetic analysis (Table 1).

For validation, we used the US PAH Biobank cohort comprising exome sequencing data from 2572 subjects diagnosed with group 1 PAH[15] and a biobank of 440 PAH patients established at Columbia University Medical Center composed of 29 FPAH, 195 IPAH and 216 APAH individuals[16].

Phenotyping of patients

Clinical phenotyping and case-control cohort using phenotypic 'tags'

Pseudonymised results of routinely performed clinical tests were stored across twenty-one electronic Clinical Case Report Forms (eCRFs) in the *OpenClinica* data capture system (Table S2). All cases were diagnosed between January 2002 to December 2017, and the diagnostic classification was made according to international guidelines using a multidisciplinary assessment that included echocardiography, comprehensive blood testing, pulmonary function testing, overnight oximetry, isotope perfusion scanning, high-resolution computed tomography, and right heart catheterisation. To aid data analysis and improve data quality, a number of quality assurance procedures were introduced (see Supplemental Material). Diagnosis in all patients was verified based on haemodynamic criteria, reported comorbidities (history of pulmonary embolism, chronic obstructive pulmonary disease, interstitial lung disease (ILD), left heart disease, connective tissue disease, structural heart abnormalities, anorexigen use) and results of pulmonary function tests, heart and lung imaging and clinical blood tests (autoantibody screen). Cases in which the diagnosis was questionable were reported back to recruiting centres for verification. Appropriate diagnostic and phenotypic tags were assigned to all recruited patients to be used in the subsequent case-control analysis (Figure S1). The full set of tags, with corresponding numbers of cases, controls and excluded relatives, can be found in Table 1.

Analysis of computerised tomography scans

Diagnostic chest computerised tomography (CT) scans were performed and reported in 613 study participants. The analysis of these scans was done in PAH centres and subsequently transcribed to study eCRFs. Of 613 scans, 294 were available for repeated analysis. The scans

were anonymised and transferred to Department of Infection, Immunity and Cardiovascular Disease, University of Sheffield, Sheffield, UK where they were reviewed by two independent cardiothoracic radiologists with expertise in pulmonary hypertension (AS and SR), who were blinded to the underlying diagnosis, mutation and smoking status. For consistency and reproducibility, all measurements were reported on a customised proforma (Table S3).

CT scans were obtained between 2002 and 2018 (n=269), CT pulmonary angiogram (CTPA, n=241), high resolution computed tomography (HRCT no CTPA, n=28). Slice thickness was less than 5mm for all studies, typically ≤1mm. Images were analysed on open source software Horos (Annapolis, MD USA). Cardiac and vascular measurements were taken by one observer (MC) and reviewed by the Consultant Radiologist (AS). Thoracic Radiological features were scored semi-quantitatively by two independent Cardiothoracic Radiologist observers each with 9 years experience in pulmonary hypertension imaging (AS, SR) with a very good interobserver agreement (see Supplement, Table S10)

Whole-genome sequencing, short read alignment and variant calling

Samples were received as either DNA extracted from whole blood or as whole blood EDTA samples that were extracted at central DNA extraction and QC laboratory in Cambridge (UK). They were subsequently tested for adequate DNA concentration, DNA degradation and purity. Next-generation paired-end whole-genome sequencing, using three read lengths 100bp (377 samples), 125bp (3,154 samples) and 150bp (9,656 samples), was performed on cases and controls using Illumina HiSeq2500 and HiSeq X (Illumina Inc, San Diego, USA).

Reads were aligned against the Genome Reference Consortium human genome build 37 (GRCh37, https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/) using the Illumina

Isaac Aligner version SAAC00776.15.01.27[17] and variants were called using the Illumina Starling software version 2.1.4.2 (https://support.illumina.com/help/BS_App_TS_Amplicon_OLH_15055858/Content/Source/Informatics/Apps/IsaacVariantCaller_appENR.htm). The variants were then left-aligned, normalized with *bcftools* and loaded into our Hbase database to produce multi-sample variant calls to undertake the genetic association studies[14].

Genetic association between rare variants and selected diagnostic and phenotypic tags

Schematic analysis pipeline is depicted in Figure 1A. We hypothesised that groups of patients who share a particular feature may also share a similar genetic aetiology and used the current diagnostic classification of pulmonary hypertension and stratification by age at diagnosis and KCO (% predicted), to define a set of phenotypic tags (Table 1). We defined cases as individuals carrying a particular tag whereas the individuals from the non-PAH domains served as controls (Figure 1, Table S1). Variants were extracted from each gene as previously described[14] including a $PMAF_x$ (for a given variant, the probability that the minor allele count is at least the observed minor allele count, given that MAF=1/X) <0.05 with x=1,000 for the recessive and x=10,000 for the dominant association model, and a CADD Phred score ≥10. The analysis was restricted to the Ensembl annotated canonical transcript. Bayesian model comparison method called BeviMed[18] was applied to the extracted rare variants from a set of unrelated individuals to test posterior probability of gene-tag associations under dominant and recessive modes of inheritance. Patients with rare deleterious variants in previously established PAH disease genes (*BMPR2*, *ACVRL1*, *ENG*, *CAV1*, *SMAD1*, *SMAD4*, *SMAD9*, *KCNK3*, *EIF2AK4*, *TBX4*, *AQP1*,

ATP13A3, *GDF2*, *SOX17*) that were deemed disease-causing by a genetic multidisciplinary team according to the ACMG Standards and Guidelines[19], were excluded from the association testing for other genes to minimise false-positive associations. To increase power in scenarios where only variants of particular consequence types were associated with the disease risk, association models were fitted to different subsets of variants according to the consequences provided by Ensembl (https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html): the High category, comprised only variants of "high" impact, including PTVs and large deletions; the Moderate category contains variants of impact "moderate", including missense variants or consequence "non_coding_transcript_exon_variant"; the combined category Moderate and High, combining the respective consequence types. The prior probability of association across all association models was set to 0.001. Our choice of prior was informed by the estimation that approximately 30 genes might be involved in the pathogenesis of pulmonary arterial hypertension out of the 32,606 protein-coding and non-coding genes (defined by the selected gene biotypes provided by Ensembl, see supplemental material) tested after applying the filtering described above.

Descriptive statistics

Statistical analysis and data visualisation were performed in R (www.r-project.org). Summary statistics are shown as mean (±SD) or median [IQR] according to data distribution (normality testing was performed with the Shapiro-Wilk test and QQ plots). The number of available data points is reported in tables. Comparisons between the categorical variables were performed using Fisher's exact and Chi-square test, comparisons between continuous non-normally distributed variables were performed with the Mann-Whitney' test (for two groups) or the Kruskal-Wallis test (three and more groups). Adjustment for multiple comparisons was performed when

appropriate. The Kaplan-Meier method was used to visualise survival curves; the log-rank test was used to compare survival between two or more groups; Cox proportional hazards regression was used to examine the effect of variables on survival. Testing for the proportional hazards assumption, influential observations and non-linearity were done, and the assumptions were met. To measure the magnitude of agreement between CT scan readers, 22 randomly selected tests were assessed by both radiologists. For categorical variables, the weighted (ordinal data) and unweighted (for non-ordinal data) Cohen's Kappa for two readers were calculated and for continuous variables, the intraclass correlation coefficient (ICC) was computed with the R package ("irr").

Results

Characterization of study cohorts and tag definition

Whole-genome sequencing was performed in 13,037 participants of the NIHRBR-RD study, of which 1148 were recruited to the PAH domain. The PAH domain included 23 unaffected parents and 3 cases with an unknown phenotype, which were subsequently removed from the analysis (Table S1, Figure 1B). Of the remaining 1122 participants, 972 (86.6%) had a clinical diagnosis of IPAH, 73 (6.5%) of HPAH, and 20 (1.8%) were diagnosed with PVOD/PCH. Verification of diagnosis based on the collected clinical information revealed that 57 participants (5%) had a diagnosis other than IPAH, HPAH or PVOD/PCH. These cases were subsequently relabelled and moved to the respective tag group for analysis (see Table S4, Table 1). The population structure of the PAH cohort was comparable to previously studied European PAH populations, with a median age at diagnosis of 49[35;63] years, and female predominance of 68% (760 individuals). Among the most common comorbidities were hypertension (24%), diabetes mellitus type 2

(12%) and hypothyroidism (12%). Most patients were treated with combination therapies (44%) followed by monotherapy with sildenafil (24%) (Table S4). Overall survival in the studied population was 97% at 1-year, 91% at 3-years and 84% at 5-years. When the cohort was divided into prevalent and incident cases 1-, 3-, and 5-year survival was 98%, 93%, 87% and 97%, 84%, 72% respectively.

Transfer coefficient of carbon monoxide (KCO) measured at diagnosis was available for 644 patients (57%) (see Supplemental Material, Table S5, Figure S1). Median KCO in the entire studied population was 71[52;86]% predicted (Figure S2). Cases in the lower tertile or below the KCO threshold of 50% predicted were more commonly male, older at diagnosis, had a current or past history of cigarette smoking and an increased number of cardiorespiratory comorbidities (Table S6, Table S7). Survival in these groups was significantly worse than in those with preserved or mildly reduced KCO (Figure S3 A and B). Even after adjusting for confounding factors (age, sex, comorbidities, smoking status and whether the case was prevalent or incident), KCO remained an independent predictor of survival (Table S8).

Age at diagnosis was calculated as age at the time of diagnostic right heart catheter (RHC) and was available in all but 10 cases. When patients were divided by age, those in higher age tertile showed more functional impairment despite milder haemodynamics, lower FEV1/FVC ratio and KCO % predicted as well as milder emphysematous and fibrotic changes on CT scans (Figure S2 and Table S9).

Rare variants in previously established genes

We identified variants in previously established genes (namely, *BMPR2*, *ACVRL1*, *ENG*, *SMAD1*, *SMAD4*, *SMAD9*, *KCNK3*, *TBX4*, *EIF2AK4*, *AQP1*, *ATP13A3*, *GDF2*, *SOX17*) in 271 (24.2%) of the 1122 cases and interpreted them based on the ACMG standards and guidelines[19]. The

majority of these variants have already been described in Gräf *et al.*[13] (see supplemental material). Larger deletions are depicted in Figure S4 A-F.

Rare variant association testing

We used the rare variant association test BeviMed to consolidate previously reported and discover novel genotype-phenotype associations. The BeviMed analysis identified 42 significant gene-tag associations with posterior probability (PP) above 0.6 (Table 2 and Figure 2A). *BMPR2, TBX4, EIF2AK4, ACVRL1* show the highest association (PP ≥0.99) and further confirmed significant associations in the majority of other previously established genes[13]. Our analysis showed that individuals with rare variants in *BMPR2*, *TBX4*, *EIF2AK4* (autosomal recessive model) and *SOX17* have a significantly earlier age of disease onset (tag: young age). We also confirmed the association of rare variants in *AQP1* with HPAH (PP=0.994). The refined phenotype approach corroborated the association between mutations in *BMPR2* and preserved KCO (KCO higher tertile, PP=1) as well as an association between biallelic *EIF2AK4* mutations and significantly reduced KCO (KCO <50% predicted, PP=1).

Under an autosomal dominant mode of inheritance, protein-truncating variants (PTVs) in kinase insert domain receptor (*KDR)* were associated with a significantly reduced KCO (KCO lower tertile, PP=0.989), as well as older age at diagnosis (tag: old age, PP=0.912). Furthermore, KCO stratification highlighted an association between isocitrate dehydrogenase (NAD(+)) 3 non-catalytic subunit gamma (*IDH3G*) and moderately reduced KCO in patients with pulmonary hypertension (PP = 0.920).

Rare variants in the new PAH risk genes: *KDR* and *IDH3G*

We identified a total of five rare protein-truncating variants in *KDR* in the study cohort. Four of these were in PAH cases: 1 frameshift variant in exon 3 of 30 (c.183del, p.Tryp61CysfsTer16), 2 nonsense variants, one in exon 3 (c.183G>A, p.Trp61Ter) and one in exon 22 (c.3064C>T, p.Arg1022Ter) and 1 splice acceptor variant in intron 4 of 29 (c.490-1G>A). In addition, one nonsense variant was identified in exon 27 (p.Glu1206Ter) in a non-PAH control subject (Table 3). This latter nonsense variant appears late in the amino acid sequence, in exon 29 of 30, which might have limited impact on *KDR* function. Furthermore, 13 PAH cases (1%) and 102 non-PAH controls (0.9%) harboured rare predicted deleterious *KDR* missense variants (Figure 3). The missense variant carriers, however, did not exhibit a reduced KCO or older age of diagnosis. Instead, these patients show the opposite trend in KCO (Table 4 and Figure 2C and D). Importantly, seven of the 13 *KDR* missense variants seen in the PAH cases also were detected in several non-PAH controls, and thus are unlikely to be playing a causal role. Furthermore, three of the *KDR* missense variants co-occurred with predicted deleterious variants in established PAH risk genes (two patients with variants in *BMPR2* and one with variant in *AQP1*).

We also identified three missense variants (c.74C>T, p.Pro25Leu; c.1037C>T, p.Thr346Ile; c.1067T>C, p.Met356Thr) and one large deletion (X:147511939-154854072) in the gene *IDH3G* in five individuals. The missense variant (c.74C>T, p.Pro25Leu) was present in two IPAH individuals, whereas the large deletion (X:147511939-154854072) was present in one IPAH and one control case. The “Moderate and high” impact category contributed to the detected association. IPAH patients harbouring variants in *IDH3G* were all females with early-onset disease, median age 34 [27;51] and relatively preserved KCO (Table 3, Table S11).

Clinical characterisation of *KDR* mutation carriers

Patients with PTV in *KDR* were older and exhibited significantly reduced KCO compared with *KDR* missense variant carriers and *BMPR2* mutation carriers (Figure 2C). In order to determine whether the reduction in KCO was the result of coexistent emphysema secondary to smoking or other parenchymal lung diseases, we performed a detailed analysis of thoracic CT imaging. Three of the four cases did not have a history of smoking. CT scans were available in all four patients harbouring PTV in *KDR* and showed a range of lung parenchymal changes in all four cases (Figure 4). W000229 had evidence of mild mainly subpleural ILD, mild emphysema and air trapping. W000274 had signs of ILD with traction bronchiectasis in the lower zones, mild air trapping, and mild diffuse ground glass opacities (GGO) and neovascularity. Also, E001392 showed mild centrilobular ground glass nodularity in addition to moderate pleural effusion and a trace of air trapping, but no ILD. In these cases it seemed likely that the observed parenchymal changes contributed to the low KCO. In contrast, E003448 had a low KCO despite only a trace of central non-specific ground glass change on the CT images. Comparisons of CT findings between patients harbouring deleterious mutations in *BMPR2*, *EIF2AK4*, *KDR,* other PAH risk genes and patients without mutations are presented in Table S11. There were no differences in the frequency of comorbidities between patients harbouring missense and PTV in *KDR* although the frequency of systemic hypertension was high in both UK and US cohorts (44% and 50%, respectively) (Table 4 and Table S11). None of the PTV carriers had a family history of PAH. Survival in this group could not be assessed because of the small number of patients harbouring the mutation, as well as only one event occurring in this group.

Additional cases in US PAH cohorts

To replicate our findings, we analysed subjects recruited to the US PAH Biobank[15] and the Pulmonary Hypertension Center at Columbia University[16] to identify additional patients carrying predicted pathogenic rare variants in the new PAH risk genes. Four individuals harbouring *KDR* PTVs were identified. These comprised, 2 nonsense variants, one in exon 3 (c.303C>A, p.Tyr101Ter) and one in exon 22 (c.3064C>T, p.Arg1022Ter) and two splice donor variants, one in intron 2 of 29 (c.161+1G>T) and one in intron 5 (c.658+1G>A). Interestingly, the nonsense variant p.Arg1022Ter appeared in both cohorts (Figure 3). Patient-level data for these individuals are summarised in Table S3. Three of the four patients were diagnosed with idiopathic PAH at 72, 65 and 42 years respectively, whereas one patient was diagnosed at age 4 with PAH associated with double outlet right ventricle. Diffusion capacity of carbon monoxide was available for one patient and was decreased at 35% predicted, with only minor pleural scarring in the left upper lobe found on CT imaging. Two out of four patients harbouring PTV in *KDR* had also been diagnosed with systemic hypertension.

Additionally, two individuals carrying missense variants in *IDH3G* locus were found in US PAH Biobank and Pulmonary Hypertension Center at Columbia University cohorts; one male neonate diagnosed with Scimitar syndrome, hypoplastic right lung and atrial septal defect (ASD) (c.1091C>T, p.Pro364Leu) and a 55-year-old female with large ASD (c.217G>C, p.Val73Leu).

Discussion

One of the critical translational steps in identifying novel, causative genes in rare disorders is the discovery of genotype-phenotype associations to inform patient care and impact outcomes. A pragmatic focus on deeply-phenotyped individuals and “smart” experimental design cannot be

overestimated[20]. With this in mind, we studied the molecular genetic architecture of PAH using BeviMed[18]. To generate case/control labels, we tagged PAH cases with diagnostic labels and stratified them by age at diagnosis and KCO. Analyses were then performed to identify associations between tags and rare gene variants.

Our findings strongly suggest a link between rare protein-truncating *KDR* variants and reduced KCO and older age at diagnosis. The human *KDR* gene, located on chromosome 4q11–q12, encodes vascular endothelial growth factor receptor 2 (VEGFR-2)[21]. VEGFR-2 is composed of an extracellular domain, which comprises seven Ig-like domains (I–VII), of which domains II and III bind VEGF-A, a critical growth factor for physiological and pathological angiogenesis in vascular endothelial cells. In mice, even though VegfA haploinsufficiency is embryonically lethal[22], heterozygosity of its receptor, Vegfr2, is compatible with life and unimpaired vascular development[23].

The role of VEGF signalling in the pathogenesis of PAH has been an area of intense interest since reports of increased expression of VEGF, VEGFR1 and VEGFR2 in rat lung tissue in response to acute and chronic hypoxia[24]. An increase in lung VEGF has also been reported in rats with PH following monocrotaline exposure[25]. In humans, VEGFA is highly expressed in plexiform lesions in patients with IPAH[26], tracheal aspirates from neonates with a persistent PH of the newborn[27] and small pulmonary arteries from infants with PH associated with a congenital diaphragmatic hernia[28]. In view of these findings, it is surprising that the overexpression of VEGFA ameliorates hypoxia-induced PAH[29]. In contrast, inhibition of VEGF signalling by SU5416 (sugen) combined with chronic hypoxia triggers severe angioproliferative PH[30]. SU5416, a small-molecule inhibitor of the tyrosine kinase segment of VEGF receptors inhibits VEGFR1[31] and VEGFR2[32] causing endothelial cell apoptosis, loss of lung capillaries and emphysema[33]. In combination with chronic hypoxia, SU5416 causes cell-death dependent compensatory pulmonary endothelial cell proliferation and severe PH[30]. Further evidence supporting the role of VEGF inhibition in the

pathobiology of PAH comes from reports of PH in patients treated with bevacizumab[34] and the multi-tyrosine kinase inhibitors, dasatinib[35] and bosutinib, have also been associated with PAH[36]. Both preclinical and patient data show that inhibition of VEGF is associated with considerable cardiovascular side effects[37]. Among common side effects of VEGF inhibitors are systemic hypertension, proteinuria, renal impairment and thyroid dysfunction. The overall incidence of systemic hypertension induced by bevacizumab and RTKIs scale from 9 to 67% and is dose-dependent[38]. Mechanisms implicated in systemic hypertensive response include impairment of nitric oxide (NO) signalling, increased arterial stiffness[39], reduced capillary density[40] or functional rarefaction[41] and activation of the endothelin system[42], all of which are relevant to the pathobiology of PAH. Notably, two out of four of our cases with PTVs at the *KDR* locus had systemic hypertension. Also, the frequency of thyroid dysfunction was higher (although not statistically significant) in patients with *KDR* PTVs (25% UK cohort, 50% US cohort) than in patients without mutations in PAH risk genes (13.2%). The proportion of patients with renal impairment was not different between *KDR* PTV and missense variant carriers or the rest of the study population. Mutations in *KDR* have also been reported in other cardiovascular diseases; Bleyl et al. reported that *KDR* might be a candidate for familial total anomalous pulmonary venous return[43]. In addition, haploinsufficiency in *KDR* locus has also been associated with tetralogy of Fallot[44]. We report one patient (US cohort) with PAH associated with congenital heart disease and *KDR* protein-truncating splice donor variant (c.161+1G>T). *IDH3G* is a protein-coding gene encoding enzyme catalyzing the decarboxylation of isocitrate (ICT) into alpha-ketoglutarate, a tricarboxylic acid (TCA) cycle intermediate. Metabolomic[3] and imaging studies[45] have previously shown disrupted bioenergetics in IPAH characterised by the accumulation of TCA cycle intermediates. This indicates suppression of mitochondrial glucose oxidation, central to which is inhibition of pyruvate dehydrogenase (PDH)[46]. Alpha-ketoglutarate is a required cofactor for PDH, the enzyme that under normal conditions causes proteasomal degradation of hypoxia-

inducible factor (HIF)[47]. Citrate and alpha-ketoglutarate have also been implicated in acetylation[48] and methylation[49] of nuclear histones. Interestingly isocitrate dehydrogenase (IDH) activity has been reported to be increased both in PAEC and serum in patients harbouring *BMPR2* pathogenic variants[50]. IDH has the capacity to catalyze against TCA flow by converting alpha-ketoglutarate to isocitrate leading to depletion of the PDH co-factor alpha-ketoglutarate and causing decreased hydroxylation of HIF leading to its proteasomal degradation[50]. Such findings have potential therapeutic implications, as pyruvate dehydrogenase kinase inhibitor (dichloroacetate) has shown some efficacy in genetically susceptible PAH patients[51].

In the present study, we highlight that deep clinical phenotyping in combination with genotype data can accelerate the identification of novel disease risk genes and disease subtypes, which may have prognostic and therapeutic implications. Of particular interest is the association of *KDR* PTVs with significantly reduced KCO. Reduced KCO, which reflects impairment of alveolar-capillary membrane function, has been noted in the analysis of early registry data[52] to be an independent predictor of survival. Decreased KCO was also found in patients with PVOD/PCH with or without biallelic EIF2AK4 mutations[53]. Although some reduction in KCO is one of the typical features of pulmonary vascular disease, PVOD patients show the lowest KCO values when compared to IPAH or CTEPH. In contrast, KCO is relatively preserved in *BMPR2* mutation carriers[54]. Strong association with survival and a link with other causative mutations makes the KCO phenotype particularly attractive for genetic studies, and KCO should be consistently collected in future PAH registries.

As lung disease should always be taken under consideration as a cause of low KCO, we applied the World Symposium on PH criteria[55] to exclude lung disease as a cause of PH: TLC ≥70% pred., FVC ≥70% pred., FEV1 ≥60% pred., and no severe fibrosis and/or emphysema on chest HRCT. None of the PTV *KDR* cases met these criteria although two of the four patients did show evidence of early ILD. Another potential reason for low KCO in the PAH population is the

diagnosis of PVOD/PCH[56]. Again, careful analysis of CT scans and clinical data did not reveal convincing evidence for this diagnosis in *KDR* PTV carriers. Cigarette smoking is a well-known factor leading to the decrease of KCO, which can be explained by increased carboxyhemoglobin levels[57] and smoking-induced emphysema[58]; only one of the 4 *KDR* PTV carriers was a previous smoker with 15 pack-years of exposure but non-smoker for over 20 years prior to diagnosis and with no signs of emphysema on CT. These findings suggest that PTVs in *KDR* are associated with a form of PAH characterised by a range of lung parenchymal abnormalities, including small airways disease, emphysema and ILD, as two of the four patients harbouring PTV in *KDR* had mild fibrotic lung changes. Of note, the patients with mutations in other PAH risk genes or those without identified genetic mutation showed less than 10% incidence of fibrotic changes on CT imaging. Further larger studies are needed to determine the full range of lung parenchymal abnormalities in PAH cases with PTVs in *KDR*.

In summary, this study shows how deep phenotyping enabled patient stratification into subgroups with shared pathobiology and increased the power to detect genotype-phenotype associations. We provided statistical evidence of a strong association between PTVs in the gene *KDR* and significantly decreased KCO as well as later age of disease onset, and moderate impact variants in *IDH3G* and preserved KCO.

Acknowledgements

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 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.

We thank the research nurses and coordinators at the specialist pulmonary hypertension centres involved in this study. We acknowledge the support of the Imperial NIHR Clinical Research Facility, the Netherlands CardioVascular Research Initiative, the Dutch Heart Foundation, Dutch Federation of University Medical Centres, the Netherlands Organisation for Health Research and Development and the Royal Netherlands Academy of Sciences. We thank all the patients and their families who contributed to this research and the Pulmonary Hypertension Association (UK) for their support.

We thank contributors, including the Pulmonary Hypertension Centers who collected samples used in this study, as well as patients and their families, whose help and participation made this work possible. Exome sequencing and genotyping data were generated by the Regeneron Genetics Center.

PAH Biobank Enrolling Centers' Investigators: Russel Hirsch, MD; R. James White, MD, PhD; Marc Simon, MD; David Badesch, MD; Erika Rosenzweig, MD; Charles Burger, MD; Murali Chakinala, MD; Thenappan Thenappan, MD; Greg Elliott, MD; Robert Simms, MD; Harrison Farber, MD; Robert Frantz, MD; Jean Elwing, MD; Nicholas Hill, MD; Dunbar Ivy, MD; James Klinger, MD; StevenNathan, MD; Ronald Oudiz, MD; Ivan Robbins, MD; Robert Schilz, DO, PhD; Terry Fortin, MD; Jeffrey Wilt, MD; Delphine Yung, MD; Eric Austin, MD; Ferhaan Ahmad, MD, PhD; Nitin Bhatt, MD; Tim Lahm, MD; Adaani Frost, MD; Zeenat Safdar, MD; Zia Rehman, MD; Robert Walter, MD; Fernando Torres, MD; Sahil Bakshi, DO; Stephen Archer, MD; Rahul Argula, MD; Christopher Barne

Sources of Funding

The UK National Cohort of Idiopathic and Heritable PAH is supported by the NIHRBR-RD, the British Heart Foundation (BHF) (SP/12/12/29836), the BHF Cambridge Centre of Cardiovascular

Research Excellence, and the UK Medical Research Council (MR/K020919/1), the Dinosaur Trust, BHF Programme grants to RCT (RG/08/006/25302) and NWM (RG/13/4/30107), and the UK NIHR National Institute for Health Research Cambridge Biomedical Research Centre. NWM is a BHF Professor and NIHR Senior Investigator. AL is supported by a BHF Senior Basic Science Research Fellowship (FS/13/48/30453).

Samples and/or data from the National Biological Sample and Data Repository for PAH, funded by an NIH investigator-initiated resources grant (R24 HL105333 to WCN), were used in this study.

Disclosures

NWM is a Director and Co-founder of Morphogen-IX. JW received personal fees from Actelion Pharmaceuticals. GK reports personal fees and non-financial support from Actelion Pharmaceuticals, Bayer, GSK, MSD, Boehringer Ingelheim, Novartis, Chiesi and Vitalaire outside the submitted work. CJR declares fees from Actelion Pharmaceuticals and United Therapeutics. AL received support and fees from GSK and Actelion Pharmaceuticals.

References

1. Voelkel NF, Gomez-Arroyo J, Abbate A, Bogaard HJ, Nicolls MR. Pathobiology of pulmonary arterial hypertension and right ventricular failure [Internet]. European Respiratory Journal. 2012;40:1555–1565. Available from: http://dx.doi.org/10.1183/09031936.00046612

2. Tuder RM, Groves B, Badesch DB, Voelkel NF. Exuberant endothelial cell growth and elements of inflammation are present in plexiform lesions of pulmonary hypertension. *Am J Pathol*. 1994;144:275–285.

3. 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 JSR, Lawrie A, Gräf 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.

4. Pullamsetti SS, Savai R, Seeger W, Goncharova EA. Translational Advances in the Field of Pulmonary Hypertension. From Cancer Biology to New Pulmonary Arterial Hypertension Therapeutics. Targeting Cell Growth and Proliferation Signaling Hubs. *Am J Respir Crit Care Med*. 2017;195:425–437.

5. Thomson J, Machado R, Pauciulo M, Morgan N, Yacoub M, Corris P, McNeil K, Loyd J, Nichols W, Trembath R. Familial and sporadic primary pulmonary hypertension is caused by BMPR2 gene mutations resulting in haploinsufficiency of the bone morphogenetic protein tùype II receptor. *J Heart Lung Transplant*. 2001;20:149.

6. Machado RD, Pauciulo MW, Thomson JR, Lane KB, Morgan NV, Wheeler L, Phillips JA 3rd, Newman J, Williams D, Galiè N, Manes A, McNeil K, Yacoub M, Mikhail G, Rogers P, Corris P, Humbert M, Donnai D, Martensson G, Tranebjaerg L, Loyd JE, Trembath RC, Nichols WC. BMPR2 haploinsufficiency as the inherited molecular mechanism for primary pulmonary hypertension. *Am J Hum Genet*. 2001;68:92–102.

7. Larkin EK, Newman JH, Austin ED, Hemnes AR, Wheeler L, Robbins IM, West JD, Phillips JA 3rd, Hamid R, Loyd JE. Longitudinal analysis casts doubt on the presence of genetic anticipation in heritable pulmonary arterial hypertension. *Am J Respir Crit Care Med*. 2012;186:892–896.

8. Trembath RC. Mutations in the TGF-beta type 1 receptor, ALK1, in combined primary pulmonary hypertension and hereditary haemorrhagic telangiectasia, implies pathway specificity. *J Heart Lung Transplant*. 2001;20:175.

9. Chaouat A. Endoglin germline mutation in a patient with hereditary haemorrhagic telangiectasia and dexfenfluramine associated pulmonary arterial hypertension [Internet]. Thorax. 2004;59:446–448. Available from: http://dx.doi.org/10.1136/thx.2003.11890

10. Shintani M, Yagi H, Nakayama T, Saji T, Matsuoka R. A new nonsense mutation of SMAD8 associated with pulmonary arterial hypertension [Internet]. Journal of Medical Genetics. 2009;46:331–337. Available from: http://dx.doi.org/10.1136/jmg.2008.062703

11. Austin ED, Ma L, LeDuc C, Berman Rosenzweig E, Borczuk A, Phillips JA 3rd, Palomero T, Sumazin P, Kim HR, Talati MH, West J, Loyd JE, Chung WK. Whole exome sequencing to identify a novel gene (caveolin-1) associated with human pulmonary arterial hypertension. *Circ Cardiovasc Genet*. 2012;5:336–343.

12. Ma L, Roman-Campos D, Austin ED, Eyries M, Sampson KS, Soubrier F, Germain M, Trégouët D-A, Borczuk A, Rosenzweig EB, Girerd B, Montani D, Humbert M, Loyd JE, Kass RS, Chung WK. A Novel Channelopathy in Pulmonary Arterial Hypertension [Internet]. New England Journal of Medicine. 2013;369:351–361. Available from: http://dx.doi.org/10.1056/nejmoa1211097

13. Gräf 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 H-A, 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.

14. The NIHR BioResource, on behalf of the 100,000 Genomes Project. Whole-genome sequencing of rare disease patients in a national healthcare system. *Nature* [Internet]. Available from: https://www.biorxiv.org/content/10.1101/507244v1

15. Zhu N, Pauciulo MW, Welch CL, Lutz KA, Coleman AW, Gonzaga-Jauregui C, Wang J, Grimes JM, Martin LJ, He H, PAH Biobank Enrolling Centers' Investigators, Shen Y, Chung WK, Nichols WC. Novel risk genes and mechanisms implicated by exome sequencing of 2572 individuals with pulmonary arterial hypertension. *Genome Med*.

2019;11:69.

16. Zhu N, Welch CL, Wang J, Allen PM, Gonzaga-Jauregui C, Ma L, King AK, Krishnan U, Rosenzweig EB, Ivy DD, Austin ED, Hamid R, Pauciulo MW, Lutz KA, Nichols WC, Reid JG, Overton JD, Baras A, Dewey FE, Shen Y, Chung WK. Rare variants in SOX17 are associated with pulmonary arterial hypertension with congenital heart disease. *Genome Med*. 2018;10:56.

17. Raczy C, Petrovski R, Saunders CT, Chorny I, Kruglyak S, Margulies EH, Chuang H-Y, Källberg M, Kumar SA, Liao A, Little KM, Strömberg MP, Tanner SW. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics*. 2013;29:2041–2043.

18. Greene D, NIHR BioResource, Richardson S, Turro E. A Fast Association Test for Identifying Pathogenic Variants Involved in Rare Diseases. *Am J Hum Genet*. 2017;101:104–114.

19. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL, ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17:405–424.

20. FitzGerald G, Botstein D, Califf R, Collins R, Peters K, Van Bruggen N, Rader D. The future of humans as model organisms [Internet]. Science. 2018;361:552–553. Available from: http://dx.doi.org/10.1126/science.aau7779

21. Terman BI, Carrion ME, Kovacs E, Rasmussen BA, Eddy RL, Shows TB. Identification of a

new endothelial cell growth factor receptor tyrosine kinase. *Oncogene*. 1991;6:1677–1683.

22. Ferrara N, Carver-Moore K, Chen H, Dowd M, Lu L, O'Shea KS, Powell-Braxton L, Hillan KJ, Moore MW. Heterozygous embryonic lethality induced by targeted inactivation of the VEGF gene. *Nature*. 1996;380:439–442.

23. Oladipupo SS, Smith C, Santeford A, Park C, Sene A, Wiley LA, Osei-Owusu P, Hsu J, Zapata N, Liu F, Nakamura R, Lavine KJ, Blumer KJ, Choi K, Apte RS, Ornitz DM. Endothelial cell FGF signaling is required for injury response but not for vascular homeostasis [Internet]. Proceedings of the National Academy of Sciences. 2014;111:13379–13384. Available from: http://dx.doi.org/10.1073/pnas.1324235111

24. Tuder RM, Flook BE, Voelkel NF. Increased gene expression for VEGF and the VEGF receptors KDR/Flk and Flt in lungs exposed to acute or to chronic hypoxia. Modulation of gene expression by nitric oxide. *J Clin Invest*. 1995;95:1798–1807.

25. Cho YJ, Han JY, Lee SG, Jeon BT, Choi WS, Hwang YS, Roh GS, Lee JD. Temporal changes of angiopoietins and Tie2 expression in rat lungs after monocrotaline-induced pulmonary hypertension. *Comp Med*. 2009;59:350–356.

26. Tuder RM, Chacon M, Alger L, Wang J, Taraseviciene-Stewart L, Kasahara Y, Cool CD, Bishop AE, Geraci M, Semenza GL, Yacoub M, Polak JM, Voelkel NF. Expression of angiogenesis-related molecules in plexiform lesions in severe pulmonary hypertension: evidence for a process of disordered angiogenesis [Internet]. The Journal of Pathology. 2001;195:367–374. Available from: http://dx.doi.org/10.1002/path.953.abs

27. Lassus P, Turanlahti M, Heikkilä P, Andersson LC, Nupponen I, Sarnesto A, Andersson S. Pulmonary Vascular Endothelial Growth Factor and Flt-1 in Fetuses, in Acute and Chronic

Lung Disease, and in Persistent Pulmonary Hypertension of the Newborn [Internet]. American Journal of Respiratory and Critical Care Medicine. 2001;164:1981–1987. Available from: http://dx.doi.org/10.1164/ajrccm.164.10.2012036

28. Shehata SM, Mooi WJ, Okazaki T, El-Banna I, Sharma HS, Tibboel D. Enhanced expression of vascular endothelial growth factor in lungs of newborn infants with congenital diaphragmatic hernia and pulmonary hypertension. *Thorax*. 1999;54:427–431.

29. Partovian C, Adnot S, Raffestin B, Louzier V, Levame M, Mavier IM, Lemarchand P, Eddahibi S. Adenovirus-mediated lung vascular endothelial growth factor overexpression protects against hypoxic pulmonary hypertension in rats. *Am J Respir Cell Mol Biol*. 2000;23:762–771.

30. Taraseviciene-Stewart L, Kasahara Y, Alger L, Hirth P, Mc Mahon G, Waltenberger J, Voelkel NF, Tuder RM. Inhibition of the VEGF receptor 2 combined with chronic hypoxia causes cell death-dependent pulmonary endothelial cell proliferation and severe pulmonary hypertension. *FASEB J*. 2001;15:427–438.

31. Itokawa T, Nokihara H, Nishioka Y, Sone S, Iwamoto Y, Yamada Y, Cherrington J, McMahon G, Shibuya M, Kuwano M, Ono M. Antiangiogenic effect by SU5416 is partly attributable to inhibition of Flt-1 receptor signaling. *Mol Cancer Ther*. 2002;1:295–302.

32. Fong TA, Shawver LK, Sun L, Tang C, App H, Powell TJ, Kim YH, Schreck R, Wang X, Risau W, Ullrich A, Hirth KP, McMahon G. SU5416 is a potent and selective inhibitor of the vascular endothelial growth factor receptor (Flk-1/KDR) that inhibits tyrosine kinase catalysis, tumor vascularization, and growth of multiple tumor types. *Cancer Res*. 1999;59:99–106.

33. Kasahara Y, Tuder RM, Taraseviciene-Stewart L, Le Cras TD, Abman S, Hirth PK, Waltenberger J, Voelkel NF. Inhibition of VEGF receptors causes lung cell apoptosis and emphysema. *J Clin Invest*. 2000;106:1311–1319.

34. Garcia AA, Hirte H, Fleming G, Yang D, Tsao-Wei DD, Roman L, Groshen S, Swenson S, Markland F, Gandara D, Scudder S, Morgan R, Chen H, Lenz H-J, Oza AM. Phase II Clinical Trial of Bevacizumab and Low-Dose Metronomic Oral Cyclophosphamide in Recurrent Ovarian Cancer: A Trial of the California, Chicago, and Princess Margaret Hospital Phase II Consortia [Internet]. Journal of Clinical Oncology. 2008;26:76–82. Available from: http://dx.doi.org/10.1200/jco.2007.12.1939

35. Montani D, Bergot E, Günther S, Savale L, Bergeron A, Bourdin A, Bouvaist H, Canuet M, Pison C, Macro M, Poubeau P, Girerd B, Natali D, Guignabert C, Perros F, O'Callaghan DS, Jaïs X, Tubert-Bitter P, Zalcman G, Sitbon O, Simonneau G, Humbert M. Pulmonary Arterial Hypertension in Patients Treated by Dasatinib [Internet]. Circulation. 2012;125:2128–2137. Available from: http://dx.doi.org/10.1161/circulationaha.111.079921

36. Ashraf O, Naddour M, Babar L, Beg M, Malik K. SECOND GENERATION TYROSINE KINASE INHIBITOR BOSUTINIB AS NOVEL CAUSE OF SEVERE PULMONARY HYPERTENSION: UNMASKING THE CULPRIT [Internet]. Chest. 2018;154:1047A. Available from: http://dx.doi.org/10.1016/j.chest.2018.08.945

37. Kamba T, McDonald DM. Mechanisms of adverse effects of anti-VEGF therapy for cancer. *Br J Cancer*. 2007;96:1788–1795.

38. Zheng Y, Chen J, Lu Y. Incidence and risk of hypertension with bevacizumab in non-small-cell lung cancer patients: a meta-analysis of randomized controlled trials [Internet].

Drug Design, Development and Therapy. 2015;4751. Available from: http://dx.doi.org/10.2147/dddt.s87258

39. Steeghs N, Gelderblom H, Roodt JO 't, Christensen O, Rajagopalan P, Hovens M, Putter H, Rabelink TJ, de Koning E. Hypertension and rarefaction during treatment with telatinib, a small molecule angiogenesis inhibitor. *Clin Cancer Res*. 2008;14:3470–3476.

40. van den Meiracker AH, Danser AHJ, Sleijfer S, Kappers MHW. Re: Hypertension as a Biomarker of Efficacy in Patients With Metastatic Renal Cell Carcinoma Treated With Sunitinib [Internet]. JNCI Journal of the National Cancer Institute. 2011;103:1557–1557. Available from: http://dx.doi.org/10.1093/jnci/djr328

41. Levy BI, Ambrosio G, Pries AR, Struijker-Boudier HAJ. Microcirculation in Hypertension [Internet]. Circulation. 2001;104:735–740. Available from: http://dx.doi.org/10.1161/hc3101.091158

42. Kappers MHW, van Esch JHM, Sluiter W, Sleijfer S, Jan Danser AH, van den Meiracker AH. Hypertension Induced by the Tyrosine Kinase Inhibitor Sunitinib Is Associated With Increased Circulating Endothelin-1 Levels [Internet]. Hypertension. 2010;56:675–681. Available from: http://dx.doi.org/10.1161/hypertensionaha.109.149690

43. Bleyl S, Nelson L, Odelberg SJ, Ruttenberg HD, Otterud B, Leppert M, Ward K. A gene for familial total anomalous pulmonary venous return maps to chromosome 4p13-q12. *Am J Hum Genet*. 1995;56:408–415.

44. Reuter MS, Jobling R, Chaturvedi RR, Manshaei R, Costain G, Heung T, Curtis M, Hosseini SM, Liston E, Lowther C, Oechslin E, Sticht H, Thiruvahindrapuram B, van Mil S, Wald RM, Walker S, Marshall CR, Silversides CK, Scherer SW, Kim RH, Bassett AS.

Haploinsufficiency of vascular endothelial growth factor related signaling genes is associated with tetralogy of Fallot. *Genet Med*. 2019;21:1001–1007.

45. Sakao S, Miyauchi H, Voelkel NF, Sugiura T, Tanabe N, Kobayashi Y, Tatsumi K. Increased Right Ventricular Fatty Acid Accumulation in Chronic Thromboembolic Pulmonary Hypertension. *Ann Am Thorac Soc*. 2015;12:1465–1472.

46. Sutendra G, Dromparis P, Bonnet S, Haromy A, McMurtry MS, Chris Bleackley R, Michelakis ED. Pyruvate dehydrogenase inhibition by the inflammatory cytokine TNFα contributes to the pathogenesis of pulmonary arterial hypertension [Internet]. Journal of Molecular Medicine. 2011;89:771–783. Available from: http://dx.doi.org/10.1007/s00109-011-0762-2

47. Salceda S, Caro J. Hypoxia-inducible Factor 1α (HIF-1α) Protein Is Rapidly Degraded by the Ubiquitin-Proteasome System under Normoxic Conditions [Internet]. Journal of Biological Chemistry. 1997;272:22642–22647. Available from: http://dx.doi.org/10.1074/jbc.272.36.22642

48. Wellen KE, Hatzivassiliou G, Sachdeva UM, Bui TV, Cross JR, Thompson CB. ATP-citrate lyase links cellular metabolism to histone acetylation. *Science*. 2009;324:1076–1080.

49. Johnson D. Cell Death Signaling in Cancer Biology and Treatment. Springer Science & Business Media; 2012.

50. Fessel JP, Hamid R, Wittmann BM, Robinson LJ, Blackwell T, Tada Y, Tanabe N, Tatsumi K, Hemnes AR, West JD. Metabolomic analysis of bone morphogenetic protein receptor type 2 mutations in human pulmonary endothelium reveals widespread metabolic reprogramming. *Pulm Circ*. 2012;2:201–213.

51. Michelakis ED, Gurtu V, Webster L, Barnes G, Watson G, Howard L, Cupitt J, Paterson I, Thompson RB, Chow K, O'Regan DP, Zhao L, Wharton J, Kiely DG, Kinnaird A, Boukouris AE, White C, Nagendran J, Freed DH, Wort SJ, Gibbs JSR, Wilkins MR. Inhibition of pyruvate dehydrogenase kinase improves pulmonary arterial hypertension in genetically susceptible patients. *Sci Transl Med* [Internet]. 2017;9. Available from: http://dx.doi.org/10.1126/scitranslmed.aao4583

52. Test VJ, Farber HW, McGoon MD, Parsons L, Channick RN. Pulmonary Arterial Hypertension in the Elderly: Baseline Characteristics and Evaluation of Therapeutics. An Examination of the Reveal Registry [Internet]. B27. FROM ALPHA TO OMEGA: ASSESSMENT AND OUTCOMES IN PULMONARY HYPERTENSION. 2009;Available from: http://dx.doi.org/10.1164/ajrccm-conference.2009.179.1_meetingabstracts.a2649

53. Hadinnapola C, Bleda M, Haimel M, Screaton N, Swift A, Dorfmüller P, Preston SD, Southwood M, Hernandez-Sanchez J, Martin J, Treacy C, Yates K, Bogaard H, Church C, Coghlan G, Condliffe R, Corris PA, Gibbs S, Girerd B, Holden S, Humbert M, Kiely DG, Lawrie A, Machado R, MacKenzie Ross R, Moledina S, Montani D, Newnham M, Peacock A, Pepke-Zaba J, Rayner-Matthews P, Shamardina O, Soubrier F, Southgate L, Suntharalingam J, Toshner M, Trembath R, Vonk Noordegraaf A, Wilkins MR, Wort SJ, Wharton J, NIHR BioResource–Rare Diseases Consortium; UK National Cohort Study of Idiopathic and Heritable PAH, Gräf S, Morrell NW. Phenotypic Characterization of Mutation Carriers in a Large Cohort of Patients Diagnosed Clinically With Pulmonary Arterial Hypertension. *Circulation*. 2017;136:2022–2033.

54. Trip P, Girerd B, Bogaard H-J, de Man FS, Boonstra A, Garcia G, Humbert M, Montani D, Vonk-Noordegraaf A. Diffusion capacity and BMPR2 mutations in pulmonary arterial

hypertension. *Eur Respir J*. 2014;43:1195–1198.

55. Nathan SD, Barbera JA, Gaine SP, Harari S, Martinez FJ, Olschewski H, Olsson KM, Peacock AJ, Pepke-Zaba J, Provencher S, Weissmann N, Seeger W. Pulmonary hypertension in chronic lung disease and hypoxia. *Eur Respir J* [Internet]. 2019;53. Available from: http://dx.doi.org/10.1183/13993003.01914-2018

56. Montani D, Dorfmuller P, Maitre S, Jaïs X, Sitbon O, Simonneau G, Humbert M. [Pulmonary veno-occlusive disease and pulmonary capillary hemangiomatosis]. *Presse Med*. 2010;39:134–143.

57. Sansores RH, Pare PD, Abboud RT. Acute effect of cigarette smoking on the carbon monoxide diffusing capacity of the lung. *Am Rev Respir Dis*. 1992;146:951–958.

58. López-Campos JL, Soler-Cataluña JJ, Miravitlles M. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2019 Report: Future Challenges [Internet]. Archivos de Bronconeumología. 2019;Available from: http://dx.doi.org/10.1016/j.arbres.2019.06.001

Figure legends

Figure 1. Design of the genetic association study. A, Overview of the analytical approach. Using deep phenotyping, data tags were assigned to patients who shared phenotypic features (see Figure S1 for more details). Rare sequence variants, called from whole-genome sequencing data, were filtered and explained cases were labelled. BeviMed was applied to a set of unrelated individuals, to test the posterior probability of gene-tag associations. **B**, Consort diagram summarising the size of the study cohort. **C**, Schematic representation of the definition of cases, exemplified by the KCO lower tertile tag. Cases were defined as individuals carrying a particular tag, whereas patients with missing information or those without a tag were removed from the gene-tag association testing. Individuals from non-PAH domains served as controls. KCO - transfer coefficient of carbon monoxide, MAF - minor allele frequency.

Figure 2. Genetic association study results revealing established and novel genotype-phenotype links. A, Figure showing phenotype tags on the x-axis and their corresponding posterior probability on the y-axis, as calculated by BeviMed. This measure predicts associations between tags and rare, predicted deleterious variants within a given gene. The definitions of the tags are listed in Table 1. Shape and colour of points indicate the mode of inheritance and consequence type/impact of variants driving the association. Box-and-whisker plots showing the distribution of transfer coefficient of carbon monoxide (**B**) and age at diagnosis (**C**) stratified by genotype across the PAH domain. The two-tailed Wilcoxon signed-rank test was used to determine differences in the medians of the distributions, which are indicated by the bars at the top of the figures providing the respective p-values. bial. - biallelic, het. - heterozygous, pt. - protein-truncating, mis. - missense.

Figure 3. Summary of single nucleotide variants (SNVs), small insertions and deletions (indels) and large deletions identified in the two novel candidate PAH disease risk genes. Only predicted deleterious variants in *KDR* (**A**) and *IDH3G* (**B**, **C**) are shown (MAF<1/10,000 and CADD>=15). SNVs and indels are represented by coloured lollipops on top of the protein sequence. Lollipop colour indicates the consequence type and size represents the variant frequency within a cohort. The domain annotations were retrieved from Uniprot (accession numbers P35968 [*KDR* (**A**)] and P51553 [*IDH3G* (**D**)]). PTVs are labelled with the respective HGVS notation. Splice variants are marked by dark grey arrows. The large deletion identified in *IDH3G* (**C**) is depicted in light blue; the respective gene locus is highlighted in red. The number of variants by predicted consequence type and cohort is provided in the tables.

Figure 4. Pulmonary computerised tomography (CT) scans of patients carrying protein-truncating *KDR* mutations. A, Axial image of pulmonary CT angiogram at the level of the right ventricle (RV) moderator band, showing flattening of intraventricular septum, leftwards bowing of the interatrial septum and the enlargement of the right atrium (RA) and RV, indicative of RV strain; bilateral pleural effusion, larger on the right side. **B**, Axial image of a pulmonary CT angiogram demonstrating enlarged pulmonary artery and mild central lung ground glass opacity (GGO). **C**, Axial high-resolution CT slice of the chest in the lung window showing a trace of non-specific GGO with a central distribution. **D**, Coronal image showing the trace of central GGO and enlarged central pulmonary arteries. Axial high-resolution CT slice of the chest in the lung window showing apical subpleural fibrosis (**E**), and very minor subpleural fibrosis at the lung bases (**F**). Axial high-resolution CT slice of the chest in the lung window showing subpleural GGO

at apical level (**G**), and mild GGO at mid-thoracic level (**H**). Patients: E001392 (**A**, **B**), E003448 (**C**, **D**), W000229 (**E**, **F**), W000274 (**G** and **H**).

Supplemental figure legends

Figure S1. Summary of missing data. A, The fraction of missing data for KCO, in comparison to diagnosis, age at diagnosis and lung function tests. **B**, The missingness pattern in KCO, in relation to diagnosis, age at diagnosis and lung function tests. KCO - transfer coefficient of carbon monoxide. FEV_1 - forced expiratory volume in 1 second, FVC - forced vital capacity, TLC - total lung capacity.

Figure S2. Flowchart describing the definition of diagnostic and phenotypic tags. A detailed description is provided in the supplemental material. The definition of tags is listed in Table 1.

Figure S3. Characterisation and survival analysis of the cohort. Distribution of transfer coefficient of carbon monoxide (KCO), coloured by KCO tertiles (**A**), and coloured by KCO below and above the 50% predicted threshold (**B**). **C**, Distribution of age tertiles. **D**, Kaplan-Meier survival curves for KCO tertiles. **E**, Kaplan-Meier survival curves for KCO below and above 50% predicted threshold. **F**, Kaplan-Meier survival curves for age tertiles.

Figure S4. Summary of large deletions identified in previously established disease genes. Deletions are indicated by light blue boxes. The protein-coding genes, annotated in the displayed region by Ensembl (GRCh37, version 75), are depicted in the bottom panels. The affected genomic regions, with the disease gene locus highlighted in red and the magnified view focusing on the gene loci, are shown for *BMPR2* (**A**, **B**), *GDF2* (**C**, **D**) and *TBX4* (**E**, **F**).

Table legends

Table 1. Definitions of labels and the number of unrelated cases and controls for genetic association analysis with BeviMed. mPAP - mean pulmonary artery pressure, PH - pulmonary hypertension, PAH - pulmonary arterial hypertension, I/HPAH - Idiopathic/Hereditary Pulmonary Arterial Hypertension, PVOD - Pulmonary veno-occlusive disease, PCH - Pulmonary capillary haemangiomatosis, APAH - Associated Pulmonary Arterial Hypertension, CHD - Congenital Heart Disease, CTD - Connective Heart Disease, PPH, LHD - Left Heart Disease, LD - Lung Disease, CTEPH - Chronic Thromboembolic Pulmonary Hypertension, KCO - transfer coefficient of carbon monoxide.

Table 2. BeviMed analysis results. Posterior probabilities and Bayes factors of gene-tag associations. The "High" category, comprise only variants of “high” impact, including PTVs and large deletions; the Moderate category contains variants of impact “moderate”, including missense variants or consequence “non_coding_transcript_exon_variant”; the combined category Moderate and High, include both respective consequence types.

Table 3. Gene changes for IPAH patients harbouring protein-truncating variants (PTV) in the KDR gene and PTV and missense variants in the *IDH3G* gene. *KDR* - Kinase insert domain receptor, *IDH3G* - Isocitrate dehydrogenase (NAD(+)) 3 non-catalytic subunit gamma, WHO FC - World Health Organisation functional class, 6MWD - 6-minute walk distance, SpO_2 - arterial oxygen saturation, mRAP - mean right atrial pressure, mPAP - mean pulmonary artery pressure, mPAWP - mean pulmonary artery wedge pressure, CO - cardiac output, PVR - pulmonary vascular resistance, FEV_1 - forced expiratory volume in 1 sec, FVC - forced vital capacity, KCO - transfer factor coefficient for carbon monoxide. None of the *KDR* variants has been previously

reported in gnomAD, ExAC or internal controls. For *KDR* HGVSc notations are based on transcript sequence ENST00000263923.4. HGVSp notations are based on amino acid sequence ENSP00000263923.4. None of the patients harbouring PTV in *KDR* had capillary hemangioma, *DLCO% predicted; For *IDH3G* HGVSc notations are based on transcript sequence ENST00000217901.5, HGVSp notations are based on amino acid sequence ENSP00000217901.5. Protein truncating variants were defined as stop gained, splice acceptor variants or frameshift variants.

Table 4. Clinical characteristics of IPAH patients harbouring protein truncating variants in the *KDR* gene. *KDR* - Kinase insert domain receptor, IPAH - idiopathic pulmonary arterial hypertension, BMI - Body Mass Index, WHO FC - World Health Organisation functional class, 6MWD - six-minute walk distance, SpO_2 - arterial oxygen saturation, mRAP - mean right atrial pressure, mPAP - mean pulmonary artery pressure, PAWP - pulmonary artery wedge pressure, CO - cardiac output, PVR - pulmonary vascular resistance, NO - nitric oxide, FEV_1 - forced expiratory volume in 1 second, FVC - forced vital capacity, KCO - transfer factor coefficient for carbon monoxide, COPD - chronic obstructive pulmonary disease, CAD - coronary artery disease, HTN - systemic hypertension, CKD - chronic kidney disease, Hb - haemoglobin, WBC - white blood cells, TSH - thyroid-stimulating hormone. Comorbidities are reported as the number and percentage of cases possessing a disease entity. None of the patients had a history of pulmonary embolism or asthma. Three of the *KDR* missense variants co-occurred with predicted deleterious variants in established PAH risk genes (*BMPR2* and *AQP1*)

Supplemental table legends

Table S1. NIHR BioResource - Rare Diseases domain definitions.

Table S2. Summary of electronic clinical report forms (CRFs) constructed to capture phenotypic information

Table S3. Reporting proforma for CT scan revision. CTPA - Computerised Tomography Pulmonary Angiogram, HRCT - High-Resolution Computerised Tomography, GGO - ground glass opacities

Table S4. Clinical characterisation of the study population. BMI - body mass index, WHO FC - World Health Organisation Functional Class, 6MWD - 6-minute walk distance, mRAP - mean right atrial pressure, mPAP - mean pulmonary artery pressure, CO - cardiac output, FEV_1 - forced expiratory capacity in 1 second, FVC - forced vital capacity, KCO - transfer factor coefficient for carbon monoxide, Hb - haemoglobin, RDW - red cell distribution width, WBC - white blood cell count, NTproBNP - N-terminal pro b-type natriuretic peptide, BNP - B-type natriuretic peptide, CRP - C-Reactive Protein Protein, HTN - hypertension, DM - diabetes mellitus, CAD - coronary artery disease, CVA - cerebrovascular accident, COPD - chronic obstructive pulmonary disease, CCB - calcium channel blocker, ERA - endothelin receptor antagonists, PA - prostacyclin analogues, PED5 - phosphodiesterase type 5, sGC - soluble guanylate cyclase; Entire cohort (n=1122) was composed of IPAH (n=972), HPAH (n=73), PVOD/PCH (n=20), PH associated with left heart disease (n=7), PH associated with lung disease (n=8), chronic thromboembolic pulmonary hypertension (n=6), multifactorial PH (n=6), hereditary hemorrhagic telangiectasia (n=1)

Table S5. Clinical differences between patients with present and missing transfer coefficient results. BMI - body mass index, WHO FC - World Health Organisation Functional Class, 6MWD - 6-minute walk distance, mRAP - mean right atrial pressure, mPAP - mean pulmonary artery pressure, PAWP - Pulmonary Artery Wedge Pressure, CI - cardiac index, PVR - pulmonary vascular resistance, SvO_2 [%] - mixed venous saturation, HRCT - High-Resolution Computerised Tomography, FEV_1 - forced expiratory capacity in 1 second, FVC - forced vital capacity, COPD - chronic obstructive pulmonary disease, OSA - obstructive sleep apnoea

Table S6. Clinical characteristics of unrelated individuals used in gene-tag association analysis by KCO threshold. BMI - body mass index, WHO FC - World Health Organisation Functional Class, 6MWD - 6-minute walk distance, SpO_2 - peripheral capillary oxygen saturation, mRAP - mean right atrial pressure, mPAP - mean pulmonary artery pressure, PAWP - Pulmonary Artery Wedge Pressure, CO - cardiac output, SvO_2 - Mixed venous oxygen saturation, FEV_1 - forced expiratory capacity in 1 second, FVC - forced vital capacity, TLC - Total Lung Capacity, KCO - transfer coefficient of carbon monoxide, HRCT - High-Resolution Computerised Tomography, NTproBNP - N-terminal pro B-Type Natriuretic Peptide, BNP - B-Type Natriuretic Peptide, CRP - C-Reactive Protein Protein, Hb - haemoglobin, WBC - white blood cell count, COPD - chronic obstructive pulmonary disease, OSA - obstructive sleep apnoea, CAD - coronary artery disease, CVA - cerebrovascular accident, PAD - peripheral artery disease, HTN - hypertension, DM - diabetes mellitus; none of the patients had systemic lupus erythematosus, systemic sclerosis, undifferentiated connective tissue disease or ankylosing spondylitis

Table S7. Clinical characteristics of unrelated individuals used in gene-tag association analysis by KCO tertiles. BMI - body mass index, WHO FC - World Health Organisation Functional Class,

6MWD - 6-minute walk distance, SpO_2 - peripheral capillary oxygen saturation, mRAP - mean right atrial pressure, mPAP - mean pulmonary artery pressure, PAWP - Pulmonary Artery Wedge Pressure, CO - cardiac output, SvO_2 - mixed venous oxygen saturation, NO - nitric oxide, FEV_1 - forced expiratory capacity in 1 second, FVC - forced vital capacity, TLC - Total Lung Capacity, KCO - transfer coefficient of carbon monoxide, HRCT - High-Resolution Computerised Tomography, NTproBNP - N-terminal pro b-type natriuretic peptide, BNP - B-type natriuretic peptide, CRP - C-Reactive Protein Protein, Hb - haemoglobin, WBC - white blood cell count, COPD - chronic obstructive pulmonary disease, OSA - obstructive sleep apnoea, CAD - coronary artery disease, CVA - cerebrovascular accident, PAD - peripheral artery disease, HTN - hypertension, DM - diabetes mellitus; none of the patients had systemic lupus erythematosus, systemic sclerosis, undifferentiated connective tissue disease or ankylosing spondylitis

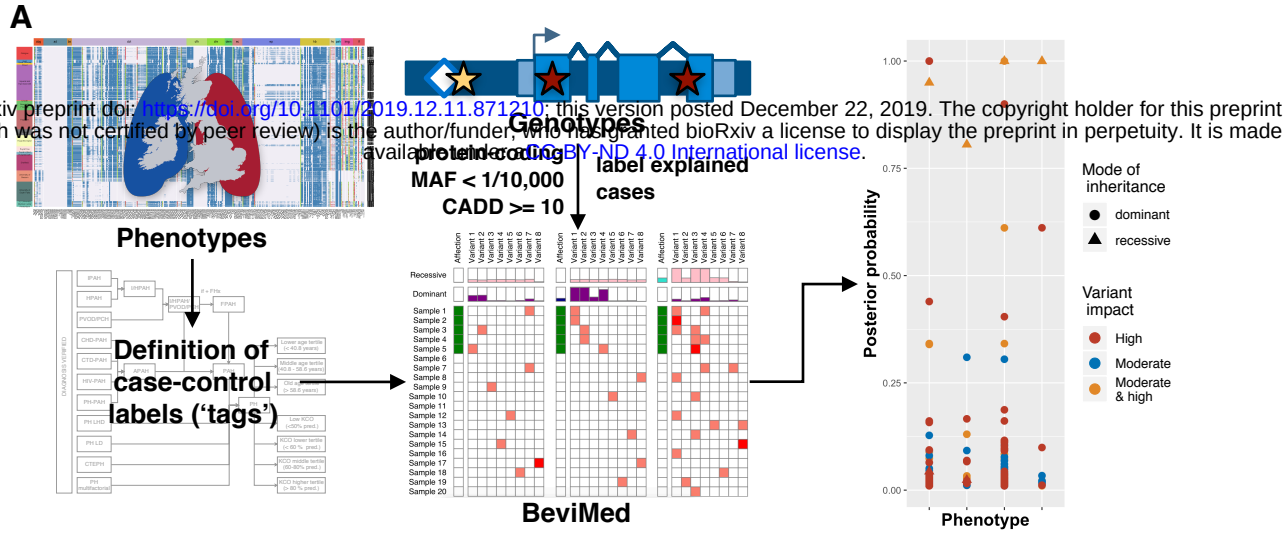
Table S8. Result of Cox regression analysis relating overall survival to selected variables at baseline. CI - Confidence interval, 6MWD - 6-minute walking distance, mPAP - mean pulmonary arterial pressure, mRAP - mean right atrial pressure, PVR - pulmonary vascular resistance, WU - Wood units, KCO - transfer coefficient of carbon monoxide, CAD - coronary artery disease, COPD - chronic obstructive pulmonary disease, HTN - systemic hypertension, HRCT - High-Resolution Computerised Tomography

Table S9. Clinical characteristics of unrelated individuals used in gene-tag association by age tertiles. BMI - body mass index, WHO FC - World Health Organisation Functional Class, 6MWD - 6-minute walk distance, SpO_2 - peripheral capillary oxygen saturation, mRAP - mean right atrial pressure, mPAP - mean pulmonary artery pressure, PAWP - Pulmonary Artery Wedge Pressure, CO - cardiac output, SvO_2 - Mixed venous oxygen saturation, NO - nitric oxide, FEV_1 - forced expiratory capacity in 1 second, FVC - forced vital capacity, TLC - Total Lung Capacity, KCO -

transfer factor coefficient for carbon monoxide, NTproBNP - N-terminal pro b-type natriuretic peptide, BNP - B-type natriuretic peptide, CRP - C-Reactive Protein Protein, Hb - haemoglobin, WBC - white blood cell count, COPD - chronic obstructive pulmonary disease, OSA - obstructive sleep apnoea, CAD - coronary artery disease, CVA - cerebrovascular accident, PAD - peripheral artery disease, HTN - hypertension, DM - diabetes mellitus; none of the patients had undifferentiated connective tissue disease, incident cases were defined as those diagnosed within 6 months from study commencement.

Table S10. Summary of imaging analysis. IPAH - idiopathic pulmonary arterial hypertension, HPAH - hereditary pulmonary arterial hypertension, PVOD - pulmonary veno-occlusive disease, PCH - Pulmonary capillary haemangiomatosis, GGO - ground glass opacities, BA - bronchial artery, C - central, U - upper, Z - zonal, D - diffuse; Intra-rater reliability: GGO centrilobular pattern severity weighted Cohen's Kappa=0.679, p-value <0.001; GGO distribution unweighted Cohen's Kappa = 1, p-value 0.046; Severity of GGO non-specific pattern - no positive findings; Pulmonary arteriovenous malformations - no positive findings; largest BA size - no positive findings; Mediastinal venous collaterals: unweighted Cohen's Kappa = 1, p-value <0.001; Intralobular septal thickening weighted Cohen's Kappa = 1, p-value <0.001; Mediastinal lymphadenopathy unweighted Cohen's Kappa=0.83, p-value <0.001; Mediastinal lymphadenopathy size [mm] intraclass correlation coefficient (ICC) 0.717, p-value 0.088; Emphysema - not enough positive findings, Bronchial wall thickening - not enough positive findings, Fibrosis - no positive findings; Pleural effusion weighted Cohen's Kappa 0.826, p-value <0.001; Air trapping weighted Cohen's Kappa 0.845, p-value <0.001; Subpleural scarring - not enough positive findings.

Table S11. Clinical characteristics of IPAH patients who harbour protein-truncating variants in *BMPR2*, *EIF2AK4*, *KDR* and *IDH3G*. BMI - body mass index, WHO FC - World Health Organisation functional class, 6MWD - 6-minute walk distance, SpO_2 - arterial oxygen saturation, mRAP - mean right atrial pressure, mPAP - mean pulmonary artery pressure, mPAWP - mean pulmonary artery wedge pressure, CO - cardiac output, PVR - pulmonary vascular resistance, NO - nitric oxide challenge, FEV_1 - forced expiratory volume in 1 second, FVC - forced vital capacity, KCO - transfer factor coefficient for carbon monoxide, COPD - chronic obstructive pulmonary disease, OSA - obstructive sleep apnea, CAD - coronary artery disease, HTN - systemic hypertension, CKD - chronic kidney disease, Hb - haemoglobin, WBC - white blood cells, TSH - thyroid-stimulating hormone. Comorbidities are reported as the number and percentage of cases possessing a disease entity.



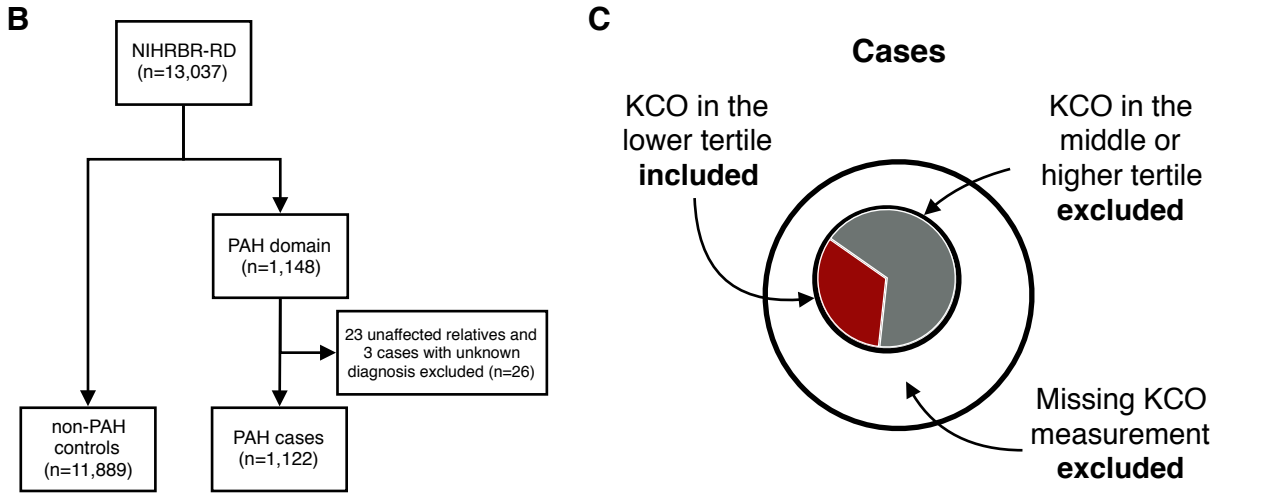


Figure 1. Design of the genetic association study. A, Overview of the analytical approach. Using deep phenotyping, data tags were assigned to patients who shared phenotypic features (see Figure S1 for more details). Rare sequence variants, called from whole-genome sequencing data, were filtered and explained cases were labelled. BeviMed was applied to a set of unrelated individuals, to test the posterior probability of gene-tag associations. **B**, Consort diagram summarising the size of the study cohort. **C**, Schematic representation of the definition of cases, exemplified by the KCO lower tertile tag. Cases were defined as individuals carrying a particular tag, whereas patients with missing information or those without a tag were removed from the gene-tag association testing. Individuals from non-PAH domains served as controls. KCO - transfer coefficient of carbon monoxide, MAF - minor allele frequency.

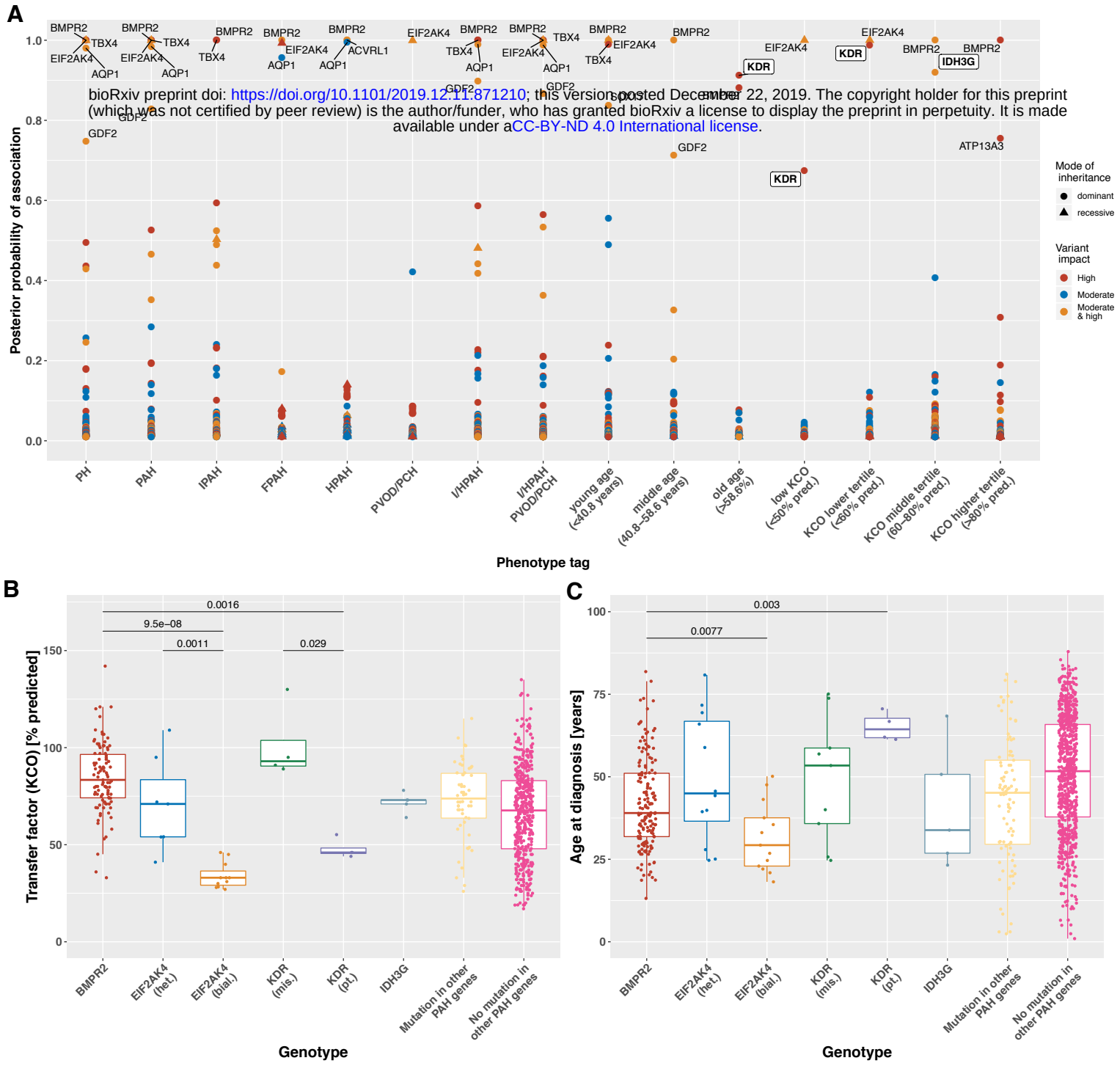
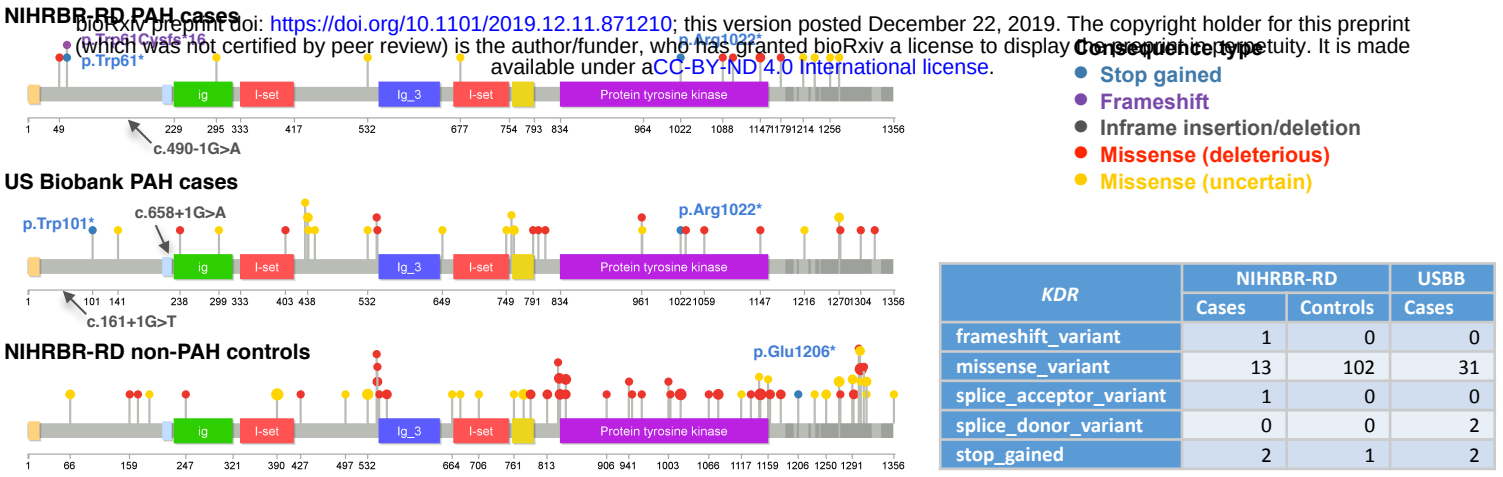


Figure 2. Genetic association study results revealing established and novel genotype-phenotype links. A, Figure showing phenotype tags on the x-axis and their corresponding posterior probability on the y-axis, as calculated by BeviMed. This measure predicts associations between tags and rare, predicted deleterious variants within a given gene. The definitions of the tags are listed in Table 1. Shape and colour of points indicate the mode of inheritance and consequence type/impact of variants driving the association. Box-and-whisker plots showing the distribution of transfer coefficient of carbon monoxide (**B**) and age at diagnosis (**C**) stratified by genotype across the PAH domain. The two-tailed Wilcoxon signed-rank test was used to determine differences in the medians of the distributions, which are indicated by the bars at the top of the figures providing the respective p-values. bial. - biallelic, het. - heterozygous, pt. - protein-truncating, mis. - missense.

A

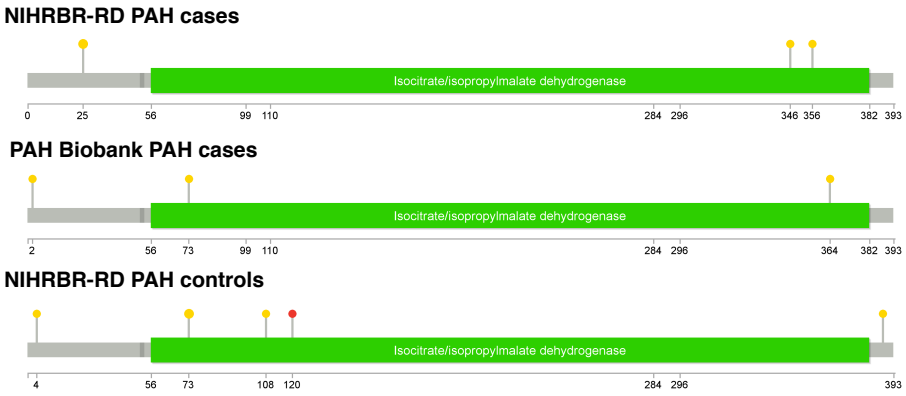
KDR SNVs and indels



KDR	NIHRBR-RD		USBB
	Cases	Controls	Cases
frameshift_variant	1	0	0
missense_variant	13	102	31
splice_acceptor_variant	1	0	0
splice_donor_variant	0	0	2
stop_gained	2	1	2

B

IDH3G SNVs and indels



IDH3G	NIHRBR-RD		USBB
	Cases	Controls	Cases
large_deletion	1	2	0
missense_variant	4	6	3

C

IDH3G large deletion

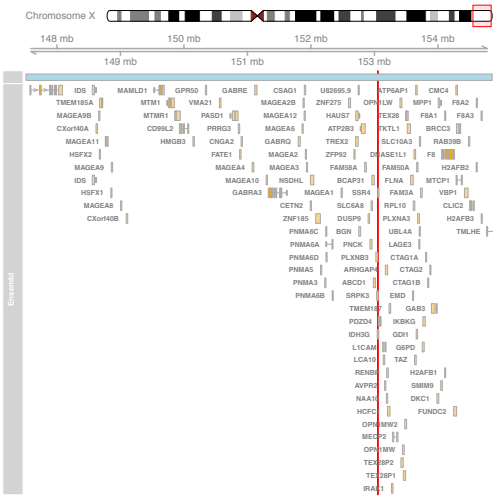


Figure 3. Summary of single nucleotide variants (SNVs), small insertions and deletions (indels) and large deletions identified in the two novel candidate PAH disease risk genes. Only predicted deleterious variants in *KDR* (**A**) and *IDH3G* (**B**, **C**) are shown (MAF<1/10,000 and CADD>=15). SNVs and indels are represented by coloured lollipops on top of the protein sequence. Lollipop colour indicates the consequence type and size represents the variant frequency within a cohort. The domain annotations were retrieved from Uniprot (accession numbers P35968 [*KDR* (**A**)] and P51553 [*IDH3G* (**D**)]). PTVs are labelled with the respective HGVS notation. Splice variants are marked by dark grey arrows. The large deletion identified in *IDH3G* (**C**) is depicted in light blue; the respective gene locus is highlighted in red. The number of variants by predicted consequence type and cohort is provided in the tables.

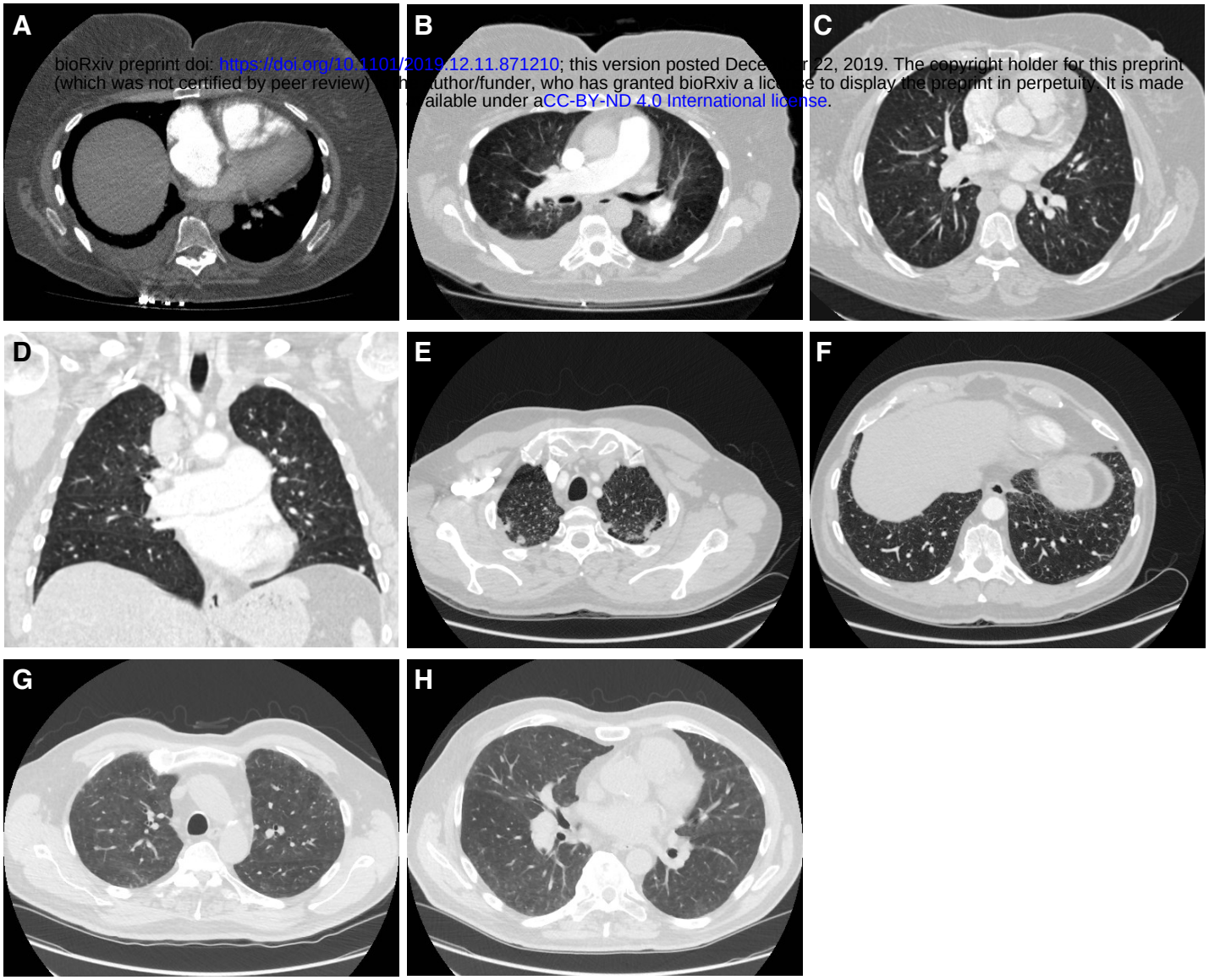


Figure 4. Pulmonary computerised tomography (CT) scans of patients carrying protein-truncating *KDR* mutations. A, Axial image of pulmonary CT angiogram at the level of the right ventricle (RV) moderator band, showing flattening of intraventricular septum, leftwards bowing of the interatrial septum and the enlargement of the right atrium (RA) and RV, indicative of RV strain; bilateral pleural effusion, larger on the right side. **B**, Axial image of a pulmonary CT angiogram demonstrating enlarged pulmonary artery and mild central lung ground glass opacity (GGO). **C**, Axial high-resolution CT slice of the chest in the lung window showing a trace of non-specific GGO with a central distribution. **D**, Coronal image showing the trace of central GGO and enlarged central pulmonary arteries. Axial high-resolution CT slice of the chest in the lung window showing apical subpleural fibrosis (**E**), and very minor subpleural fibrosis at the lung bases (**F**). Axial high-resolution CT slice of the chest in the lung window showing subpleural GGO at apical level (**G**), and mild GGO at mid-thoracic level (**H**). Patients: E001392 (**A**, **B**), E003448 (**C**, **D**), W000229 (**E**, **F**), W000274 (**G** and **H**).

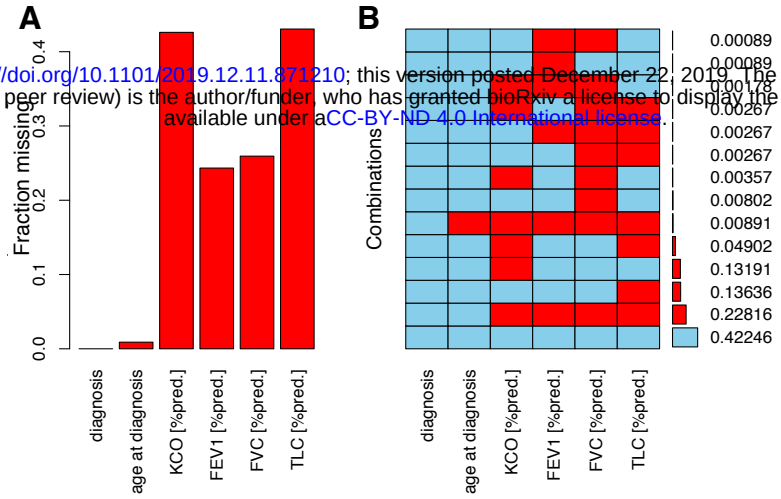


Figure S1. Summary of missing data. A, The fraction of missing data for KCO, in comparison to diagnosis, age at diagnosis and lung function tests. **B**, The missingness pattern in KCO, in relation to diagnosis, age at diagnosis and lung function tests. KCO - transfer coefficient of carbon monoxide. FEV_1 - forced expiratory volume in 1 second, FVC - forced vital capacity, TLC - total lung capacity.

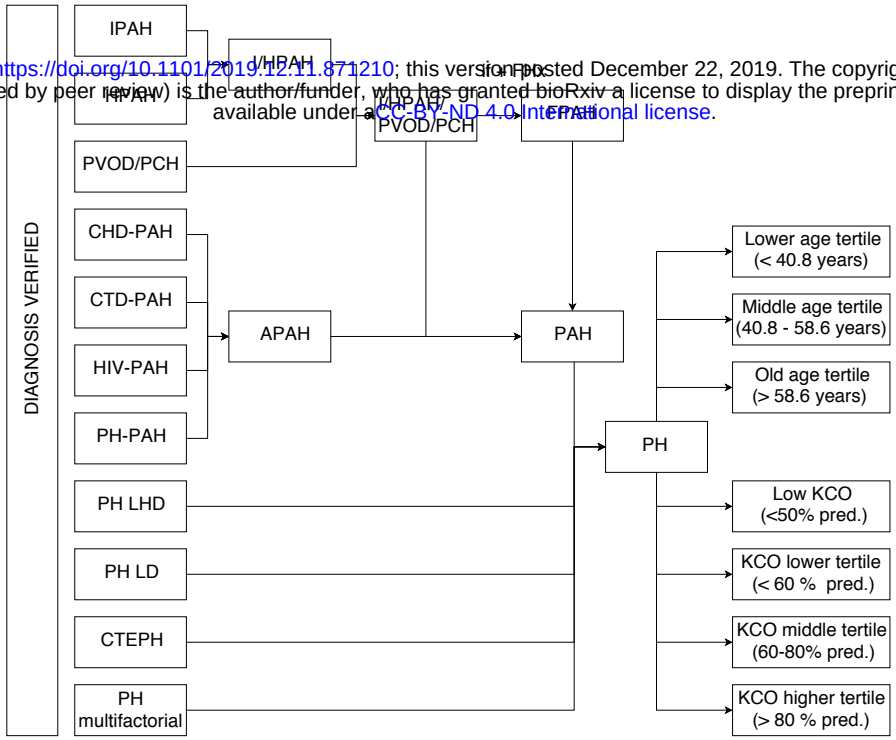


Figure S2. Flowchart describing the definition of diagnostic and phenotypic tags. A detailed description is provided in the supplemental material. The definition of tags is listed in Table 1.

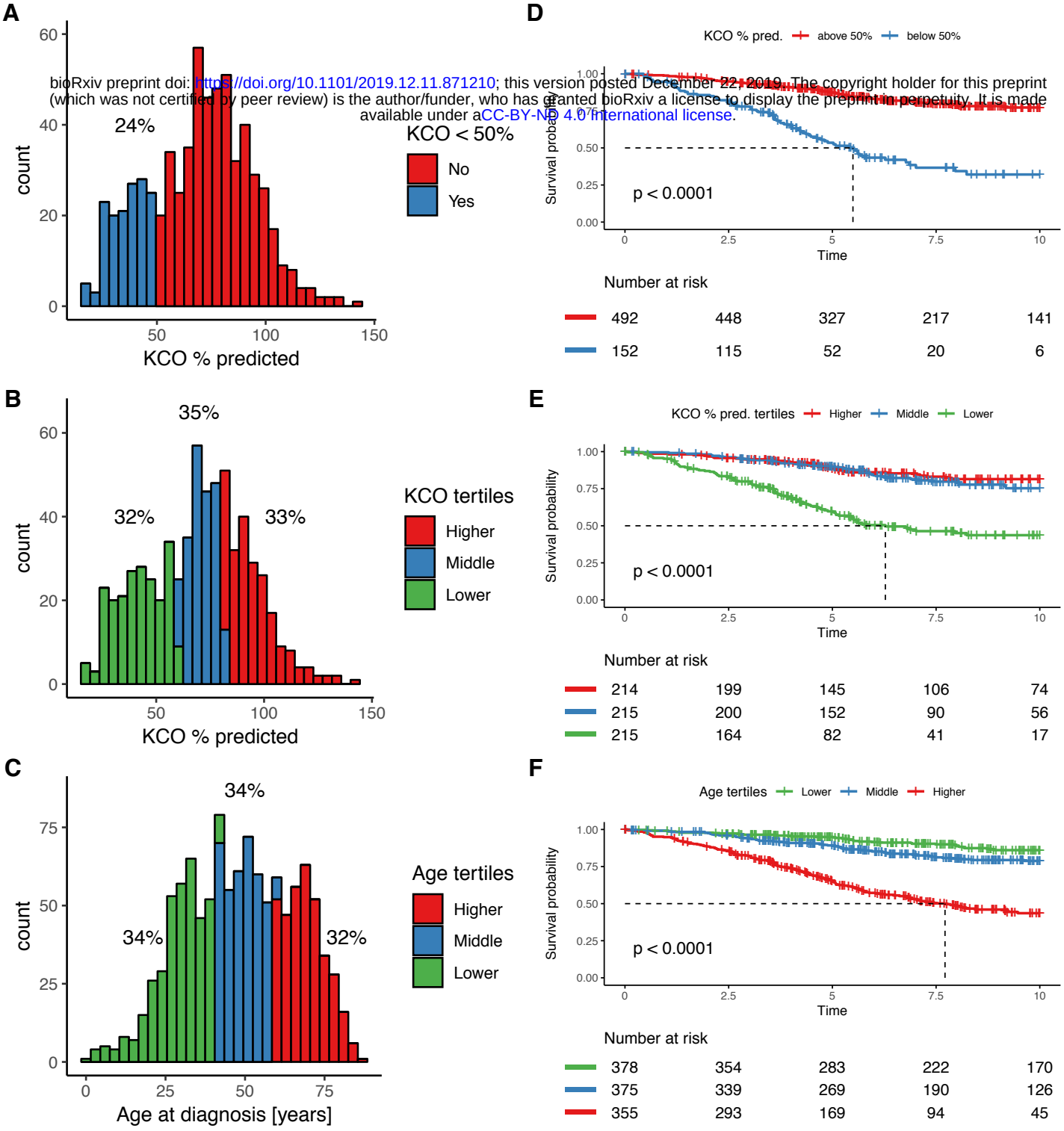


Figure S3. Characterisation and survival analysis of the cohort. Distribution of transfer coefficient of carbon monoxide (KCO), coloured by KCO tertiles (**A**), and coloured by KCO below and above the 50% predicted threshold (**B**). **C**, Distribution of age tertiles. **D**, Kaplan-Meier survival curves for KCO tertiles. **E**, Kaplan-Meier survival curves for KCO below and above 50% predicted threshold. **F**, Kaplan-Meier survival curves for age tertiles.

Figure S4. Summary of large deletions identified in previously established disease genes. Deletions are indicated by light blue boxes. The protein-coding genes, annotated in the displayed region by Ensembl (GRCh37, version 75), are depicted in the bottom panels. The affected genomic regions, with the disease gene locus highlighted in red and the magnified view focusing on the gene loci, are shown for *BMPR2* (**A**, **B**), *GDF2* (**C**, **D**) and *TBX4* (**E**, **F**).

Table 1 Definitions of tags and number of unrelated cases and controls genetic association analysis with Bernoulli mixed model

Abbreviations: mPAP - mean pulmonary artery pressure, PH - pulmonary hypertension, PAH - pulmonary arterial hypertension, I/HPAH - Idiopathic/Hereditary Pulmonary Arterial Hypertension, PVOD - Pulmonary veno-occlusive disease, PCH - Pulmonary capillary haemangiomatosis, APAH - Associated Pulmonary Arterial Hypertension, CHD - Congenital Heart Disease, CTD - Connective Heart Disease, PPH, LHD - Left Heart Disease, LD - Lung Disease, CTEPH - Chronic Thromboembolic Pulmonary Hypertension, KCO - transfer coefficient of carbon monoxide

Tag	Tag description	Cases	Controls	Excluded relatives
PH	Individuals with mPAP > 25 mmHg	1112	9134	2786
PAH	Patients with one of the following diagnoses: IPAH, HPAH, PVOD, PCH, APAH: CHD-PAH, APAH:CTD-PAH, APAH:HIV-PAH, APAH:PH-PAH	1085	9134	2786
I/HPAH	Patients with a clinical diagnosis of IPAH or HPAH	1036	9134	2786
IPAH	Patients with a clinical diagnosis of IPAH	972	9134	2785
HPAH	Patients with a clinical diagnosis of HPAH	67	9136	2779
PVOD/PCH	Patients with a clinical diagnosis of PVOD/PCH	20	9136	2778
I/HPAH/PVOD/PCH	Patients with one of the following diagnoses: IPAH, HPAH, PVOD, PCH	1056	9134	2786
FPAH	Patients with one of the following diagnoses: IPAH, HPAH, PVOD, PCH and a positive family history	80	9136	2781
APAH	Patients with one of the following diagnoses: APAH:CHD_PAH, APAH:CTD-PAH,	29	9136	2778
APAH: CHD-PAH	Patients with PAH associated with congenital heart disease	17	9136	2778
APAH: CTD-PAH	Patients with PAH associated with connective tissue disease	10	9136	2778
APAH: PoPH	Patients with PAH associated with portopulmonary hypertension	1	9136	2778
APAH: HIV-PAH	Patients with PAH associated with HIV	1	9136	2778
PH-LHD	Patients with pulmonary hypertension associated with left heart disease (Group 2)	7	9136	2778
PH-LD	Patients with pulmonary hypertension associated with lung disease (Group 3)	8	9136	2778
CTEPH	Chronic thromboembolic pulmonary hypertension (Group 4)	6	9136	2778
PH-multifactorial	Multifactorial pulmonary hypertension (Group 5)	6	9136	2778
young age	Lower age tertile (<40.8 years)	378	9136	2785
middle age	Middle age tertile (40.8 - 58.6 years)	376	9134	2779
old age	Higher age tertile (>58.6 years)	355	9136	2778
low KCO	KCO < 50% pred.	152	9136	2778
KCO lower tertile	KCO <60% pred.	211	9136	2778
KCO middle tertile	KCO 60-80% pred.	215	9136	2778
KCO higher tertile	KCO >80% pred.	215	9134	2779

Table 2. BeviMed analysis results. Posterior probabilities and Bayes factors of gene-tag associations. The "High" category, comprise only variants of "high" impact, including PTVs and large deletions; the Moderate category contains variants of impact "moderate", including missense variants or consequence "non_coding_transcript_exon_variant"; the combined category Moderate and High, include both respective consequence types.

Gene	Transcript	Tag	Log Bayes factor	Posterior probability	Consequence type	Mode of inheritance
BMPR2	ENST00000374580	I/HPAH	265.762	1.000	High	dominant
BMPR2	ENST00000374580	PAH	265.639	1.000	High	dominant
BMPR2	ENST00000374580	I/HPAH/PVOD/PCH	263.481	1.000	High	dominant
BMPR2	ENST00000374580	young age	149.576	1.000	Moderate and high	dominant
BMPR2	ENST00000374580	HPAH	149.091	1.000	Moderate and high	dominant
BMPR2	ENST00000374580	FPAH	147.822	1.000	Moderate and high	dominant
BMPR2	ENST00000374580	IPAH	144.582	1.000	High	dominant
BMPR2	ENST00000374580	KCO higher tertile	99.923	1.000	High	dominant
BMPR2	ENST00000374580	middle age	63.119	1.000	Moderate and high	dominant
BMPR2	ENST00000374580	KCO middle tertile	52.706	1.000	Moderate and high	dominant
EIF2AK4	ENST00000263791	low KCO	29.741	1.000	Moderate and high	recessive
EIF2AK4	ENST00000263791	KCO lower tertile	26.247	1.000	Moderate and high	recessive
TBX4	ENST00000240335	I/HPAH	23.783	1.000	High	dominant
TBX4	ENST00000240335	I/HPAH/PVOD/PCH	23.549	1.000	High	dominant
TBX4	ENST00000240335	PAH	23.141	1.000	High	dominant
EIF2AK4	ENST00000263791	young age	20.547	1.000	Moderate and high	recessive
TBX4	ENST00000240335	IPAH	19.990	1.000	High	dominant
EIF2AK4	ENST00000263791	I/HPAH/PVOD/PCH	15.718	1.000	Moderate and high	recessive
ACVRL1	ENST00000388922	HPAH	15.501	1.000	Moderate and high	dominant
EIF2AK4	ENST00000263791	PAH	15.407	1.000	Moderate and high	recessive
EIF2AK4	ENST00000263791	PVOD/PCH	14.441	0.999	Moderate and high	recessive
AQP1	ENST00000311813	HPAH	12.075	0.994	Moderate	dominant
EIF2AK4	ENST00000263791	FPAH	11.858	0.993	High	recessive
TBX4	ENST00000240335	young age	11.500	0.990	High	dominant
AQP1	ENST00000311813	I/HPAH	11.466	0.990	Moderate and high	dominant
KDR	ENST00000263923	KCO lower tertile	11.362	0.989	High	dominant
AQP1	ENST00000311813	I/HPAH/PVOD/PCH	11.291	0.988	Moderate and high	dominant
AQP1	ENST00000311813	PAH	11.047	0.984	Moderate and high	dominant
AQP1	ENST00000311813	FPAH	10.023	0.958	Moderate	dominant
IDH3G	ENST00000217901	KCO middle tertile	9.346	0.920	Moderate and high	dominant
KDR	ENST00000263923	old age	9.249	0.912	High	dominant
GDF2	ENST00000249598	I/HPAH	9.091	0.899	Moderate and high	dominant
BMPR2	ENST00000374580	old age	8.913	0.881	High	dominant
GDF2	ENST00000249598	I/HPAH/PVOD/PCH	8.775	0.866	Moderate and high	dominant
SOX17	ENST00000297316	young age	8.554	0.839	Moderate and high	dominant
GDF2	ENST00000249598	PAH	8.478	0.828	Moderate and high	dominant
ATP13A3	ENST00000439040	KCO higher tertile	8.035	0.755	High	dominant
GDF2	ENST00000249598	middle age	7.818	0.713	Moderate and high	dominant
KDR	ENST00000263923	low KCO	7.636	0.675	High	dominant

Table 3. Gene changes for IPAH patients harbouring protein-truncating variants (PTV) in the *KDR* gene and PTV and missense variants in the IDH3G gene. *KDR* - Kinase insert domain receptor, *IDH3G* - Isocitrate dehydrogenase (NAD(+)) 3 non-catalytic subunit gamma, WHO FC - World Health Organisation functional class, 6MWD - 6-minute walk distance, SpO_2 - arterial oxygen saturation, mRAP - mean right atrial pressure, mPAP - mean pulmonary artery pressure, mPAWP - mean pulmonary artery wedge pressure, CO - cardiac output, PVR - pulmonary vascular resistance, FEV_1 - forced expiratory volume in 1 sec, FVC - forced vital capacity, KCO - transfer factor coefficient for carbon monoxide. None of the *KDR* variants has been previously reported in gnomAD, ExAC or internal controls. For *KDR* HGVSc notations are based on transcript sequence ENST00000263923.4. HGVSp notations are based on amino acid sequence ENSP00000263923.4. None of the patients harbouring PTV in *KDR* had capillary hemangioma, *DLCO% predicted; For *IDH3G* HGVSc notations are based on transcript sequence ENST00000217901.5, HGVSp notations are based on amino acid sequence ENSP00000217901.5. Protein truncating variants were defined as stop gained, splice acceptor variants or frameshift variants.

Gene	***KDR***								***IDH3G***						
Cohort	UK				US				UK					US	
WGS ID	W000229	E003448	W000274	E001392	CUMC-JM161	CCHMC12-190	CCHMC-19-023	CCHMC-27-015	E004190	E004149	E004194	E001063	W000031	CCHMC_22-105	CCHMC_10-074
Exon	3		22	3	2	3	5	22	1-13	1	1	12	12	13	4
HGVSc	c.183G>A	c.490-1G>A	c.3064C>T	c.183del	c.161+1G>T	c.303C>A	c.658+1G>A	c.3064C>T		c.1067T>C	c.1037C>T	c.74C>T	c.74C>T	c.1091C>T	c.217G>C
HGVSp	p.Trp61Ter	-	p.Arg1022Ter	p.Trp61CysfsTer16		p.Tyr101Ter		p.Arg1022Ter		p.Met356Thr	p.Thr346Ile	p.Pro25Leu	p.Pro25Leu	p.Pro364Leu	p.Val73Leu
Consequence type	stop gained	splice acceptor variant	stop gained	frameshift variant	splice donor variant	stop gained	stop gained	stop gained	large deletion	missense variant	missense variant	missense variant	missense variant	missense variant	missense variant
Shared	PAH(1)	PAH(1)	PAH(1)	PAH(1)	No	No	No	No	GEL(1); PAH(1)	PAH(1)	PAH(1)	PAH(2)	PAH(2)	NA	NA
gnomAD	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	5.47E-06	1.09E-05
CADD_PHRED_v1.4	40	34	36	33	26	38	24	37		23.9	17.15	23.7	23.7	23.3	21.7
GerpN	5.93	5.75	5.95	5.93	5.83	5.83	5.8	5.95		5.46	5.46	5.22	5.22	5.18	4.72
Ansestry	European	European	European	European	East-Asian	European	European	European	East-Asian	European	European	European	European	European	European
Sex	male	female	male	female	female	male	female	female	female	female	female	female	female	female	male
Diagnosis	IPAH	IPAH	IPAH	IPAH	APAH-CHD secondary to double outlet RV	IPAH	IPAH	IPAH	IPAH	IPAH	IPAH	IPAH	IPAH	CHD-PAH	CHD-PAH
Age at diagnosis [years]	71	62	67	61	4	72	65	42	23	27	34	51	68	0	55
WHO FC	2	3	3	3	2	NA	NA	NA	4	3	4	4	2	3	3
6MWD [m]	472	422	660	180	NA	380	NA	245	350		414		414	NA	316
SpO_2 pre [%]	95	97	98	97	NA	NA	NA	NA	99	96	95	98	96	NA	
SpO_2 post [%]	86	86	NA	91	NA	NA	NA	NA	97		99	96	95	NA	
FEV_1 [% pred.]	116	90	83	67.3	85%	NA	77%	NA	74	87	104	95	99.1	NA	
FVC [% pred.]	115	94	91	72.8	92%	NA	83%	NA	76	90	109	95.8	96.3	NA	
TLC [% pred.]	NA	NA	NA	NA	NA	NA	65%	NA	NA	NA	105	76	98	NA	
KCO [% pred.]	44	46	46	55.2	NA	NA	35%*	NA	73	71	64	78	73	NA	
Smoking history	Never	Never	Ex-smoker	Never	Never	Never	Ex-smoker	Never	Never	Ex-smoker	Never	Never	Never	Never	
mRAP [mmHg]	5	8	8	3	NA	5	29	14	15	14	8	12	6	3	7
mPAP [mmHg]	62	57	41	44	NA	49	66	60	58	64	49	50	62	46	69
PAWP [mmHg]	4	15	12	9	NA	5	16	15	15	8	10	12	7	NA	10
CO [L/min]	3.6	4.58	5.97	5.23	NA	4.33	1.8	4.6	2.37	3.23	NA	3.29	4.1	4.4	
PVR	16.11	9.17	4.86	6.69	NA	NA	27.9	9.8	18.1	17.3	NA	11.6	13.4	NA	
Comrobidities	hyperlipidemia, HTN, DM type 2	HTN, hypothyrodism	DM type 2	CAD, DM type 2	No	HTN, hyperlipidemia,	HTN, hypothyrpoidism, OA	Obesity, CAD, DM type 2, hypothyroidism	No	No	No	PFO	No	Scimitar syndrome, hypoplastic rght lung, ASD with spontaneous closure	Large ASD
Family history	No	No	No	No	No	No	No	No	No	No	No	No	No	?	?
Status	alive	alive	alive	dead	alive	alive	alive	alive	alive	alive	alive	dead	alive	alive	alive

Table 4. Clinical characteristics of IPAH patients harbouring protein truncating variants in the *KDR* gene. KDR - kinase insert domain receptor, IPAH - idiopathic pulmonary arterial hypertension, BMI - Body Mass Index, WHO FC - World Health Organisation functional class, 6MWD - six-minute walk distance, SpO_2 - arterial oxygen saturation, mRAP - mean right atrial pressure, mPAP - mean pulmonary artery pressure, PAWP - pulmonary artery wedge pressure, CO - cardiac output, PVR - pulmonary vascular resistance, NO - nitric oxide, FEV_1 - forced expiratory volume in 1 second, FVC - forced vital capacity, KCO - transfer factor coefficient for carbon monoxide, COPD - chronic obstructive pulmonary disease, CAD - coronary artery disease, HTN - systemic hypertension, CKD - chronic kidney disease, Hb - haemoglobin, WBC - white blood cells, TSH - thyroid-stimulating hormone. Comorbidities are reported as the number and percentage of cases possessing a disease entity. None of the patients had a history of pulmonary embolism or asthma. Three of the KDR missense variants co-occurred with predicted deleterious variants in established PAH risk genes (*BMPR2* and *AQP1*)

	***KDR* missense N=13**	***KDR* PTV N=4**	**p-value**	**N**
Diagnosis verified: IPAH	13 (100%)	4 (100%)	.	17
Age [years]	46 [36;59]	64 [62;68]	0.113	17
Sex: female	9 (69%)	2 (50%)	0.584	17
BMI [kg/m^2]	29 [24;32]	26 [26;30]	1	13
WHO FC: II/III/IV [%]	23.1/9.2/7.7	25/75/0	1	17
6MWD [m]	312 [150;355]	301 [240;362]	0.814	11
SpO_2 pre [%]	95 [93;97]	97 [96;97]	0.335	11
SpO_2 post [%]	90 [80;96]	86 [86;88]	0.926	12
mRAP [mmHg]	8 [6;13]	6 [4;8]	0.431	14
mPAP [mmHg]	53 [42;62]	50 [43;58]	0.896	15
PAWP [mmHg]	10 [8;13]	10 [8;13]	0.642	13
CO [L/min]	4.0 [3.0;5.5]	4.9 [4.3;5.4]	0.514	15
PVR [WU]	10.2 [4.56;14.3]	7.93 [6.23;10.9]	1	13
Acute NO challenge: vasoresponder	1 (33.3%)	1 (25.0%)	1	7
FEV_1 [% pred.]	84 [65;94]	86 [79;96]	0.48	14
FVC [% pred.]	86 [72;97]	92 [86;99]	0.723	14
FEV_1/FVC ratio	0.78 [0.75;0.87]	0.78 [0.76;0.79]	0.671	14
KCO [% pred.]	89 [74;93]	46 [46;48]	0.008	11
Smoking history	6 (54.5%)	1 (25.0%)	0.677	15
COPD	1 (7.69%)	0 (0.00%)	1	17
Pulmonary fibrosis	0 (0.00%)	2 (50.0%)	0.044	17
CAD	1 (7.69%)	1 (25.0%)	0.426	17
HTN	5 (38.5%)	2 (50.0%)	1	17
CKD	1 (7.69%)	0 (0.00%)	1	17
Hb [g/l]	154 [140;166]	148 [135;152]	0.214	15
WBC [x10e9/l]	9.20 [6.30;11.0]	8.80 [8.23;9.55]	0.844	15
Platelets [x10e9/l]	262 [209;294]	216 [188;251]	0.361	15
Creatinine [umol/l]	78.0 [61.5;98.0]	67.0 [66.5;96.5]	0.866	13
TSH [mU/l]	3.65 [1.80;6.90]	1.76 [1.72;1.84]	0.234	12

Table S1. NIHR BioResource - Rare Diseases domain definitions

Project acronym	Project	Number of individuals in the project
GEL	Genomics England Ltd	3058
BPD	Bleeding, Thrombotic and Platelet Disorders	986
PID	Primary Immune Disorders	1027
CNTRL	Processed Controls	50
IRD	Inherited Retinal Disorders	717
NDD	Neurological and Developmental Disorders	518
EDS	Ehlers Danlos Syndrome	15
HCM	Hypertrophic Cardiomyopathy	239
PMG	Primary Membranoproliferative Glomerulonephritis	181
SRNS	Steroid Resistant Nephrotic Syndrome	234
CSVD	Cerebral Small Vessel Disease	134
NPD	Neuropathic Pain Disorder	185
ICP	Intrahepatic Cholestasis of Pregnancy	267
LHON	Leber Hereditary Optic Neuropathy	54
MPMT	Multiple Primary Tumours	554
SMD	Stem Cell & Myeloid Disorders	153
PAH	Pulmonary arterial hypertension	1123
UKBio	UK BioBank	764
		10259

Table S2. Summary of electronic clinical report forms (CRFs) constructed to capture phenotype information

ID capture
Demographics
Functional class
Clinical features by history
Clinical features by examination
Risk factors
Haemodynamics
Echocardiography
Electrocardiogram
Lung function
Associated Diseases
Clinical blood tests
Survival
Arterial blood gases
Imaging
Exercise performance
Body system
Drug treatment history (PAH)
Drug treatment history (other)
Family history
Epidemiology questionnaire

Table S3. Reporting proforma for CT scan revision. CTPA - Computerised Tomography Pulmonary Angiogram, HRCT - High Resolution Computerised Tomography, GGO - ground glass opacities

Parameter	Response
ID	character
Reader	character
CT scan date	date
Slice thickness	numeric
Number of slices	numeric
CTPA	done/not done
HRCT	done/not done
Expiratory CT	done/not done
Pleural effusion	Nil; Trace; Mild; Moderate; Severe
Subcutaneous oedema	present, absent
Severity of GGO centrilobular parttern	Nil; Trace; Mild; Moderate; Severe
Severity of GGO non-specific mosaic pattern	Nil; Trace; Mild; Moderate; Severe
Distribution of GGO	C-central; U-upper; Z-zonal; D-diffuse
Pulmonary arteriovenous malformations	present, absent
Largest bronchial artery size	numeric [mm]
Mediastinal venous collaterals	present, absent
Intralobular septal thickening	Nil; Trace; Mild; Moderate; Severe
Mediastinal lymphadenopathy	present, absent
Mediastinal lymphadenopathy subcarinal	[mm]
Emphysema	Nil; Mild; Moderate; Severe
Bronchial wall thickening	Nil; Trace; Mild; Moderate; Severe
Fibrosis	Nil; Mild; Moderate; Severe
Air trapping	Nil; Trace; Mild; Moderate; Severe
Subpleural scarring	Nil; Mild; Moderate; Severe

Table S4. Clinical characterisation of the study population. BMI - body mass index, WHO FC - World Health Organisation Functional Class, 6MWD - 6-minute walk distance, mRAP - mean right atrial pressure, mPAP - mean pulmonary artery pressure, CO - cardiac output, FEV_1 - forced expiratory capacity in 1 second, FVC - forced vital capacity, KCO - transfer factor coefficient for carbon monoxide, Hb - haemoglobin, RDW - red cell distribution width, WBC - white blood cell count, NTproBNP - N-terminal pro b-type natriuretic peptide, BNP - B-type natriuretic peptide, CRP - C-Reactive Protein Protein, HTN - hypertension, DM - diabetes mellitus, CAD - coronary artery disease, CVA - cerebrovascular accident, COPD - chronic obstructive pulmonary disease, CCB - calcium channel blocker, ERA - endothelin receptor antagonists, PA - prostacyclin analogues, PED5 - phosphodiesterase type 5, sGC - soluble guanylate cyclase; Entire cohort (n=1122) was composed of IPAH (n=972), HPAH (n=73), PVOD/PCH (n=20), PH associated with left heart disease (n=7), PH associated with lung disease (n=8), chronic thromboembolic pulmonary hypertension (n=6), multifactorial PH (n=6), hereditary hemorrhagic telangiectasia (n=1)

	ALL N=1122	**N**	**I/HPAH and PVOD/PCH N=1065**	**N**
Demographics and functional status				
Sex: female	760 (68%)	1116	732 (69%)	1064
Age [years]	49 [35;63]	1112	49 [35;63]	1061
BMI [kg/m^2]	27 [23;32]	1015	27 [23;31]	970
WHO FC: I/II/III/IV	21 (2%)/217 (20%)/703 (65%)/138 (13%)	1079	21 (2%)/210 (20%)/663 (64%)/135 (13%)	1029
6MWD [m]	335 [220;415]	953	336 [220;415]	906
Haemodynamics				
mRAP [mmHg]	8 [5;12]	985	8 [5;12]	939
mPAP [mmHg]	53 [44;61]	1052	53 [44;61]	1004
CO [L/min]	3.9 [3.1;4.9]	1003	3.9 [3.1;4.9]	960
FEV_1 [% pred.]	85 [73;97]	849	86 [74;97]	811
FVC [% pred.]	94 [81;106]	831	95 [82;106]	793
KCO [% pred.]	71 [52;86]	644	71 [52;86]	610
Clinical blood tests				
Hb [g/l]	151 [138;165]	847	152 [138;164]	805
RDW [%]	14 [14;16]	413	14 [14;16]	392
WBC [x10e9/l]	8.2 [6.8;9.8]	839	8.2 [6.8;9.8]	797
Platelets [x10e9/l]	224 [182;272]	836	225 [183;274]	795
Creatinine [umol/l]	86 [70;102]	832	86 [70;102]	790
NTproBNP [ng/l]	926 [215;2637]	276	963 [217;2672]	265
BNP [ng/l]	195 [72;432]	271	197 [74.6;454]	252
CRP [mg/l]	4 [2;8]	639	4 [2;8]	604
Comorbidities				
HTN	265 (24%)	1122	256 (24%)	1065
DM type 1	20 (2%)	1122	19 (2%)	1065
DM type 2	138 (12%)	1122	132 (12%)	1065
CAD	45 (4%)	1122	42 (4%)	1065
CVA	17 (2%)	1122	15 (1%)	1065
Hypothyroidism	135 (12%)	1122	130 (12%)	1065
COPD	66 (6%)	1122	57 (5%)	1065
Asthma	78 (7%)	1122	74 (7%)	1065
Cancer	4 (0.4%)	1122	3 (0.3%)	1065
Medication				
Initial therapy:		631		604
CCB	73 (12%)		72 (12%)	
combination therapy	279 (44%)		269 (45%)	
ERA	86 (14%)		82 (14%)	
PA	42 (7%)		41 (7%)	
PDE5 inhibitor	150 (24%)		139 (23%)	
sGC stimulator	1 (0%)		1 (0%)	

Table S5. Clinical differences between patients with present and missing transfer coefficient of the lung for carbon monoxide (KCO) [% pred.]

BMI - body mass index, WHO FC - World Health Organisation Functional Class, 6MWD - 6-minute walk distance, mRAP - mean right atrial pressure, mPAP - mean pulmonary artery pressure, PAWP - Pulmonary Artery Wedge Pressure, CI - cardiac index, PVR - pulmonary vascular resistance, SvO_2 - mixed venous saturation, HRCT - High-Resolution Computerised Tomography, FEV_1 - forced expiratory capacity in 1 second, FVC - forced vital capacity, COPD - chronic obstructive pulmonary disease, OSA - obstructive sleep apnoea

	[ALL] N=1122	KCO [% pred.] missing N=478	KCO [% pred.] present N=644	p-value	N
Sex: Female	760 (68.1%)	326 (69.1%)	434 (67.4%)	0.597	1116
Prevalent cases	852 (75.9%)	411 (86.0%)	441 (68.5%)	<0.001	1122
BMI [kg/m^2]	26.9 [23.1;31.5]	25.6 [22.0;30.1]	27.7 [24.1;32.5]	<0.001	1015
Age [years]	48.8 [34.8;62.7]	44.6 [31.0;58.5]	51.3 [38.1;65.5]	<0.001	1112
WHO FC				<0.001	1079
I	21 (1.95%)	13 (2.88%)	8 (1.27%)		
II	217 (20.1%)	117 (25.9%)	100 (15.9%)		
III	703 (65.2%)	269 (59.6%)	434 (69.1%)		
IV	138 (12.8%)	52 (11.5%)	86 (13.7%)		
6MWD [m]	340 [230;418]	364 [289;432]	314 [192;405]	<0.001	702
FEV_1 [% pred.]	85.0 [73.0;97.0]	83.2 [69.6;97.0]	86.0 [74.0;97.0]	0.2	849
FVC [% pred.]	94.0 [81.4;106]	89.0 [73.0;103]	96.0 [82.6;107]	<0.001	831
TLC [% pred.]	95.0 [85.0;104]	93.2 [83.7;104]	95.0 [86.0;103]	0.625	639
mRAP [mmHg]	8.00 [5.00;12.0]	7.00 [5.00;12.0]	9.00 [6.00;12.0]	<0.001	985
mPAP [mmHg]	53.0 [44.0;61.0]	53.0 [44.0;61.0]	53.0 [45.0;61.0]	0.669	1052
PAWP [mmHg]	9.00 [7.00;12.0]	9.00 [6.00;11.0]	10.0 [7.00;12.0]	0.001	934
CI [L/min/m^2]	2.17 [1.72;2.67]	2.31 [1.81;2.82]	2.07 [1.68;2.59]	<0.001	946
PVR [WU]	11.0 [7.69;15.1]	10.6 [7.60;15.2]	11.1 [7.69;15.1]	0.622	893
SvO_2 [%]	64.0 [58.0;70.0]	64.7 [57.8;70.0]	63.8 [58.2;70.0]	0.732	817
Fibrosis [HRCT report]:				0.445	614
none	586 (95.4%)	169 (96.0%)	417 (95.2%)		
minimal/mild	26 (4.23%)	6 (3.41%)	20 (4.57%)		
moderate	1 (0.16%)	0 (0.00%)	1 (0.23%)		
severe	1 (0.16%)	1 (0.57%)	0 (0.00%)		
Emphysema [HRCT report]:				0.029	612
none	560 (91.5%)	169 (96.0%)	391 (89.7%)		
minimal/mild	33 (5.39%)	3 (1.70%)	30 (6.88%)		
moderate	15 (2.45%)	4 (2.27%)	11 (2.52%)		
severe	4 (0.65%)	0 (0.00%)	4 (0.92%)		
Smoking history: past/current	435 (50.6%)	97 (38.5%)	338 (55.7%)	<0.001	859
COPD	66 (5.88%)	24 (5.02%)	42 (6.52%)	0.353	1122
OSA	58 (5.17%)	20 (4.18%)	38 (5.90%)	0.251	1122
Asthma	78 (6.95%)	15 (3.14%)	63 (9.78%)	<0.001	1122

Table S6 Characteristics of unrelated individuals used in genetic association analysis by KCO. Abbreviations: BMI - body mass index, WHO FC - World Health Organisation Functional Class, 6MWD - 6-minute walk distance, SpO_2 - peripheral capillary oxygen saturation, mRAP - mean right atrial pressure, mPAP - mean pulmonary artery pressure, PAWP - Pulmonary Artery Wedge Pressure, CO - cardiac output, SvO_2 - Mixed venous oxygen saturation, FEV_1 - forced expiratory capacity in 1 second, FVC - forced vital capacity, TLC - Total Lung Capacity, KCO - transfer coefficient of carbon monoxide, HRCT - High-Resolution Computerised Tomography, NTproBNP - N-terminal pro B-Type Natriuretic Peptide, BNP - B-Type Natriuretic Peptide, CRP - C-Reactive Protein Protein, Hb - haemoglobin, WBC - white blood cell count, COPD - chronic obstructive pulmonary disease, OSA - obstructive sleep apnoea, CAD - coronary artery disease, CVA - cerebrovascular accident, PAD - peripheral artery disease,

	[ALL] N=644	KCO > 50% pred N=492	KCO =< 50 % pred. N=152	p-value	N
Age [years]	51 [38;66]	47 [36;60]	66 [54;71]	<0.001	644
Sex: female	434 (67%)	352 (72%)	82 (54%)	<0.001	644
Incident cases	203 (31.5%)	134 (27.2%)	69 (45.4%)	<0.001	644
BMI [kg/m²]	27.7 [24.1;32.5]	27.7 [23.7;32.7]	27.7 [24.8;32.3]	0.806	629
WHO FC				0.013	628
I	8 (1%)	7 (1%)	1 (1%)		
II	100 (16%)	88 (18%)	12 (8%)		
III	434 (69%)	320 (67%)	114 (77%)		
IV	86 (14%)	64 (13%)	22 (15%)		
6MWD [m]	313 [190;404]	334 [229;414]	219 [120;348]	<0.001	599
SpO_2 pre [%]	95.0 [93.0;97.0]	96.0 [93.0;98.0]	92.0 [89.0;95.0]	<0.001	575
SpO_2 post [%]	91.0 [85.0;96.0]	93.0 [88.0;96.0]	83.0 [76.0;88.0]	<0.001	529
mRAP [mmHg]	9 [6;12]	9 [6;13]	9 [6;12]	0.375	601
mPAP [mmHg]	53 [45;61]	54 [46;63]	50 [42;57]	0.001	625
PAWP [mmHg]	10 [7;12]	10 [7;12]	10 [8;12]	0.301	560
CO [L/min]	3.8 [3.1;4.8]	3.9 [3.1;4.9]	3.6 [3.1;4.7]	0.282	614
SvO_2 [%]	64 [58;70]	64 [59;71]	61 [55;67]	<0.001	572
Acute NO challenge: vasoresponder	43 (17%)	38 (18%)	5 (10%)	0.204	257
FEV_1 [% pred.]	86.0 [74.0;97.0]	85.0 [73.4;96.0]	87.0 [77.0;98.0]	0.119	639
FVC [% pred.]	96.0 [82.6;107]	94.0 [81.0;106]	101 [87.0;113]	<0.001	628
FEV_1/FVC ratio	0.76 [0.69;0.81]	0.76 [0.71;0.81]	0.70 [0.63;0.77]	<0.001	614
TLC [% pred.]	95.0 [86.0;103]	95.0 [85.0;103]	95.0 [87.5;104]	0.564	485
KCO [%pred.]	71 [52;86]	78 [67;90]	37 [30;44]	<0.001	644
Emphysema (HRCT scan)				<0.001	436
none	391 (90%)	307 (95%)	84 (74%)		
minimal/mild	30 (7%)	12 (4%)	18 (16%)		
moderate	11 (3%)	2 (1%)	9 (8%)		
severe	4 (1%)	1 (0%)	3 (3%)		
Fibrosis (HRCT scan)				<0.001	438
none	417 (95%)	319 (98%)	98 (87%)		
minimal/mild	20 (5%)	6 (2%)	14 (12%)		
moderate	1 (0%)	0 (0%)	1 (1%)		
Smoking history: past/current	338 (55.7%)	233 (50.0%)	105 (74.5%)	<0.001	607
NTproBNP [ng/l]	980 [248;2673]	866 [257;2590]	1225 [249;2878]	0.359	220
BNP [ng/l]	200 [72.9;432]	198 [69.7;456]	200 [82.2;328]	0.798	143
Uric acid [mmol/l]	0.42 [0.32;0.53]	0.40 [0.30;0.51]	0.48 [0.39;0.56]	0.002	231
CRP [mg/l]	5.00 [2.00;8.60]	5.00 [2.00;8.00]	4.15 [2.00;9.00]	0.953	450
Hb [g/l]	154 [139;166]	153 [138;166]	154 [142;166]	0.658	603
WBC [x10e9/l]	8 [7;10]	8 [7;10]	9 [7;10]	0.011	597
Platelets [x10e9/l]	220 [179;268]	220 [179;270]	220 [182;254]	0.516	595
Sodium [mmol/l]	139 [138;141]	139 [138;141]	140 [138;141]	0.858	597
Potassium [mmol/l]	4.20 [4.00;4.50]	4.20 [3.90;4.50]	4.20 [4.00;4.50]	0.412	591
Urea [mmol/l]	5.80 [4.60;7.70]	5.60 [4.40;7.10]	7.00 [5.30;9.20]	<0.001	594
Creatinine [umol/l]	88.0 [73.2;105]	87.0 [73.0;102]	95.0 [78.5;114]	0.002	598
COPD	42 (7%)	23 (5%)	19 (12%)	0.001	644
Asthma	63 (10%)	54 (11%)	9 (6%)	0.093	644
OSA	38 (6%)	29 (6%)	9 (6%)	1	644
CAD	26 (4%)	10 (2%)	16 (11%)	<0.001	644
CVA	10 (2%)	5 (1%)	5 (3%)	0.061	644
PAD	2 (0%)	0 (0%)	2 (1%)	0.055	644
HTN	170 (26%)	116 (24%)	54 (36%)	0.005	644
DM type 1	9 (1%)	8 (2%)	1 (1%)	0.693	644
DM type 2	93 (14%)	60 (12%)	33 (22%)	0.005	644
Hypothyroidism	73 (11%)	60 (12%)	13 (9%)	0.275	644
Sjogren syndrome	3 (0%)	2 (0%)	1 (1%)	0.555	644

Table S1. Clinical characteristics of patients used in derivation cohort, KCO - transfer coefficient of carbon monoxide, 6MWD - 6-minute walk distance, SpO2 - peripheral oxygen saturation, mPAP - mean pulmonary artery pressure, PAWP - Pulmonary Artery Wedge Pressure, CO - cardiac output, SvO2 - mixed venous oxygen saturation, NO - nitric oxide, FEV_1 - forced expiratory capacity in 1 second, FVC - forced vital capacity, TLC - Total Lung Capacity, KCO - transfer coefficient of carbon monoxide, HRCT - High-Resolution Computerised Tomography, NTproBNP - N-terminal pro b-type natriuretic peptide, BNP - B-type natriuretic peptide, CRP - C-Reactive Protein Protein, Hb - haemoglobin, WBC - white blood cell count, COPD - chronic obstructive pulmonary disease, OSA - obstructive sleep apnoea, CAD - coronary artery disease, CVA

	[ALL] N=644	Higher tertile N=214	Middle tertile N=215	Lower tertile N=215	p.overall	N
Age [years]	51 [38;66]	44 [37;58]	49 [34;61]	64 [50;71]	<0.001	644
Sex: female	434 (67%)	149 (70%)	160 (74%)	125 (58%)	0.001	644
Incident cases	203 (31.5%)	58 (27.1%)	54 (25.1%)	91 (42.3%)	<0.001	644
BMI [kg/m^2]	27.7 [24.1;32.5]	29.1 [25.3;34.5]	26.8 [23.1;30.9]	27.4 [24.2;32.3]	<0.001	629
WHO FC					0.028	628
I	8 (1%)	4 (2%)	2 (1%)	2 (1%)		
II	100 (16%)	40 (19%)	40 (19%)	20 (10%)		
III	434 (69%)	143 (68%)	135 (65%)	156 (74%)		
IV	86 (14%)	22 (11%)	32 (15%)	32 (15%)		
6MWD [m]	313 [190;404]	331 [240;420]	340 [230;418]	240 [131;360]	<0.001	599
SpO_2 pre [%]	95.0 [93.0;97.0]	96.0 [94.0;98.0]	96.0 [93.0;98.0]	93.0 [90.0;96.0]	<0.001	575
SpO_2 post [%]	91.0 [85.0;96.0]	94.0 [89.0;96.0]	93.0 [88.0;96.0]	85.0 [79.0;91.0]	<0.001	529
mRAP [mmHg]	9 [6;12]	9 [7;13]	9 [6;12]	8 [5;12]	0.224	601
mPAP [mmHg]	53 [45;61]	55 [47;65]	53 [46;62]	51 [42;57]	<0.001	625
PAWP [mmHg]	10 [7;12]	10 [8;12]	10 [7;12]	10 [7;12]	0.487	560
CO [L/min]	3.8 [3.1;4.8]	3.8 [3.0;5.0]	4.0 [3.1;4.8]	3.8 [3.1;4.8]	0.967	614
SvO_2 [%]	64 [58;70]	64 [60;71]	65 [59;70]	62 [56;68]	0.001	572
Acute NO challenge: vasoresponder	43 (17%)	15 (17%)	21 (24%)	7 (9%)	0.024	257
FEV_1 [% pred.]	86.0 [74.0;97.0]	83.8 [72.9;95.0]	86.0 [73.9;97.4]	87.0 [77.0;98.0]	0.08	639
FVC [% pred.]	96.0 [82.6;107]	90.0 [79.8;101]	95.9 [83.0;108]	100 [86.9;112]	<0.001	628
FEV_1/FVC ratio	0.76 [0.69;0.81]	0.78 [0.72;0.82]	0.76 [0.69;0.81]	0.71 [0.65;0.78]	<0.001	614
TLC [% pred.]	95.0 [86.0;103]	94.0 [86.2;103]	96.0 [83.0;104]	97.0 [87.0;102]	0.787	485
KCO [% pred.]	71 [52;86]	92 [86;101]	71 [67;76]	42 [33;52]	<0.001	644
Emphysema (HRCT scan):					<0.001	436
none	391 (90%)	144 (99%)	128 (96%)	119 (76%)		
minimal/mild	30 (7%)	2 (1%)	4 (3%)	24 (15%)		
moderate	11 (3%)	0 (0%)	0 (0%)	11 (7%)		
severe	4 (1%)	0 (0%)	1 (1%)	3 (2%)		
Fibrosis (HRCT scan):					<0.001	438
none	417 (95%)	145 (99%)	132 (99%)	140 (89%)		
minimal/mild	20 (5%)	2 (1%)	2 (1%)	16 (10%)		
moderate	1 (0%)	0 (0%)	0 (0%)	1 (1%)		
Smoking history: past/current	338 (55.7%)	89 (42.4%)	106 (52.7%)	143 (73.0%)	<0.001	607
NTproBNP [ng/l]	980 [248;2673]	842 [167;2358]	902 [310;2593]	1225 [237;2811]	0.539	220
BNP [ng/l]	200 [72.9;432]	193 [71.7;392]	145 [67.5;427]	214 [82.2;448]	0.659	143
Uric acid [mmol/l]	0.42 [0.32;0.53]	0.40 [0.28;0.47]	0.40 [0.32;0.51]	0.48 [0.38;0.55]	0.011	231
CRP [mg/l]	5.00 [2.00;8.60]	5.00 [2.00;8.00]	5.00 [2.00;9.32]	4.00 [2.00;8.57]	0.496	450
Hb [g/l]	154 [139;166]	160 [145;169]	149 [136;164]	151 [140;165]	<0.001	603
WBC [x10e9/l]	8 [7;10]	8 [7;9]	8 [7;10]	9 [7;10]	0.005	597
Platelets [x10e9/l]	220 [179;268]	224 [183;262]	219 [174;276]	217 [184;262]	0.89	595
Sodium [mmol/l]	139 [138;141]	140 [138;141]	139 [137;141]	140 [138;141]	0.623	597
Potassium [mmol/l]	4.20 [4.00;4.50]	4.30 [4.00;4.50]	4.20 [3.90;4.40]	4.30 [4.00;4.50]	0.13	591
Urea [mmol/l]	5.80 [4.60;7.70]	5.50 [4.43;7.00]	5.70 [4.30;7.10]	6.70 [5.15;8.80]	<0.001	594
Creatinine [µmol/l]	88.0 [73.2;105]	87.0 [73.0;99.5]	86.0 [72.0;102]	92.5 [77.0;111]	0.003	598
COPD	42 (7%)	5 (2%)	12 (6%)	25 (12%)	<0.001	644
Asthma	63 (10%)	26 (12%)	25 (12%)	12 (6%)	0.039	644
OSA	38 (6%)	12 (6%)	11 (5%)	15 (7%)	0.698	644
CAD	26 (4%)	2 (1%)	4 (2%)	20 (9%)	<0.001	644
CVA	10 (2%)	1 (0%)	3 (1%)	6 (3%)	0.174	644
PAD	2 (0%)	0 (0%)	0 (0%)	2 (1%)	0.332	644
HTN	170 (26%)	46 (21%)	53 (25%)	71 (33%)	0.02	644
DM type 1	9 (1%)	2 (1%)	5 (2%)	2 (1%)	0.526	644
DM type 2	93 (14%)	22 (10%)	22 (10%)	49 (23%)	<0.001	644
Hypothyroidism	73 (11%)	25 (12%)	30 (14%)	18 (8%)	0.185	644
Sjogren syndrome	3 (0%)	2 (1%)	0 (0%)	1 (0%)	0.331	644

Table S8. Result of Cox regression analysis relating overall survival to selected variables at baseline. CI - confidence interval, 6MWD - 6-minute walking distance, mPAP - mean pulmonary arterial pressure, mRAP - mean right atrial pressure, PVR - pulmonary vascular resistance, WU - Wood units, KCO - transfer coefficient of carbon monoxide, CAD - coronary artery disease, COPD - chronic obstructive pulmonary disease, HTN - systemic hypertension, HRCT - High-Resolution Computerised Tomography

			Univariate		Multivariate	
	No event N=481	Event N=163	HR [95%CI]	p-value	HR [95%CI]	p-value
Sex:				<0.001		<0.001
female	343 (71.3%)	91 (55.8%)	Ref.		Ref.	
male	138 (28.7%)	72 (44.2%)	1.98 [1.45;2.70]		2.93 [1.81;4.75]	
Age [years]	4.80 (1.50)	6.17 (1.59)	1.75 [1.56;1.96]	<0.001	1.57 [1.27;1.94]	<0.001
Incident/Prevalent:				<0.001		0.131
incident	151 (31.4%)	52 (31.9%)	Ref.		Ref.	
prevalent	330 (68.6%)	111 (68.1%)	0.40 [0.28;0.58]		0.65 [0.37;1.14]	
6MWD [m]	32.9 (15.0)	21.7 (12.9)	0.95 [0.94;0.96]	<0.001	0.97 [0.95;0.99]	0.002
mRAP [mmHg]	1.84 (1.07)	2.15 (1.13)	1.26 [1.11;1.44]	0.001	1.28 [1.05;1.57]	0.016
mPAP [mmHg]	10.9 (2.75)	10.4 (2.34)	0.94 [0.88;1.00]	0.038	1.08 [0.97;1.2]	0.142
CI [L/min/m^2]	2.28 (0.78)	2.10 (0.68)	0.67 [0.53;0.86]	0.002	0.98 [0.62;1.55]	0.923
PVR [WU]	12.1 (5.96)	11.8 (4.94)	1.00 [0.97;1.03]	0.89		
KCO [%pred]	7.35 (2.22)	5.71 (2.41)	0.71 [0.66;0.77]	<0.001	0.79 [0.7;0.88]	<0.001
Smoking history:				0.002		0.692
no	215 (47.5%)	54 (35.1%)	Ref.		Ref.	
past/current	238 (52.5%)	100 (64.9%)	1.67 [1.20;2.33]		1.11 [0.66;1.88]	
CAD:				0.002		0.081
no	467 (97.1%)	151 (92.6%)	Ref.		Ref.	
yes	14 (2.91%)	12 (7.36%)	2.51 [1.39;4.53]		0.38 [0.13;1.13]	
COPD:				0.141		0.268
no	452 (94.0%)	150 (92.0%)	Ref.		Ref.	
yes	29 (6.03%)	13 (7.98%)	1.53 [0.87;2.69]		0.59 [0.23;1.5]	
HTN:				<0.001		
no	374 (77.8%)	100 (61.3%)	Ref.		Ref.	
yes	107 (22.2%)	63 (38.7%)	1.92 [1.40;2.63]		1.12 [0.68;1.84]	
Emphysema (HRCT scan):				0.012		
none	300 (91.5%)	91 (84.3%)	Ref.		Ref.	
minimal/mild	19 (5.79%)	11 (10.2%)	2.12 [1.13;3.98]		0.85 [0.38;1.92]	0.696
moderate	6 (1.83%)	5 (4.63%)	2.83 [1.15;6.98]		0.7 [0.19;2.61]	0.594
severe	3 (0.91%)	1 (0.93%)	2.36 [0.33;17.0]		0 [0;Inf]	0.996
Fibrosis (HRCT scan):				0.003		
none	321 (97.0%)	96 (89.7%)	Ref.		Ref.	
minimal/mild	10 (3.02%)	10 (9.35%)	2.79 [1.45;5.36]		0.83 [0.36;1.93]	0.672
moderate	0 (0.00%)	1 (0.93%)	3.29 [0.46;23.6]		3.98 [0.47;33.58]	0.204

Table S9. Clinical characteristics of undifferentiated individuals

mRAP - mean right atrial pressure, mPAP - mean pulmonary artery pressure, PAWP - pulmonary artery Wedge Pressure, CO - cardiac output, SvO_2 - Mixed venous oxygen saturation, NO - nitric oxide, FEV_1 - forced expiratory capacity in 1 second, FVC - forced vital capacity, TLC - Total Lung Capacity, KCO - transfer factor coefficient for carbon monoxide, NTproBNP - N-terminal pro b-type natriuretic peptide, BNP - B-type natriuretic peptide, CRP - C-Reactive Protein Protein, Hb - haemoglobin, WBC - white blood cell count, COPD - chronic obstructive pulmonary disease, OSA - obstructive sleep apnoea, CAD - coronary artery disease, CVA - cerebrovascular accident, PAD - peripheral artery disease, HTN - hypertension, DM - diabetes mellitus; none of the patients had undifferentiated connective tissue disease, incident cases were defined as those diagnosed within 6 months from study commencement.

	[ALL] N=1112	Lower tertile N=381	Middle tertile N=376	Higher tertile N=355	p.overall	N
Age [years]	49 [35;63]	30 [25;35]	50 [45;54]	68 [64;73]	<0.001	1112
Sex: female	757 (68%)	276 (72%)	273 (73%)	208 (59%)	<0.001	1112
Incident cases	270 (24.3%)	77 (20.2%)	74 (19.7%)	119 (33.5%)	<0.001	1112
BMI [kg/m^2]	26.9 [23.1;31.5]	24.2 [21.0;29.3]	28.0 [24.3;32.5]	28.1 [25.1;32.0]	<0.001	1015
WHO FC					<0.001	1078
I	21 (2%)	15 (4%)	3 (1%)	3 (1%)		
II	217 (20%)	95 (26%)	79 (21%)	43 (12%)		
III	703 (65%)	201 (56%)	248 (67%)	254 (73%)		
IV	137 (13%)	50 (14%)	40 (11%)	47 (14%)		
6MWD [m]	335 [220;415]	375 [302;460]	343 [250;420]	236 [134;348]	<0.001	953
SpO_2 pre [%]	96.0 [93.0;97.0]	97.0 [95.0;98.0]	96.0 [93.0;97.0]	94.0 [90.0;96.0]	<0.001	890
SpO_2 post [%]	91.0 [85.0;95.0]	94.0 [88.0;97.0]	92.0 [86.0;96.0]	88.0 [82.0;92.8]	<0.001	830
mRAP [mmHg]	8 [5;12]	8 [5;12]	9 [6;13]	8 [5;12]	0.046	984
mPAP [mmHg]	53 [44;61]	55 [47;66]	55 [48;62]	48 [40;57]	<0.001	1051
PAWP [mmHg]	9 [7;12]	9 [6;11]	9 [7;12]	10 [7;13]	0.004	933
CO [L/min]	3.9 [3.1;4.9]	4.0 [3.1;5.0]	3.9 [3.1;4.9]	3.8 [3.2;4.8]	0.651	1003
SvO_2 [%]	64 [58;70]	67 [60;72]	64 [58;70]	62 [57;67]	<0.001	817
Acute NO challenge	59 (14%)	31 (17%)	21 (14%)	7 (7%)	0.078	435
FEV_1 [% pred.]	85.0 [73.0;97.0]	87.0 [77.0;97.0]	84.0 [71.0;96.0]	85.2 [71.1;97.9]	0.34	849
FVC [% pred.]	94.0 [81.4;106]	90.0 [81.0;102]	96.0 [82.0;108]	96.3 [82.0;108]	0.018	831
FEV_1/FVC ratio	0.76 [0.69;0.81]	0.81 [0.77;0.86]	0.75 [0.69;0.80]	0.72 [0.66;0.77]	<0.001	760
TLC [% pred.]	95.0 [85.0;104]	95.0 [85.0;103]	95.0 [87.0;106]	93.6 [83.0;102]	0.082	639
KCO [% pred.]	71 [52;86]	76 [65;91]	76 [62;88]	57 [40;78]	<0.001	644
Emphysema:					<0.001	611
none	559 (91%)	193 (99%)	198 (93%)	168 (82%)		
minimal/mild	33 (5%)	1 (1%)	10 (5%)	22 (11%)		
moderate	15 (2%)	0 (0%)	4 (2%)	11 (5%)		
severe	4 (1%)	0 (0%)	0 (0%)	4 (2%)		
Fibrosis:					<0.001	613
none	585 (95%)	193 (98%)	208 (98%)	184 (90%)		
minimal/mild	26 (4%)	3 (2%)	2 (1%)	21 (10%)		
moderate	1 (0%)	0 (0%)	1 (0%)	0 (0%)		
severe	1 (0%)	0 (0%)	1 (0%)	0 (0%)		
Smoking history: past/current	435 (50.8%)	107 (36.9%)	153 (53.5%)	175 (62.3%)	<0.001	857
NTproBNP [ng/l]	926 [215;2637]	345 [122;1640]	763 [158;1356]	1996 [501;3706]	<0.001	276
BNP [ng/l]	195 [72.4;432]	117 [30.0;394]	181 [85.1;398]	236 [112;481]	0.005	271
Uric acid [mmol/l]	0.41 [0.31;0.52]	0.37 [0.26;0.46]	0.41 [0.30;0.50]	0.48 [0.36;0.56]	<0.001	358
CRP [mg/l]	4.30 [2.00;8.50]	4.00 [2.00;7.00]	4.15 [2.00;8.50]	5.00 [2.50;9.10]	0.151	639
Hb [g/l]	151 [138;165]	152 [138;164]	154 [141;166]	149 [133;163]	0.007	847
WBC [x10e9/l]	8 [7;10]	8 [6;10]	8 [7;10]	8 [7;10]	0.73	839
Platelets [x10e9/l]	224 [182;272]	231 [186;280]	219 [181;261]	221 [179;272]	0.072	836
Sodium [mmol/l]	140 [138;141]	140 [138;141]	139 [138;141]	140 [137;141]	0.728	835
Potassium [mmol/l]	4.20 [3.90;4.50]	4.20 [3.98;4.40]	4.20 [3.90;4.40]	4.30 [4.00;4.50]	0.03	830
Urea [mmol/l]	5.70 [4.40;7.60]	4.80 [3.88;5.81]	5.50 [4.30;6.70]	7.60 [5.90;10.1]	<0.001	830
Creatinine [µmol/l]	85.5 [70.0;102]	79.0 [68.0;93.5]	82.0 [69.0;96.0]	96.0 [80.0;121]	<0.001	832
COPD	65 (6%)	2 (1%)	22 (6%)	41 (12%)	<0.001	1112
Asthma	78 (7%)	32 (8%)	32 (9%)	14 (4%)	0.023	1112
OSA	57 (5%)	5 (1%)	21 (6%)	31 (9%)	<0.001	1112
CAD	44 (4%)	0 (0%)	9 (2%)	35 (10%)	<0.001	1112
CVA	17 (2%)	2 (1%)	6 (2%)	9 (3%)	0.084	1112
PAD	5 (0%)	0 (0%)	0 (0%)	5 (1%)	0.003	1112
HTN	264 (24%)	19 (5%)	80 (21%)	165 (46%)	<0.001	1112
DM type 1	19 (2%)	7 (2%)	6 (2%)	6 (2%)	0.967	1112
DM type 2	137 (12%)	5 (1%)	37 (10%)	95 (27%)	<0.001	1112
Hypothyroidism	135 (12%)	38 (10%)	49 (13%)	48 (14%)	0.274	1112
Systemic lupus erythormatosus	1 (0%)	0 (0%)	0 (0%)	1 (0%)	0.319	1112
Systemic sclerosis	1 (0%)	0 (0%)	0 (0%)	1 (0%)	0.319	1112
Ankylosing spondylitis	1 (0.09%)	0 (0.00%)	0 (0.00%)	1 (0.28%)	0.319	1112
Sjogren syndrome	5 (0%)	2 (1%)	1 (0%)	2 (1%)	0.871	1112

Table S10. Summary of imaging analysis. IPAH - idiopathic pulmonary arterial hypertension, HPAH - hereditary pulmonary arterial hypertension, PVOD - pulmonary veno-occlusive disease, PCH - Pulmonary capillary haemangiomatosis, GGO - ground glass opacities, BA - bronchial artery, C - central, U - upper, Z - zonal, D - diffuse. Intra-rater reliability: GGO centrilobular pattern severity weighted Cohen's Kappa=0.679, p-value <0.001; GGO distribution unweighted Cohen's Kappa = 1, p-value 0.046; Severity of GGO non-specific pattern - no positive findings; Pulmonary arteriovenous malformations - no positive findings; largest BA size - no positive findings; Mediastinal venous collaterals: unweighted Cohen's Kappa = 1, p-value <0.001; Intralobular septal thickening weighted Cohen's Kappa = 1, p-value <0.001; Mediastinal lymphadenopathy unweighted Cohen's Kappa=0.83, p-value <0.001; Mediastinal lymphadenopathy size [mm] intraclass correlation coefficient (ICC) 0.717, p-value 0.088; Emphysema - not enough positive findings, Bronchial wall thickening - not enough positive findings, Fibrosis - no positive findings; Pleural effusion weighted Cohen's Kappa 0.826, p-value <0.001; Air trapping weighted Cohen's Kappa 0.845, p-value <0.001; Subpleural scarring - not enough positive findings.

	[ALL] N=269	*BMPR2* N=44	*EIF2AK4* N=6	*EIF2AK4* bial. N=7	KDR missense N=5	*KDR* PTV N=4	no mutation N=185	other mutations N=18	p.overall	N
Sex: female	179 (66.5%)	28 (63.6%)	5 (83.3%)	3 (42.9%)	3 (60.0%)	2 (50.0%)	129 (69.7%)	9 (50.0%)	0.341	269
Age of diagnosis	50.2 (17.0)	44.2 (13.7)	51.0 (16.7)	28.5 (10.8)	36.6 (13.1)	65.2 (4.3)	53.1 (17.2)	44.4 (14.1)	<0.001	268
Diagnosis verified:									.	269
HPAH	20 (7.4%)	13 (29.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	4 (2.2%)	3 (16.7%)		
IPAH	237 (88.1%)	31 (70.5%)	5 (83.3%)	4 (57.1%)	5 (100.0%)	4 (100.0%)	173 (93.5%)	15 (83.3%)		
PCH	1 (0.4%)	0 (0.0%)	0 (0.0%)	1 (14.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)		
PVOD	11 (4.1%)	0 (0.0%)	1 (16.7%)	2 (28.6%)	0 (0.0%)	0 (0.0%)	8 (4.3%)	0 (0.0%)		
Severity of GGO centrilobular parttern:									.	269
Nil	167 (62.1%)	22 (50.0%)	4 (66.7%)	2 (28.6%)	4 (80.0%)	2 (50.0%)	121 (65.4%)	12 (66.7%)		
Trace	29 (10.8%)	3 (6.8%)	0 (0.0%)	1 (14.3%)	0 (0.0%)	0 (0.0%)	23 (12.4%)	2 (11.1%)		
Mild	29 (10.8%)	9 (20.5%)	0 (0.0%)	1 (14.3%)	0 (0.0%)	2 (50.0%)	16 (8.6%)	1 (5.6%)		
Moderate	24 (8.9%)	3 (6.8%)	0 (0.0%)	1 (14.3%)	1 (20.0%)	0 (0.0%)	17 (9.2%)	2 (11.1%)		
Severe	20 (7.4%)	7 (15.9%)	2 (33.3%)	2 (28.6%)	0 (0.0%)	0 (0.0%)	8 (4.3%)	1 (5.6%)		
Severity of GGO non-specific pattern:									.	269
Nil	240 (89.2%)	42 (95.5%)	3 (50.0%)	5 (71.4%)	5 (100.0%)	2 (50.0%)	167 (90.3%)	16 (88.9%)		
Trace	10 (3.7%)	1 (2.3%)	2 (33.3%)	0 (0.0%)	0 (0.0%)	1 (25.0%)	6 (3.2%)	0 (0.0%)		
Mild	11 (4.1%)	1 (2.3%)	0 (0.0%)	1 (14.3%)	0 (0.0%)	1 (25.0%)	7 (3.8%)	1 (5.6%)		
Moderate	7 (2.6%)	0 (0.0%)	1 (16.7%)	1 (14.3%)	0 (0.0%)	0 (0.0%)	4 (2.2%)	1 (5.6%)		
Severe	1 (0.4%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.5%)	0 (0.0%)		
Distribution of GGO:									0.108	122
C	11 (9.0%)	0 (0.0%)	1 (20.0%)	2 (40.0%)	0 (0.0%)	2 (66.7%)	6 (7.7%)	0 (0.0%)		
D	81 (66.4%)	19 (90.5%)	3 (60.0%)	3 (60.0%)	2 (100.0%)	1 (33.3%)	48 (61.5%)	5 (62.5%)		
U	14 (11.5%)	1 (4.8%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	12 (15.4%)	1 (12.5%)		
Z	16 (13.1%)	1 (4.8%)	1 (20.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	12 (15.4%)	2 (25.0%)		
Pulmonary arteriovenous malformations: Yes	4 (1.6%)	3 (7.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.6%)	0 (0.0%)	0.152	246
Largest BA size [mm]	3.0 (0.6)	3.4 (0.5)	3.0 (.)	2.5 (0.2)	. (.)	. (.)	2.5 (0.4)	4.0 (.)	0.037	12
Mediastinal venous collaterals: Yes	206 (94.1%)	37 (100.0%)	1 (100.0%)	0 (.%)	5 (100.0%)	3 (100.0%)	147 (91.9%)	13 (100.0%)	0.379	219
Intralobular septal thickening:									0.111	243
Nil	216 (88.9%)	37 (92.5%)	0 (0.0%)	0 (.%)	4 (80.0%)	4 (100.0%)	155 (88.6%)	16 (88.9%)		
Trace	15 (6.2%)	2 (5.0%)	0 (0.0%)	0 (.%)	1 (20.0%)	0 (0.0%)	12 (6.9%)	0 (0.0%)		
Mild	9 (3.7%)	1 (2.5%)	0 (0.0%)	0 (.%)	0 (0.0%)	0 (0.0%)	7 (4.0%)	1 (5.6%)		
Moderate	3 (1.2%)	0 (0.0%)	1 (100.0%)	0 (.%)	0 (0.0%)	0 (0.0%)	1 (0.6%)	1 (5.6%)		
Mediastinal lymphoaenopathy: Yes	218 (81.3%)	40 (90.9%)	3 (50.0%)	3 (42.9%)	2 (50.0%)	1 (25.0%)	151 (81.6%)	18 (100.0%)	<0.001	268
Mediastinal lymphoadenopathy [mm]	14.9 (4.0)	12.5 (1.3)	17.0 (3.6)	14.8 (2.4)	15.5 (0.7)	11.0 (0.0)	15.3 (4.4)	. (.)	0.354	50
Emphysema:									0.813	269
Nil	235 (87.4%)	38 (86.4%)	5 (83.3%)	7 (100.0%)	4 (80.0%)	3 (75.0%)	161 (87.0%)	17 (94.4%)		
Trace	20 (7.4%)	4 (9.1%)	1 (16.7%)	0 (0.0%)	1 (20.0%)	0 (0.0%)	13 (7.0%)	1 (5.6%)		
Mild	8 (3.0%)	1 (2.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (25.0%)	6 (3.2%)	0 (0.0%)		
Moderate	6 (2.2%)	1 (2.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	5 (2.7%)	0 (0.0%)		
Bronchial wall thickening:									0.057	243
Nil	211 (86.8%)	33 (82.5%)	0 (0.0%)	0 (.%)	5 (100.0%)	3 (75.0%)	157 (89.7%)	13 (72.2%)		
Trace	19 (7.8%)	5 (12.5%)	1 (100.0%)	0 (.%)	0 (0.0%)	0 (0.0%)	11 (6.3%)	2 (11.1%)		
Mild	11 (4.5%)	2 (5.0%)	0 (0.0%)	0 (.%)	0 (0.0%)	1 (25.0%)	6 (3.4%)	2 (11.1%)		
Moderate	2 (0.8%)	0 (0.0%)	0 (0.0%)	0 (.%)	0 (0.0%)	0 (0.0%)	1 (0.6%)	1 (5.6%)		
Fibrosis:									0.059	269
Nil	257 (95.5%)	43 (97.7%)	6 (100.0%)	7 (100.0%)	5 (100.0%)	2 (50.0%)	177 (95.7%)	17 (94.4%)		
Trace	5 (1.9%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	5 (2.7%)	0 (0.0%)		
Mild	6 (2.2%)	1 (2.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (50.0%)	3 (1.6%)	0 (0.0%)		
Moderate	1 (0.4%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (5.6%)		
Pleural effusion:									.	269
Nil	242 (90.0%)	41 (93.2%)	5 (83.3%)	7 (100.0%)	4 (80.0%)	3 (75.0%)	167 (90.3%)	15 (83.3%)		
Trace	11 (4.1%)	1 (2.3%)	1 (16.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	6 (3.2%)	3 (16.7%)		
Mild	7 (2.6%)	2 (4.5%)	0 (0.0%)	0 (0.0%)	1 (20.0%)	0 (0.0%)	4 (2.2%)	0 (0.0%)		
Moderate	7 (2.6%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (25.0%)	6 (3.2%)	0 (0.0%)		
Severe	2 (0.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (1.1%)	0 (0.0%)		
Air trapping:									.	269
Nil	216 (80.3%)	40 (90.9%)	5 (83.3%)	7 (100.0%)	4 (80.0%)	1 (25.0%)	146 (78.9%)	13 (72.2%)		
Trace	26 (9.7%)	3 (6.8%)	1 (16.7%)	0 (0.0%)	0 (0.0%)	1 (25.0%)	17 (9.2%)	4 (22.2%)		
Mild	18 (6.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (50.0%)	15 (8.1%)	1 (5.6%)		
Moderate	4 (1.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	4 (2.2%)	0 (0.0%)		
Severe	5 (1.9%)	1 (2.3%)	0 (0.0%)	0 (0.0%)	1 (20.0%)	0 (0.0%)	3 (1.6%)	0 (0.0%)		
Subpleural scarring:									0.736	243
Nil	237 (97.5%)	39 (97.5%)	1 (100.0%)	0 (.%)	5 (100.0%)	4 (100.0%)	170 (97.1%)	18 (100.0%)		
Trace	2 (0.8%)	1 (2.5%)	0 (0.0%)	0 (.%)	0 (0.0%)	0 (0.0%)	1 (0.6%)	0 (0.0%)		
Mild	4 (1.6%)	0 (0.0%)	0 (0.0%)	0 (.%)	0 (0.0%)	0 (0.0%)	4 (2.3%)	0 (0.0%)		

Table S11. Clinical characteristics of IPAH patients who harbour protein-truncating variants in *BMPR2*, *EIF2AK4*, *KDR* and *IDH3G*. BMI - body mass index, WHO FC - World Health Organisation functional class, 6MWD - 6-minute walk distance, SpO_2 - arterial oxygen saturation, mRAP - mean right atrial pressure, mPAP - mean pulmonary artery pressure, mPAWP - mean pulmonary artery wedge pressure, CO - cardiac output, PVR - pulmonary vascular resistance, NO - nitric oxide challenge, FEV_1 - forced expiratory volume in 1 second, FVC - forced vital capacity, KCO - transfer factor coefficient for carbon monoxide, COPD - chronic obstructive pulmonary disease, OSA - obstructive sleep apnea, CAD - coronary artery disease, HTN - systemic hypertension, CKD - chronic kidney disease, Hb - haemoglobin, WBC - white blood cells, TSH - thyroid-stimulating hormone. Comorbidities are reported as the number and percentage of cases possessing a disease entity.

	BMPR2 N=162	Biallelic *EIF2AK4* N=14	*KDR* PTV N=4	*IDH3G* N=5	No mutation N=818	p.overall	N
Age [years]	39 [32;51]	31 [23;42]	64 [62;68]	34 [27;51]	52 [38;66]	<0.001	994
Sex: female	107 (66%)	7 (50%)	2 (50%)	5 (100%)	571 (70%)	0.146	998
BMI [kg/m^2]	27 [23;32]	24 [20;27]	26 [26;30]	24 [21;24]	27 [23;32]	0.017	909
WHO FC						0.067	965
I	2 (1.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	16 (2.0%)		
II	32 (19.9%)	2 (14.3%)	1 (25.0%)	1 (20.0%)	153 (19.6%)		
III	96 (59.6%)	9 (64.3%)	3 (75.0%)	1 (20.0%)	522 (66.8%)		
IV	31 (19.3%)	3 (21.4%)	0 (0.0%)	3 (60.0%)	90 (11.5%)		
6MWD [m]	355 [288;421]	302 [210;466]	301 [240;362]	414 [382;414]	335 [218;412]	0.314	625
SpO_2 pre [%]	96 [94;98]	92 [90;96]	97 [96;97]	96 [96;98]	95 [93;97]	0.003	801
SpO_2 post [%]	94 [89;97]	83 [76;86]	86 [86;88]	96 [96;98]	90 [84;95]	<0.001	740
mRAP [mmHg]	10 [6;14]	8 [6;10]	6 [4;8]	12 [8;14]	8 [5;12]	0.019	882
mPAP [mmHg]	57 [52;68]	52 [44;59]	50 [43;58]	58 [50;62]	52 [42;61]	<0.001	946
mPAWP [mmHg]	10 [7;12]	11 [8;12]	10 [8;13]	10 [8;12]	9 [7;12]	0.902	838
CO [L/min]	3.3 [2.7;4.0]	4.5 [3.0;4.9]	4.9 [4.3;5.4]	3.3 [3.0;3.5]	4.0 [3.2;5.1]	<0.001	903
PVR [WU]	14.4 [10.8;20.3]	9.56 [8.16;11.1]	7.93 [6.23;10.9]	15.4 [12.9;17.5]	10.3 [7.14;13.9]	<0.001	806
Acute NO challenge: vasoresponder	1 (1.28%)	0 (0.00%)	1 (25.0%)	0 (0.00%)	52 (17.3%)	<0.001	392
FEV_1 [% pred.]	91 [79;100]	93 [84;100]	86 [79;96]	95 [87;99]	84 [71;95]	<0.001	764
FVC [% pred.]	100 (17)	101 (16)	93 (17)	93 (12)	93 (20)	0.003	748
FEV_1/FVC ratio	0.77 [0.73;0.82]	0.79 [0.69;0.81]	0.78 [0.76;0.79]	0.82 [0.78;0.84]	0.75 [0.68;0.81]	0.021	681
KCO [%pred.]	83 [74;96]	33 [30;35]	46 [46;48]	73 [71;73]	68 [48;83]	<0.001	580
Smoking history: yes	53 (40.8%)	4 (30.8%)	1 (25.0%)	1 (20.0%)	330 (53.4%)	0.012	770
COPD	6 (3.70%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	49 (5.99%)	0.678	1003
Asthma	20 (12.3%)	4 (28.6%)	0 (0.00%)	0 (0.00%)	47 (5.75%)	0.003	1003
OSA	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	55 (6.72%)	0.001	1003
Pulmonary fibrosis	0 (0.00%)	0 (0.00%)	2 (50.0%)	0 (0.00%)	13 (1.59%)	0.002	1003
CAD	3 (1.85%)	1 (7.14%)	1 (25.0%)	0 (0.00%)	32 (3.91%)	0.115	1003
HTN	27 (16.7%)	0 (0.00%)	2 (50.0%)	0 (0.00%)	210 (25.7%)	0.005	1003
CKD	4 (2.47%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	41 (5.01%)	0.568	1003
Hypothyroidism	14 (8.64%)	1 (7.14%)	1 (25.0%)	0 (0.00%)	108 (13.2%)	0.338	1003
Hb [g/l]	162 [152;173]	165 [154;179]	148 [135;152]	151 [148;164]	149 [135;161]	<0.001	760
WBC [x10e9/l]	8.74 [7.30;10.8]	7.43 [6.50;10.8]	8.80 [8.23;9.55]	7.30 [6.60;7.40]	8.10 [6.70;9.70]	0.03	753
Platelets [x10e9/l]	210 [174;251]	219 [206;234]	216 [188;251]	208 [160;211]	228 [181;276]	0.272	749
Creatinine [umol/l]	93.0 [77.2;102]	79.0 [72.2;95.0]	67.0 [66.5;96.5]	79.0 [73.0;85.0]	84.0 [70.0;103]	0.25	745
TSH [mU/l]	2.37 [1.67;3.65]	2.08 [1.09;3.69]	1.76 [1.72;1.84]	1.44 [1.43;2.02]	2.00 [1.10;3.16]	0.038	588

Abbreviation	explanation
WGS ID	whole-genome sequencing identifier
WES ID	whole-exome sequencing identifier
HGVSc	the HGVS coding sequence name
HGVSp	the HGVS protein sequence name
gnomAD	Genome Aggregation Database
CADD_PHRED_v1.3	Combined Annotation Dependent Depletion
CADD_PHRED_v1.4	Combined Annotation Dependent Depletion
SIFT	Sorting Intolerant From Tolerant prediction score
PolyPhen	Polymorphism Phenotyping v2 score
GerpN	conservation score of each nucleotide in multi-species alignment
KDR	Kinase Insert Domain Receptor
IDH3G	Isocitrate Dehydrogenase (NAD(+)) 3 Non-Catalytic Subunit Gamma
BMPR2	Bone Morphogenetic Protein Type 2 Receptor
EIF2AK4	Eukaryotic Translation Initiation Factor 2 Alpha Kinase 4
ATP13A3	ATPase 13A3
SOX17	SRY-box 17
AQP1	Aquaporin 1
ENG	Endoglin
ACVRL1	Activin-Like Kinase 1
CAV1	Caveolin-1
SMAD9	SMAD family member 9
SMAD1	SMAD family member 1
GDF2	Growth Differentiation Factor 2
TBX4	T-Box Transcription Factor 4
KCNK3	Potassium Two Pore Domain Channel Subfamily K Member 3
SMAD4	SMAD family member 4
shared	indicates if a variant appears in other cohort individuals, i.e. PAH(2), BPD(1) means 2 PAH cases and 1 BPD case harbour this variant
REVEL	Rare Exome Variant Ensemble Learner
NIHRBR-RD	The National Institute for Health Research BioResource - Rare Diseases study (NIHRBR-RD),
SvO_2 [%]	Mixed venous oxygen saturation
WHO FC	World Health Organisation functional class
6MWD [m]	six minute walking distance
SpO_2 [%]	peripheral capillary oxygen saturation
FEV_1 [% pred.]	Forced Expiratory Volume in one second
FVC [% pred.]	Forced Vital Capacity
TLC [% pred.]	Total Lung Capacity
KCO [% pred.]	Transfer Coefficient of Carbon Monoxide
mRAP [mmHg]	mean Right Atrial Pressure
mPAP [mmHg]	mean Pulmonary Artery Pressure
PAWP [mmHg]	Pulmonary Artery Wedge Pressure
CO [L/min]	cardiac output
CI [L/min/m^2]	cardiac index
PVR [WU]	pulmonary vascular resistance
ILD	interstitial lung disease
COPD	chronic obstructive pulmonary disease
OSA	obstructive sleep apnoea

DM	diabetes mellitus
HTN	systemic hypertension
CKD	chronic kidney disease
CAD	coronary artery disease
CVA	Celebro-Vascular Accident
PAD	Peripheral Artery Disease
GGO	Ground Glass Opacities
BA	Bronchial Artery
Hb [g/l]	Haemoglobin
NTproBNP [ng/l]	N-terminal pro B-Type Natriuretic Peptide
BNP [ng/l]	B-Type Natriuretic Peptide
CRP [mg/l]	C reactive protein
WBC [x10e9/l]	White Blood Cell Count
TSH [mU/l]	Thryroid Stymulating Hormon
PH	Pulmonary Hypertension
PAH	Pulmonary Arterial Hypertension
I/HPAH	Idiopathic/Hereditary Pulmonary Arterial Hypertnsion
PVOD/PCH	Pulmonary veno-occlusive disease/ Pulmonary capillary hemangiomatosis
APAH	Associated Pulmonary Arterial Hypertension
APAH: CHD-PAH	PAH associated with congenital heart disease
APAH: CTD-PAH	PAH associated with connective tissue disease
APAH: PPH-PAH/ should be	PAH associated with portopulmonary hypertension
APAH: HIV-PAH	PAH associated with HIV
PH-LHD	pulmonary hypertension associated with left heart disease
PH-LD	pulmonary hypertension associated with lung disease
CTEPH	Chronic thromboembolic pulmonary hypertension
PH-multifactorial	Multifactorial pulmonary hypertension
GEL	Genomics England Ltd
BPD	Bleeding, Thrombotic and Platelet Disorders
PID	Primary Immune Disorders
CNTRL	Processed Controls
IRD	Inherited Retinal Disorders
NDD	Neurological and Developmental Disorders
EDS	Ehlers Danlos Syndrome
HCM	Hypertrophic Cardiomyopathy
PMG	Primary Membranoproliferative Glomerulonephritis
SRNS	Steroid Resistant Nephrotic Syndrome
CSVD	Cerebral Small Vessel Disease
NPD	Neuropathic Pain Disorder
ICP	Intrahepatic Cholestasis of Pregnancy
LHON	Leber Hereditary Optic Neuropathy
MPMT	Multiple Primary Tumours
SMD	Stem Cell & Myeloid Disorders
PAH	Pulmonary arterial hypertension
UKBio	UK BioBank
FHx	family history
CTPA	Computerised Tomography Pulmonary Angiogram
HRCT	High-Resolution Computerised Tomography
BMI	Body Mass Index

NO	nitric oxide
BeviMed	Bayesian Evaluation of Variant Involvement in Mendelian Disease
PTV	Protein Truncating Variants
PP	Posterior Probability
VEGFR2	vascular endotheial growth factor receptor 2
eCRF	electronic Clinical Case Report Form
GRCh37	Genome Reference Consortium human genome build 37
PMAF	The probability that the minor allele count is at least the observed minor allele count
ICC	Intraclass Correlation Coefficient
CT	Computerised Tomography
RHC	Right Heart Cathetetherisation
ASD	Atrial Septal Defect
SU5416	sugen
PDH	Pyruvate dehydrogenase
IDH	isocitrate dehydrogenase
BHF	British Heart Foundation
SNV	Single Nucleotide Variants
MAF	Minor Allele Frequency