Genetic determinants of risk and survival in pulmonary arterial hypertension

Authors:

Christopher J. Rhodes, Ph.D.* [1], Ken Batai, Ph.D.* [2], Marta Bleda, Ph.D.* [3], Matthias Haimel, B.Sc.* [3,4,5], Laura Southgate, Ph.D.* [6,7], Marine Germain, Ph.D.* [8], Michael W. Pauciulo, B.S., M.B.A.* [9], Charaka Hadinnapola, M.B. B.Chir. [3], Jurjan Aman, M.D., Ph.D. [1], Barbara Girerd, Ph.D. [10], Amit Arora, MD MPH [2], Jo Knight, Ph.D. [11], Ken B. Hanscombe, Ph.D. [7], Jason H. Karnes, PharmD, PhD [12], Marika Kaakinen, Ph.D. [1], Henning Gall, M.D., Ph.D. [13], Anna Ulrich, MSc [1], Lars Harbaum, M.D. [1], Inês Cebola, Ph.D. [14], Jorge Ferrer, Ph.D. [14], The NIHR BioResource – Rare Diseases Consortium, UK PAH Cohort Study Consortium and the US PAH Biobank Consortium, Ferhaan Ahmad, M.D., Ph.D. [15], Philippe Amouyel, M.D, Ph.D. [16], Stephen L. Archer, M.D. [17], Rahul Argula, M.D. [18], Eric D. Austin, M.D. [19], David Badesch, M.D. [20], Sahil Bakshi, DO [21], Christopher F. Barnett, M.D. [22], Raymond Benza, M.D. [23], Nitin Bhatt, M.D. [24], Harm J. Bogaard, M.D., Ph.D. [25], Charles D. Burger, M.D. [26], Murali M. Chakinala, M.D. [27], Colin Church, Ph.D. [28], John G. Coghlan, M.D. [29], Robin Condliffe, M.D. [30], Paul A. Corris, M.B.B.S. [31], Cesare Danesino, M.D. [32,33], Stéphanie Debette, M.D, Ph.D. [34], C. Gregory Elliott, M.D. [35], Jean Elwing, M.D. [36], Melanie Eyries, Ph.D. [8], Terry Fortin, M.D. [37], Andre Franke, Ph.D. [38], Robert P. Frantz, M.D. [39], Adaani Frost, M.D. [40], Joe G. N. Garcia, M.D. [41], Stefano Ghio, M.D. [33], Hossein-Ardeschir Ghofrani, M.D. [13,1], J. Simon R. Gibbs, M.D. [42], John B. Harley, M.D., Ph.D. [43,79], Hua He, Ph.D. [9], Nicholas S. Hill, M.D. [44], Russel Hirsch, M.D. [45], Arjan C. Houweling, M.D., Ph.D. [25], Luke S. Howard, M.D., Ph.D. [42], Dunbar Ivy, M.D. [46], David G. Kiely, M.D. [30], James Klinger, M.D. [47], Gabor Kovacs, M.D. [48,49], Tim Lahm, M.D. [50], Matthias Laudes, M.D. [51], Katie Lutz, B.S. [9], Rajiv D. Machado, Ph.D. [52], Robert V. MacKenzie Ross, M.B. B.Chir. [53], Keith Marsolo, Ph.D. [54], Lisa J. Martin, Ph.D. [9], Shahin Moledina, M.B. B.Chir. [55], David Montani, M.D., Ph.D. [10], Steven D. Nathan, M.D. [56], Michael Newnham, M.B.B.S. [3], Andrea Olschewski, M.D. [48], Horst Olschewski, M.D. [48,49], Ronald J. Oudiz, M.D. [57], Willem H. Ouwehand, M.D., Ph.D. [4,5], Andrew J. Peacock, M.D. [28], Joanna Pepke-Zaba, Ph.D. [58], Zia Rehman, M.D. [59], Ivan M. Robbins, M.D. [60], Dan M. Roden, M.D. [61,62], Erika B. Rosenzweig, M.D. [63], Ghulam Saydain, M.D. [64], Laura Scelsi, M.D. [33], Robert Schilz, M.D. [65], Werner Seeger, M.D. [13], Christian M. Shaffer, M.Sc. [61], Robert W. Simms, M.D. [66], Marc Simon, M.D. [67], Olivier Sitbon, M.D., Ph.D. [10], Jay Suntharalingam, M.D. [53], Emilia Swietlik, M.D. [3], Haiyang Tang, Ph.D. [41], Alexander Y. Tchourbanov, Ph.D. [68], Thenappan Thenappan, M.D. [69], Fernando Torres, M.D. [70], Mark R. Toshner, M.D. [3], Carmen M. Treacy, B.Sc. [3,58], Anton Vonk Noordegraaf, M.D. [25], Quinten Waisfisz, Ph.D. [25], Anna K. Walsworth, B.S. [9], Robert E Walter, M.D. [71], John Wharton, Ph.D. [1], R. James White, M.D., Ph.D. [72], Jeffrey Wilt, M.D. [73], Stephen J. Wort, Ph.D. [74,7], Delphine Yung, M.D. [75], Allan Lawrie, Ph.D. [76], Marc Humbert, M.D., Ph.D. [10], Florent Soubrier, M.D., Ph.D. [8], David-Alexandre Trégouët, Ph.D. [8], **Inga Prokopenko, Ph.D.# [1], Richard Kittles, Ph.D.# [77], Stefan Gräf, Ph.D.# [3,4,5], William C. Nichols, Ph.D.# [9], Richard C. Trembath, F.R.C.P.# [7,78], Ankit A. Desai, M.D.#$ [41], Nicholas W. Morrell, M.D.#$ [3,5], Martin R. Wilkins, M.D.#$ [1]**

*** these authors contributed equally to this work, # these authors jointly supervised this work**
$ corresponding authors

Corresponding authors contact details:
Ankit A. Desai, University of Arizona, Tucson, AZ, United States: adesai@shc.arizona.edu;
Nicholas W. Morrell, University of Cambridge, Cambridge, United Kingdom: nwm23@cam.ac.uk;
Martin R. Wilkins, Imperial College London, London, United Kingdom: m.wilkins@imperial.ac.uk.

Affiliations

[1] Centre for Pharmacology & Therapeutics, Department of Medicine, Hammersmith Campus, Imperial College London, London, United Kingdom;
[2] Division of Urology, Department of Surgery, The University of Arizona College of Medicine, Tucson, AZ, United States;
[3] Department of Medicine, University of Cambridge, Cambridge, United Kingdom;
[4] Department of Haematology, University of Cambridge, Cambridge, United Kingdom;
[5] NIHR BioResource - Rare Diseases, Cambridge, United Kingdom;
[6] Molecular and Clinical Sciences Research Institute, St George's University of London, London, United Kingdom;
[7] Division of Genetics and Molecular Medicine, King's College London, London, United Kingdom;
[8] Sorbonne Universités, UPMC Univ. Paris 06, Institut National pour la Santé et la Recherche Médicale (INSERM), Unité Mixte de Recherche en Santé (UMR_S) 1166, Team Genomics & Pathophysiology of Cardiovascular Diseases; ICAN Institute for Cardiometabolism and Nutrition, Paris, France;
[9] Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, United States;
[10] Université Paris-Sud, Faculté de Médecine, Université Paris-Saclay; AP-HP, Service de Pneumologie, Centre de référence de l'hypertension pulmonaire, Hôpital Bicêtre, Le Kremlin-Bicêtre; INSERM UMR_S 999, Hôpital Marie Lannelongue, Le Plessis Robinson, Paris, France;
[11] Data Science Institute, Lancaster University, Lancaster, United Kingdom;
[12] Department of Pharmacy Practice & Science, University of Arizona College of Pharmacy, Sarver Heart Center and Center for Applied Genetics and Genomic Medicine (TCAG2M), University of Arizona College of Medicine, Tucson, AZ, United States;
[13] University of Giessen and Marburg Lung Center (UGMLC), member of the German Center for Lung Research (DZL) and of the Excellence Cluster Cardio-Pulmonary System (ECCCPS), Giessen, Germany;
[14] Section of Epigenomics and Disease, Department of Medicine, Hammersmith Campus, Imperial College London, London, United Kingdom;
[15] Division of Cardiovascular Medicine, University of Iowa, Iowa City IA, United States;
[16] Univ. Lille, Inserm, CHU Lille, Institut Pasteur de Lille, U1167 - RID-AGE - Risk Factors and Molecular Determinants of Aging-related Diseases, F-59000, Lille, France;
[17] Queen's University, Kingston ON, Canada;
[18] Medical University of South Carolina, Charleston SC, United States;
[19] Vanderbilt University-Peds, Nasville TN, United States;
[20] University of Colorado Denver, Denver CO, United States;
[21] Baylor Research Institute, Plano TX, United States;
[22] Medstar Health, Washington D.C., United States;
[23] Allegheny-Singer Research Institute, Pittsburgh PA, United States;
[24] Ohio State University, Columbus OH, United States;
[25] VU University Medical Center, Amsterdam, The Netherlands;
[26] Mayo Clinic Florida, Jacksonville FL, United States;
[27] Washington University, St. Louis MO, United States;
[28] Golden Jubilee National Hospital, Glasgow, United Kingdom;
[29] Royal Free Hospital, London, United Kingdom;
[30] Sheffield Pulmonary Vascular Disease Unit, Royal Hallamshire Hospital, Sheffield, United Kingdom;
[31] University of Newcastle, Newcastle, United Kingdom;
[32] Department of Molecular Medicine, University of Pavia, Pavia, Italy;
[33] Fondazione IRCCS Policlinico San Matteo, Pavia, Italy;
[34] INSERM UMR_S 1219, Bordeaux Population Health Research Center, University of Bordeaux, France; Department of Neurology, Bordeaux University Hospital, Bordeaux, France;
[35] Department of Medicine at Intermountain Medical Center and the University of Utah, Murray UT, United States;
[36] University of Cincinnati, Cincinnati OH, United States;
[37] Duke University Medical Center, Durham NC, United States;
[38] Institute of Clinical Molecular Biology, University of Kiel, Kiel, Germany;
[39] Mayo Clinic, Rochester MN, United States;
[40] Weill Cornell Medical College and The Houston Methodist Hospital, Houston TX, United States;
[41] Department of Medicine and Arizona Health Sciences Center, University of Arizona, Tucson, AZ, United States;
[42] National Heart & Lung Institute, Imperial College London and National Pulmonary Hypertension Service, Hammersmith Hospital, London, United Kingdom;
[43] CAGE, Cincinnati Children's Hospital Medical Center, Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, United States;
[44] Tufts Medical Center, Boston MA, United States;
[45] Cincinnati Children's Hospital, Cincinnati OH, United States;
[46] University of Colorado HSC, Aurora CO, United States;
[47] Rhode Island Hospital, Providence RI, United States;
[48] Ludwig Boltzmann Institute for Lung Vascular Research, Graz, Austria;
[49] Medical University of Graz, Graz, Austria;
[50] Indiana University, Indianapolis IN, United States;
[51] Department of Internal Medicine 1, University of Kiel, Kiel, Germany;
[52] Institute of Medical and Biomedical Education, St George's University of London, London, United Kingdom;
[53] Royal United Hospitals Bath NHS Foundation Trust, Bath, United Kingdom;
[54] Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, United States;
[55] Great Ormond Street Hospital, London, United Kingdom;
[56] Inova Heart and Vascular Institute, Falls Church VA, United States;
[57] LA Biomedical Research Institute at Harbor-UCLA, Torrance CA, United States;
[58] Papworth Hospital, Papworth, United Kingdom;
[59] East Carolina University, Greenville NC, United States;
[60] Vanderbilt University Medical Center, Nashville TN, United States;
[61] Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN, United States;
[62] Departments of Biomedical Informatics and Pharmacology, Vanderbilt University School of Medicine, Nashville, TN, United States;

[63] Columbia University, New York NY, United States;
[64] Wayne State University, Detroit MI, United States;
[65] University Hospital of Cleveland, Cleveland OH, United States;
[66] Boston University School of Medicine, Boston MA, United States;
[67] University of Pittsburgh, Pittsburgh PA, United States;
[68] Department of Clinical Genomics, Ambry Genetics, Aliso Viejo, CA, United States;
[69] University of Minnesota, Minneapolis MN, United States;
[70] UT Southwestern, Dallas TX, United States;
[71] LSU Health, Shreveport LA, United States;
[72] University of Rochester Medical Center, Rochester NY, United States;
[73] Spectrum Health Hospitals, Grand Rapids MI, United States;
[74] Royal Brompton Hospital, London, United Kingdom;
[75] Seattle Children's Hospital, Seattle WA, United States;
[76] Department of Infection, Immunity & Cardiovascular Disease, University of Sheffield, Sheffield, United Kingdom;
[77] Division of Health Equities, Department of Population Sciences, City of Hope, Duarte, CA, United States;
[78] Department of Clinical Genetics, Guy's Hospital, London, United Kingdom
[79] US Department of Veterans Affairs Medical Center, Cincinnati, Ohio, USA

Main text word count: 2803

Abstract

Background

Pulmonary arterial hypertension (PAH) is a rare disorder leading to premature death. Rare genetic variants contribute to disease etiology but the contribution of common genetic variation to disease risk and outcome remains poorly characterized.

Methods

We performed two separate genome-wide association studies of PAH using data across 11,744 European-ancestry individuals (including 2,085 patients), one with genotypes from 5,895 whole genome sequences and another with genotyping array data from 5,849 further samples. Cross-validation of loci reaching genome-wide significance was sought by meta-analysis. We functionally annotated associated variants and tested associations with duration of survival.

Findings

A locus at *HLA-DPA1/DPB1* within the class II major histocompatibility (MHC) region and a second near *SOX17* were significantly associated with PAH. The *SOX17* locus contained two independent signals associated with PAH. Functional and epigenomic data indicate that the risk variants near *SOX17* alter gene regulation via an enhancer active in endothelial cells. PAH risk variants determined haplotype-specific enhancer activity and CRISPR-inhibition of the enhancer reduced *SOX17* expression. Analysis of median survival showed that PAH patients with two copies of the *HLA-DPA1/DPB1* risk variant had a two-fold difference (>16 years versus 8 years), compared to patients homozygous for the alternative allele.

Interpretation

We have found that common genetic variation at loci in *HLA-DPA1/DPB1* and an enhancer near *SOX17* are associated with PAH. Impairment of Sox17 function may be more common in PAH than suggested by rare mutations in *SOX17*. Allelic variation at *HLA-DPB1* stratifies PAH patients for survival following diagnosis, with implications for future therapeutic trial design.

Funding

UK NIHR, BHF, UK MRC, Dinosaur Trust, NIH/NHLBI, ERS, EMBO, Wellcome Trust, EU, AHA, ACClinPharm, Netherlands CVRI, Dutch Heart Foundation, Dutch Federation of UMC, Netherlands OHRD and RNAS, German DFG, German BMBF, APH Paris, Inserm, Université Paris-Sud, and French ANR.

Introduction

Pulmonary arterial hypertension (PAH) refers to an uncommon but devastating disorder characterized by obliterative pulmonary vascular remodelling, leading to a progressive increase in pulmonary vascular resistance and right heart failure. The annual mortality rate for idiopathic (IPAH) and heritable PAH (HPAH) remains around 10%, despite the use of modern therapies[1,2]. In part this reflects the limited impact of licensed treatments upon the underlying pulmonary vascular pathology, which includes vascular smooth muscle and fibroblast hyperplasia, endothelial cell proliferation and inflammation[3]. Substantial variation between patients in their response to available treatments highlights underlying and inadequately characterized heterogeneity in the etiology of PAH.

Recent gene sequencing studies have revealed rare mutations in a number of genes including bone morphogenetic protein type II receptor (*BMPR2)*, potassium channels, and most recently the transcription factor *SOX17*[4]. While influencing both the risk of developing PAH and survival, rare genetic variation is found in at most 25% of patients with PAH. In the majority of PAH patients the extent of genetic contribution, including that attributable to common variation, remains largely unknown[5,6]. Therefore, we tested for genome-wide association for PAH in large international cohorts and assessed the contribution of associated regions to patient outcomes. Given the rarity of PAH, we aggregated four cohorts across North America and Europe and used a two-stage, discovery and cross-validation by meta-analysis design to assess the strength of the results.

Methods

PAH cohorts and genotyping

PAH was defined by hemodynamic criteria according to international guidelines[2]. Unrelated individuals with IPAH and HPAH or anorexigen-associated PAH were included. Subjects with evidence of other known causes of PAH were excluded (appendix p.2-3). All enrolled individuals provided written informed consent from their respective institutions or were included as anonymous controls under the DNA databank at Vanderbilt University-BioVU (https://victr.vanderbilt.edu/pub/biovu/) opt-out policy (appendix p.2).

Four studies were used for the analyses as summarised in Figure 1: In the *UK National Institute for Health Research BioResource (NIHRBR) for Rare Diseases study*, whole-genome sequencing (WGS, Illumina, mean depth ~35X, appendix p.2) was performed in a total of 5,895 individuals of European descent, each with a rare disorder from 16 categories or their unaffected relatives and included 847 PAH cases (Tables S1 and S2, Figure S1 appendix p.24). The concept of this study was to sequence patients with rare diseases to identify genetic influences on the pathogenesis of one rare disorder using the other rare diseases as controls, assuming that distinct rare diseases are highly unlikely to share common genetic mechanisms. This assumption was tested by repeating analyses excluding each major control group (see results below and appendix p.8).

Three studies used genome-wide genotyping arrays: the *US National Biological Sample and Data Repository for Pulmonary Arterial Hypertension/PAH biobank (PAHB) study*, including 694 PAH cases and 1,560 controls ascertained for a large pharmacogenomic study at Vanderbilt University[7]; the *Pulmonary Hypertension Allele-Associated Risk* (PHAAR) study[5], including 269 PAH cases and 1,068 population-based controls; and the *British Heart Foundation Pulmonary Arterial Hypertension (BHFPAH)* study, consisting of 275 PAH cases and 1,983 population-based controls, (Table S1/appendix p.12). All genotyping studies were imputed (appendix p.12) and SNPs with good imputation quality (Rsq<=0.3) taken forward for testing. Other QC steps are detailed in Table S1/appendix p.12 and appendix p.4-5.

Association analyses

We used logistic regression to test single marker variants for genetic association with a diagnosis of PAH assuming a log-additive genetic model and adjusting for sex, read length chemistry (NIHRBR only) and for population structure using principal components analysis. Genomic inflation factor was calculated and verified to be between 1 and 1.05 for each study.

Discovery was performed in two independent sets: 1) WGS data from NIHRBR (n=5,895, including 847 PAH cases) and 2) meta-analysis of genotyping studies PAHB, PHAAR and BHFPAH (n=5,849, including 1,238 PAH cases). Cross-validation was performed and loci confirmed in a meta-analysis of all four studies using inverse-variance weighted fixed-effect meta-analyses, implemented in the GWAMA software tool[8]. We performed a conditional analysis on the lead variant in each locus to test for independent distinct signals reaching $P<5\text{x}10^{-8}$.

LDlink was used to assess linkage disequilibrium (LD) of variants in all European populations from the 1000 Genomes Project, (https://analysistools.nci.nih.gov/LDlink/; accessed 18/07/17). Credible sets of variants considered 99% likely to include the functional causal variants were calculated by summing ranked posterior probabilities (appendix p.5&8).

Annotation and functional assessment of the locus near *SOX17*

The locus near *SOX17* was assessed against publically available functional annotation datasets (including ENCODE and Blueprint). The locus was repressed using CRISPR-mediated inhibition in human pulmonary artery endothelial cells (hPAECs, PromoCell GmbH, Heidelberg, Germany) by transduction with a lentivirus containing a plasmid encoding the nuclease-deficient Cas9 (dCas9) fused to the repressor KRAB and a 20bp guide RNA (appendix p.6). Cells were harvested following blasticidin selection, and gene expression of *SOX17* as well as neighbouring *MRPL15* and *TMEM68* were assessed by quantitative PCR.

In vitro enhancer activity of the loci and variants near *SOX17* was investigated using a luciferase reporter assay. Specifically, genomic DNA (gDNA) was isolated from blood-outgrowth endothelial cells derived from a PAH patient heterozygous for the lead SNP at *SOX17* and used to clone 100bp putative enhancer regions containing the *SOX17* PAH variants. The cloned products were inserted into a luciferase reporter plasmid, which was subsequently used for transformation of stable bacteria. Picking various bacterial colonies allowed for isolation of luciferase reporter plasmids containing gDNA inserts differing only by the allele of the SNP of interest. Reporter plasmids were transfected into hPAECs by electroporation and luciferase activity was measured to quantify the enhancer function of the inserts with relevant haplotype.

Survival analysis for lead variants and HLA alleles

All-cause mortality was used as the primary endpoint in survival analyses using Kaplan-Meier estimates and Cox regression in the '*survival*' package in R[9]. Survival was calculated from diagnosis to date of death, or censoring (NIHRBR: 31/10/16, PAHB: 01/08/17, PHAAR: 27/09/17, BHFPAH: 12/10/17), with left-truncation using date of genetic consent, and patients were censored at lung/heart-and-lung transplantation. Age and sex were included as covariates to correct for their known association with prognosis[2].

Analytical HLA type inference

HLA alleles and amino acids totalling 1873 features were determined by imputation from genotyped and high-quality imputed variants in the HLA region using the SNP2HLA software and the type 1 diabetes genetics consortium reference database[10]. HLA alleles and amino acids were then tested for association with the novel lead variants or case-control status by chi-squared test with FDR correction.

Results

Identification of PAH loci

In two separate GWAS discovery analyses, one based on a large case-control cohort that had undergone whole genome sequencing and the other comprising three genotyped case-control studies (Figure 1), we identified two loci associated with PAH reaching genome-wide significance ($p<5x10^{-8}$, Table 1 and Figure S1/appendix p.22). One locus was within *HLA-DPA1/DPB1*, which encodes the Major Histocompatibility Complex (MHC) class II, DP alpha- and beta-chains. The second locus was 100-200kb upstream of *SOX17* which encodes the transcription factor SRY-related HMG box 17 (known as Sox17).

Cross-validation of PAH loci and genome-wide meta-analysis

As both the *HLA-DPA1/DPB1* and *SOX17* loci reached genome-wide significance in both discovery analyses, our cross-validation strategy simply confirmed the same alleles were more frequent in PAH in both analyses (Table 1). Next we performed genome-wide meta-analysis of all four studies, totalling 2,085 cases and 9,655 controls, which confirmed their associations with PAH (*HLA-DPA1/DPB1*, rs2856830, $p=7.65x10^{-20}$; *SOX17*, rs10103692, $p=5.13x10^{-15}$, Table 1 and Figure 2), and detected no further loci at genome-wide significance. Allele frequencies in the different control groups were similar between studies and to non-Finnish Europeans in the public database gnomAD (Table 1).

Definition of key variants and independent signals within PAH loci

To determine if there was more than one signal at each locus, we performed a conditional analysis (see Methods). This confirmed that the *HLA-DPA1/DPB1* locus contained a single signal of association, but showed that the *SOX17* locus was composed of two independent signals; signal 1 is 100-103kb upstream of *SOX17* (conditional $p_{conditional}=9.82x10^{-9}$) and signal 2 is 106-200kb upstream of *SOX17* ($p_{conditional}=4.16x10^{-11}$, Figure 2 and Figure S3/appendix p.27). To narrow the variants in these loci to those 99% likely to be causal, we performed a Bayesian credible set analysis (Table S3/appendix p.14). The *HLA-DPA1/DPB1* locus included 9 SNPs (all $p<9.1x10^{-18}$), *SOX17* signal 1 included 4 SNPs 100-103kb upstream of *SOX17* (all $p<3.3x10^{-8}$) and *SOX17* signal 2 included 31 SNPs 106-142kb upstream of *SOX17* (all $p<5.7x10^{-10}$).

Testing of published loci associated with PAH and sensitivity analyses

Previous studies have reported the association of variants near *CBLN2* and *PDE1A|DNAJC10* with PAH[5,6]. These common variant signals showed no association with PAH in the combined NIHRBR, PAHB and BHFPAH cohorts (p=0.17; and p=0.24, respectively; Table S2/appendix p.13). Sensitivity analyses excluding pathogenic *BMPR2* variant carriers, all pathogenic rare variant carriers or controls from different disease groups yielded similar results to the main analyses (appendix p.8).

Functional impact of PAH locus upstream of *SOX17*

To search for evidence of regulatory elements in relevant tissues at *SOX17* signal 1 and signal 2, we examined publically available epigenomic data (including histone modifications, Figure 3 and Figure S2/appendix p.23). This identified several putative enhancer elements active in both lung tissue and endothelial cells (Figure 3). One of these (around hg19-chr8:55.270Mb) contains a cluster of three out of four credible variants from *SOX17* signal 1 (Figure 3). Another (around hg19-chr8:55.252Mb) contains 1 credible variant from *SOX17* signal 2. Of these variants, rs10958403 in signal 1 and rs765727 in signal 2 overlap a DNAse hypersensitivity signal, which indicates accessible chromatin (allowing binding of transcription factors), detected in human pulmonary artery endothelial cells (hPAECs, Figure 3).

To study the effects of the PAH risk variants on the putative enhancers defined by the epigenomic signals, we developed reporter constructs containing 100bp of the regions containing either the risk allele or non-risk alleles at each of the 4 SNPs using genomic DNA from a patient heterozygous for both *SOX17* signals. A haplotype-specific reporter assay in hPAECs confirmed that the regions containing either rs10958403 or rs765727 exhibited enhancer activity (between 3 and 6-fold induction of luciferase/Renilla ratio, p<0.001), whereas constructs containing rs12674755 or rs12677277 had no effect compared to the empty vector control. We also observed haplotype-specific activity with the active constructs, which differed only by the alleles at PAH-associated risk variants rs10958403 or rs765727 (both p<0.05, Figure 4B).

DNA folding patterns determined by Hi-C data from lung tissue and endothelial cells (human umbilical vein endothelial cells, human microvascular endothelial cells, Figure 3A) indicate that the *SOX17* PAH locus resides in a defined topologically associated domain (TAD) in which the only gene found, and thus likely target of any regulatory elements in this region, is *SOX17*. To test this, we performed CRISPR-mediated inhibition of the *SOX17* signal 1 region in hPAECs. This resulted in selective down-regulation of *SOX17* expression but not the expression of neighbouring *MRPL15* and *TMEM68* genes, suggesting that the enhancers in this locus specifically regulate *SOX17* (Figure 4C-D and Figure S3/appendix p.25).

Associations of PAH loci with clinical outcomes

We investigated whether the *HLA-DPA1/DPB1* and *SOX17* variants influence clinical outcomes in PAH, specifically all-cause mortality. The *HLA-DPA1/DPB1* rs2856830 genotype, but not the *SOX17* locus, was strongly associated with survival (Figure 5). Median survival from diagnosis in the NIHRBR and PAHB patients with the C/C homozygous genotype was double (median[95%CI] =16.34[12.34->16.34] years) that of the T/T genotype (median[95%CI] = 8.05[5.76-11.3] years). Cox regression survival analyses showed that the rs2856830 T/T genotype conferred an increased annual risk of death in PAH of 97% (Figure 5B).

Sensitivity analyses excluding pathogenic *BMPR2* variant carriers, all pathogenic rare variant carriers and patients diagnosed in previous decades who may have been exposed to different treatment regimens gave results similar to the main analyses (appendix p.8).

We tested both loci for association with other clinical variables, including disease severity measures and comorbidities (Tables S5 and S6/appendix p.16-17). The C allele at *HLA-DPA1/DPB1* lead SNP rs2856830 was associated with younger age at diagnosis (Figure 5A), with C/C homozygotes presenting a decade earlier (Table S5/appendix p.16). The rs2856830 genotype was not associated with vasoresponder status.

PAH locus at *HLA-DPA1/DPB1*

The *HLA-DPA1/DPB1* locus included a missense variant rs1042140 in *HLA-DPB1* reaching genome-wide significance, (Table 1) in partial LD (r^2=0.45 with lead rs2856830 in Europeans). The SNP, rs1042140, determines a glutamic acid (Glu^{69}) or a lysine at amino acid residue 69. To determine specific HLA alleles associated with the lead variant, rs2856830, we imputed HLA types from the genotype data. These types are represented by digit codes, where the first 2 digits represent related groups of similar alleles (e.g. *DPB1**02), and 4 digits represent specific proteins with distinct amino acid sequences (e.g. *DPB1**02:01). We found that the PAH-enriched C allele of rs2856830 was associated with *HLA-DPB1**02:01/02:02/16:01 (all $p<1\times10^{-9}$ after FDR correction, Table 2 and Table S7/appendix p.18), which all contain the Glu^{69} residue. The most numerous *DPB1**02:01 and *DPB1**04:01 alleles were associated with survival in PAH patients (hazard ratio, HR[95%CI]=0.70[0.49-1.00] and HR[95%CI]=1.33[1.04-1.70], respectively, Table 2).

Frequency of PAH risk alleles

The risk alleles at both signals within the *SOX17* locus are common (risk allele frequencies are rs13266183-C=74% and rs9298503-C=92%, respectively), such that 59% of PAH cases were homozygous for the risk allele at both *SOX17* SNPs, compared to only 46% of controls.

The alleles at *HLA-DPB1* associated with the poorest outcomes are also common (risk allele frequency of rs2856830-T=86%), such that 69% of PAH patients had the T/T genotype associated with the poorest outcomes and 95% had at least one T allele.

Discussion

Through a meta-analysis of 11,744 individuals we have established loci at *HLA-DPA1/DPB1* and at an enhancer upstream of *SOX17* associated with PAH disease risk. Polymorphic variation at the *HLA-DPA1/DPB1* locus is strongly associated with both the age at diagnosis and prognosis in PAH. Common genetic variants in the enhancer region of *SOX17* are biologically plausible candidates for susceptibility to pulmonary vascular disease.

Both *in silico* and experimental analyses of the common variants upstream of the *SOX17* gene suggest they influence susceptibility to PAH through regulation of *SOX17* expression. We have recently reported enrichment and familial segregation in PAH of causal rare deleterious variation in *SOX17*, implicating this gene in the pathogenesis of PAH[4]. Sox17 is involved in the development of the endoderm[11-13], vascular endothelium, haematopoietic cells[14] and cardiomyocytes[15,16]. Sox17 also determines the endothelial fate of CD34+ progenitor cells de-differentiated from fibroblasts[17]. Deletion in the mouse leads to abnormal pulmonary vascular development, poor distal lung perfusion and biventricular hypertrophy[18]. Sox17 is a pro-angiogenic transcription factor and interacts with well-established endothelial molecular mediators[19,20]; reduction of Sox17 in endothelial cells through Notch activation (itself associated with BMPR2 signalling[21]) restricts angiogenesis[19]. Conversely, vascular endothelial growth factor (VEGF) upregulates Sox17 and, as part of a positive feedback loop, Sox17 promotes expression of VEGF receptor 2 (VEGFR2)[20]. This is relevant as inhibition of VEGFR2 results in severe pulmonary hypertension in established preclinical models[22].

We report *HLA-DPB1* alleles associate with PAH and have a pivotal role in determining disease progression. The beneficial effect of the C/C genotype at rs2856830 on survival is greater than that of any current PAH-specific drug treatment[23], with the exception of calcium channel blockers which are effective in a small subgroup (less than 10%) of PAH patients classified as “vasoresponders”[2]. Patients with the C allele at rs2856830 presented at a significantly younger age, but the association of the *HLA-DPB1* SNP with survival remains significant after correction for both age and sex. Clinical HLA typing or rs2856830 genotyping could improve risk stratification both in clinical practice and in clinical trials, where over-representation of the C/C genotype in one treatment arm could significantly impact outcomes.

The mechanism of rs2856830 involvement in PAH is likely through its association with specific *HLA-DPB1* alleles. Class II (*HLA-DRB1*, *-DQB1*, and *-DPB1*) antigen-presenting proteins play critical roles in the adaptive immune response[24,25]. The *HLA-DPB1* alleles associated with rs2856830 (*HLA-DPB1**02:01/02:02/16:01) in the current study have also previously been linked to susceptibility to hard metal lung diseases such as berylliosis[26,27]. A number of individual amino acid residues in the peptide-binding pockets of the *HLA-DPB1* molecule influence its function and T-cell recognition, either by changing peptide antigen binding or the conformation of the peptide-binding groove[28]. *HLA-DPB1**02:01/02:02/16:01 all contain a glutamate at position 69 and a valine at position 36 that reduce the risk of clinical deterioration. These same residues are essential for T-cell activation and cytokine production in berylliosis[29,30]. The potential role of this modification in antigen binding, autoimmune response and vascular damage in PAH demands further investigation.

We have shown in a rare disorder that common variation can drive significant clinical differences in presentation and outcomes. Furthermore, a common non-coding variant can regulate expression of a gene linked by rare, deleterious mutations to the same pathology. *HLA-DPB1*, and wider immune regulatory pathways, should be considered a priority for patient stratification and investigation of new treatments in PAH. *SOX17* is a key endothelial regulator and its dysfunction in PAH may be more common than heritable cases suggest.

Acknowledgements

We gratefully acknowledge the participation of patients recruited to the UK National Institute of Health Research BioResource - Rare Diseases (NIHR BR-RD) Study. We thank the NIHR BR-RD staff and co-ordination teams at the University of Cambridge, and the research nurses and coordinators at the specialist pulmonary hypertension centres involved in this study. We are also grateful to Dr Jenny Thomson and Caroline Langman for invaluable assistance in BHFPAH patient recruitment. The UK National Cohort of Idiopathic and Heritable PAH is supported by the NIHR BR-RD, the British Heart Foundation (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), NWM (RG/13/4/30107) and MRW (RG/10/16/28575), and the UK National Institute for Health Research Cambridge Biomedical Research Centre. We also gratefully acknowledge the participation of patients recruited to the National Institutes of Health/National Heart, Lung, and Blood Institute-sponsored National Biological Sample and Data Repository for PAH (aka PAH Biobank). We thank the physicians, research nurses and coordinators at the 40 pulmonary hypertension centers across the US involved in the PAH Biobank. For a listing of investigators and enrolling centers, please visit www.pahbiobank.org. Vanderbilt University Medical Center's BioVU projects are supported by numerous sources: institutional funding, private agencies, and federal grants. These include the NIH funded Shared Instrumentation Grant S10RR025141; CTSA grants UL1TR002243, UL1TR000445, and UL1RR024975. The genotyping of the VESPA samples was supported by RC2GM092618. The authors acknowledge use of BRC Core Facilities provided by the financial support from the Department of Health via the National Institute for Health Research (NIHR) comprehensive Biomedical Research Centre award to Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London and King's College Hospital NHS Foundation Trust. NWM is a British Heart Foundation Professor and National Institute of Health Research (NIHR) Senior Investigator. CH is a NIHR Rare Disease Translational Research Collaboration Clinical PhD Fellow. CJR is supported by a British Heart Foundation Intermediate Basic Science Research Fellowship (FS/15/59/31839). LH and JA are the recipients of ERS and joint ERS/EMBO Long-Term Research fellowships LTRF 2016–6884 and LTRF 201701-00072. AL is supported by a British Heart Foundation Senior Basic Science Research Fellowship (FS/13/48/30453). LS is supported by the Wellcome Trust Institutional Strategic Support Fund (204809/Z/16/Z) awarded to St. George's, University of London. IP is supported by the Wellcome Trust (WT205915), and the EU H2020 (DYNAhealth, project number 633595). Funding for the PAH Biobank is provided by NIH/NHLBI HL105333. WCN and MWP are supported by NIH NHLBI HL105333. JHK receives support from the American Heart Association (16SDG29090005) and the American College of Clinical Pharmacy Research Institute (Futures Grant). AAD receives support from NIH NHLBI R01HL136603. JF is supported by the Wellcome Trust (WT101033). JBH is supported by eMERGE U01 (NHGRI U01 HG008666). We acknowledge the support of the Imperial NIHR Clinical Research Facility and Biomedical Research Centre, 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. MRW and HAG receive funding from German Research Foundation (DFG) SFB1213, project A09. The popgen 2.0 network is supported by a grant from the German Ministry for Education and Research (01EY1103). We thank all the patients and their families who contributed to this research and the Pulmonary Hypertension Association (UK) for their support.

References

1. McGoon MD, Benza RL, Escribano-Subias P, et al. Pulmonary arterial hypertension: epidemiology and registries. J Am Coll Cardiol 2013;62:D51-9.
2. Galie N, Humbert M, Vachiery JL, et al. 2015 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension: The Joint Task Force for the Diagnosis and Treatment of Pulmonary Hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS): Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC), International Society for Heart and Lung Transplantation (ISHLT). Eur Respir J 2015;46:903-75.
3. Stacher E, Graham BB, Hunt JM, et al. Modern age pathology of pulmonary arterial hypertension. Am J Respir Crit Care Med 2012;186:261-72.
4. Graf S, Haimel M, Bleda M, et al. Identification of rare sequence variation underlying heritable pulmonary arterial hypertension. Nat Commun 2018;9:1416.
5. Germain M, Eyries M, Montani D, et al. Genome-wide association analysis identifies a susceptibility locus for pulmonary arterial hypertension. Nat Genet 2013;45:518-21.
6. Kimura M, Tamura Y, Guignabert C, et al. A genome-wide association analysis identifies PDE1A|DNAJC10 locus on chromosome 2 associated with idiopathic pulmonary arterial hypertension in a Japanese population. Oncotarget 2017;8:74917-26.
7. Bowton E, Field JR, Wang S, et al. Biobanks and electronic medical records: enabling cost-effective research. Sci Transl Med 2014;6:234cm3.
8. Magi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics 2010;11:288.
9. Therneau T. A Package for Survival Analysis in S version 2.38, https://CRAN.R-project.org/package=survival. 2015.
10. Jia X, Han B, Onengut-Gumuscu S, et al. Imputing amino acid polymorphisms in human leukocyte antigens. PLoS One 2013;8:e64683.
11. Hudson C, Clements D, Friday RV, Stott D, Woodland HR. Xsox17alpha and -beta mediate endoderm formation in Xenopus. Cell 1997;91:397-405.
12. Alexander J, Stainier DY. A molecular pathway leading to endoderm formation in zebrafish. Curr Biol 1999;9:1147-57.
13. Kanai-Azuma M, Kanai Y, Gad JM, et al. Depletion of definitive gut endoderm in Sox17-null mutant mice. Development 2002;129:2367-79.
14. Kim I, Saunders TL, Morrison SJ. Sox17 dependence distinguishes the transcriptional regulation of fetal from adult hematopoietic stem cells. Cell 2007;130:470-83.
15. Zhang C, Basta T, Klymkowsky MW. SOX7 and SOX18 are essential for cardiogenesis in Xenopus. Dev Dyn 2005;234:878-91.
16. Liu Y, Asakura M, Inoue H, et al. Sox17 is essential for the specification of cardiac mesoderm in embryonic stem cells. Proc Natl Acad Sci U S A 2007;104:3859-64.
17. Zhang L, Jambusaria A, Hong Z, et al. SOX17 Regulates Conversion of Human Fibroblasts Into Endothelial Cells and Erythroblasts by Dedifferentiation Into CD34+ Progenitor Cells. Circulation 2017;135:2505-23.
18. Lange AW, Haitchi HM, LeCras TD, et al. Sox17 is required for normal pulmonary vascular morphogenesis. Dev Biol 2014;387:109-20.
19. Lee SH, Lee S, Yang H, et al. Notch pathway targets proangiogenic regulator Sox17 to restrict angiogenesis. Circ Res 2014;115:215-26.
20. Kim K, Kim IK, Yang JM, et al. SoxF Transcription Factors Are Positive Feedback Regulators of VEGF Signaling. Circ Res 2016;119:839-52.
21. Hurst LA, Dunmore BJ, Long L, et al. TNFalpha drives pulmonary arterial hypertension by suppressing the BMP type-II receptor and altering NOTCH signalling. Nat Commun 2017;8:14079.
22. Abe K, Toba M, Alzoubi A, et al. Formation of plexiform lesions in experimental severe pulmonary arterial hypertension. Circulation 2010;121:2747-54.

23. Zheng Y, Yang T, Chen G, Hu E, Gu Q, Xiong C. Prostanoid therapy for pulmonary arterial hypertension: a meta-analysis of survival outcomes. Eur J Clin Pharmacol 2014;70:13-21.

24. Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map: expression, interaction, diversity and disease. J Hum Genet 2009;54:15-39.

25. Trowsdale J, Knight JC. Major histocompatibility complex genomics and human disease. Annu Rev Genomics Hum Genet 2013;14:301-23.

26. Richeldi L, Sorrentino R, Saltini C. HLA-DPB1 glutamate 69: a genetic marker of beryllium disease. Science 1993;262:242-4.

27. Potolicchio I, Mosconi G, Forni A, Nemery B, Seghizzi P, Sorrentino R. Susceptibility to hard metal lung disease is strongly associated with the presence of glutamate 69 in HLA-DP beta chain. Eur J Immunol 1997;27:2741-3.

28. Diaz G, Amicosante M, Jaraquemada D, et al. Functional analysis of HLA-DP polymorphism: a crucial role for DPbeta residues 9, 11, 35, 55, 56, 69 and 84-87 in T cell allorecognition and peptide binding. Int Immunol 2003;15:565-76.

29. Fontenot AP, Torres M, Marshall WH, Newman LS, Kotzin BL. Beryllium presentation to CD4+ T cells underlies disease-susceptibility HLA-DP alleles in chronic beryllium disease. Proc Natl Acad Sci U S A 2000;97:12717-22.

30. Lombardi G, Germain C, Uren J, et al. HLA-DP allele-specific T cell responses to beryllium account for DP-associated susceptibility to chronic beryllium disease. J Immunol 2001;166:3549-55.

Tables and Figures

Variant	Chromosome and position, hg19 : Effect/ Non-effect alleles	Effect allele frequency in non-Finnish Europeans in gnomAD	Effect allele frequency in NIHRBR controls	UK NIHRBR whole genome sequencing study (n=847 cases v 5,048 controls)		Effect allele frequency in genotyping controls	Meta-analysis of genotyping studies US PAHB, Paris PHAAR and London BHFPAH (n= 1,238 cases v 4,611 controls)		Meta-analysis of all cohorts (n=2085 cases, 9659 controls, effective n=6648)	
				Odds ratio (95% confidence intervals)	P-value		Odds ratio (95% confidence intervals)	P-value	Odds ratio (95% confidence intervals)	Meta-analysis P-value
Lead SNPs										
HLA-DPA1/DPB1, rs2856830	6:33041734:C/T	0.12	0.12	1.71 (1.48 - 1.96)	**4.41x10^{-14}**	0.13	1.44 (1.26 - 1.64)	5.35x10^{-8}	1.56 (1.42 - 1.71)	**7.65x10^{-20}**
SOX17, signal 1 rs13266183	8:55267612:C/T	0.73	0.73	1.44 (1.26 - 1.64)	**4.44x10^{-8}**	0.74	1.31 (1.17 - 1.46)	4.1x10^{-6}	1.36 (1.25 - 1.48)	**1.69x10^{-12}**
SOX17, signal 2 rs10103692	8:55258127:G/A	0.90	0.90	1.85 (1.47 - 2.31)	9.51x10^{-8}	0.91	1.76 (1.45 - 2.14)	**9.84x10^{-9}**	1.80 (1.55 - 2.08)	**5.13x10^{-15}**
Other SNPs										
HLA-DPB1 missense SNP, rs1042140	6:33048640:G/A	0.23	0.23	1.38 (1.22 - 1.55)	9.21x10^{-8}	0.23	1.44 (1.29 - 1.61)	**9.73x10^{-11}**	1.41 (1.30 - 1.53)	7.13x10^{-17}
SOX17, genotyping lead SNP rs28576721	8:55265980:T/C	0.91	0.92	1.55 (1.23 - 1.95)	1.57x10^{-4}	0.92	1.96 (1.57 - 2.43)	**1.54x10^{-9}**	1.75 (1.50 - 2.05)	3.07x10^{-12}

Table 1 - Novel loci associated with PAH in sequenced and genotyped cohorts. Odds ratios are for association between effect allele and PAH. gnomAD is the Genome Aggregation Database, which provides information including allele frequencies in different populations.

Associations with *HLA-DPB1* alleles and lead SNP rs2856830																						
Amino acid residues in DPB1 alleles																			Frequencies by GWAS SNP rs2856830			Sig.
Position / Allele	8	9	11	33	35	36	55	56	57	65	69	76	84	85	86	87	96	178	194	T/T	T/C	C/C	q after FDR correction
*DPB1**02:01	L	F	G	E	F	V	D	E	E	I	E	M	G	G	P	M	R	L	R	3%	44%	90%	<5x10^{-247}
*DPB1**02:02	L	F	G	E	L	V	E	A	E	I	E	M	G	G	P	M				0%	3%	7%	2.77x10^{-87}
*DPB1**16:01	L	F	G	E	F	V	D	E	E	I	E	M	D	E	A	V				0%	2%	2%	7.08x10^{-41}
*DPB1**03:01	V	Y	L	E	F	V	D	E	D	L	K	V	D	E	A	V	K	L	R	12%	6%	0%	5.50x10^{-23}
*DPB1**04:01	L	F	G	E	F	A	A	A	E	I	K	M	G	G	P	M	R	L	R	48%	26%	0%	2.40x10^{-138}
*DPB1**04:02	L	F	G	E	F	V	D	E	E	I	K	M	G	G	P	M	R	M	R	13%	6%	0%	2.08x10^{-23}
*DPB1**01:01	V	Y	G	E	Y	A	A	A	E	I	K	V	D	E	A	V	K	L	Q	7%	4%	0%	1.17x10^{-8}

Table 2 - Associations of *HLA-DPB1* alleles with the lead SNP rs2856830. Orange indicates alleles and residues depleted in PAH cases and green indicates those enriched in PAH cases.

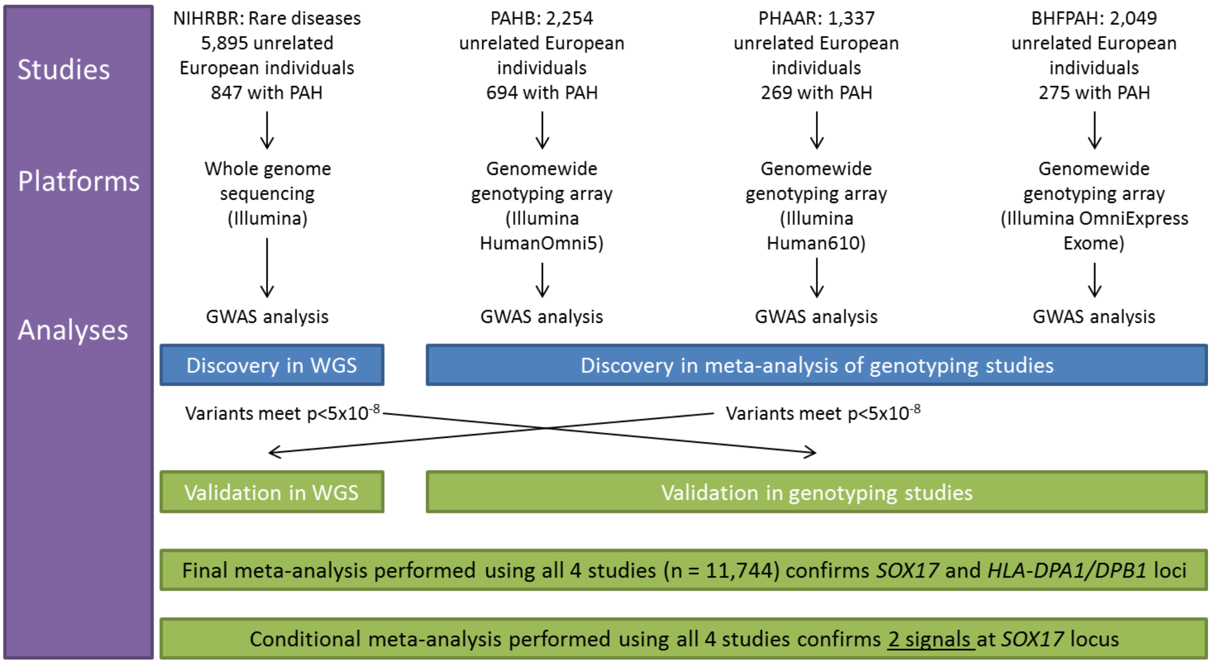

Studies
Platforms
Analyses
NIHRBR: Rare diseases 5,895 unrelated European individuals 847 with PAH
PAHB: 2,254 unrelated European individuals 694 with PAH
PHAAR: 1,337 unrelated European individuals 269 with PAH
BHFPAH: 2,049 unrelated European individuals 275 with PAH
Whole genome sequencing (Illumina)
Genomewide genotyping array (Illumina HumanOmni5)
Genomewide genotyping array (Illumina Human610)
Genomewide genotyping array (Illumina OmniExpress Exome)
GWAS analysis
GWAS analysis
GWAS analysis
GWAS analysis
Discovery in WGS
Discovery in meta-analysis of genotyping studies
Variants meet p<5x10-8
Variants meet p<5x10-8
Validation in WGS
Validation in genotyping studies
Final meta-analysis performed using all 4 studies (n = 11,744) confirms SOX17 and HLA-DPA1/DPB1 loci
Conditional meta-analysis performed using all 4 studies confirms 2 signals at SOX17 locus

Figure 1: **Flowchart showing study design.** Four studies of European-ancestry individuals were included; one NIHRBR included rare disease patients and relatives for whom whole genome sequencing was performed. PAH patients were compared with non-PAH patients and their relatives in one discovery GWAS. Three studies, PAHB, PHAAR and BHFPAH, included PAH patients and non-PAH controls from the US, France and a mixture of European countries, respectively, for whom genome-wide array data were acquired. PAH patients were compared with non-PAH controls in each study and the results were meta-analysed in another discovery GWAS. Genome-wide significant hits from each GWAS were selected for cross-validation. Finally, all four studies were meta-analysed to provide overall associations, and conditional analysis correcting for most significant variants at each locus were used to resolve signals for multiple associations.

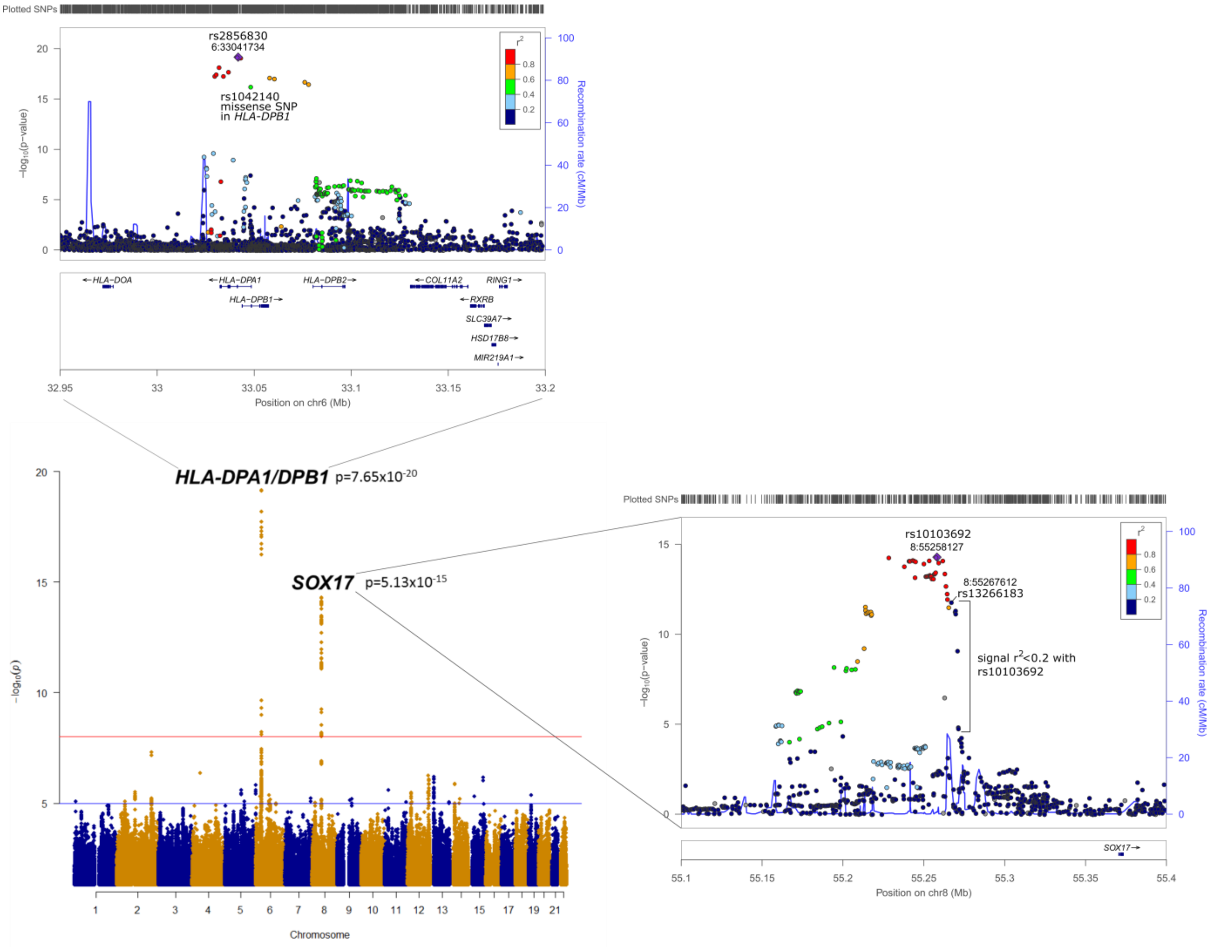


Figure 2 – A meta-analysis of all cohorts and regional plots of novel loci. The regional plots indicate variant location and linkage disequilibrium (LD) structure at the *HLA-DPA1/DPB1* and *SOX17* loci, respectively. At the *SOX17* locus, several variants associated with PAH are in very weak or no LD (r^2<0.2) with the lead SNP, rs10103692. We refer to these variants as *SOX17* signal 1 and the most significant, rs13266183, is indicated. The variants coloured as in LD with rs10103692 comprise signal 2.

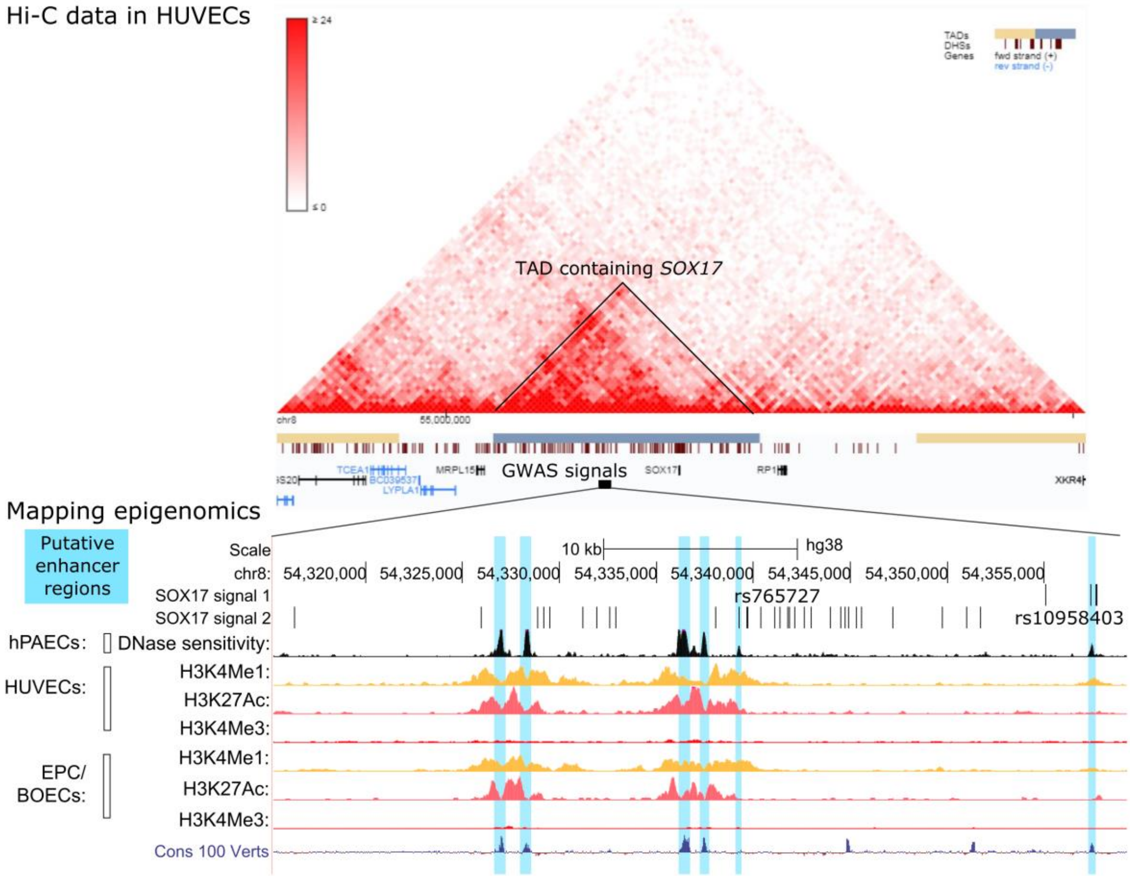
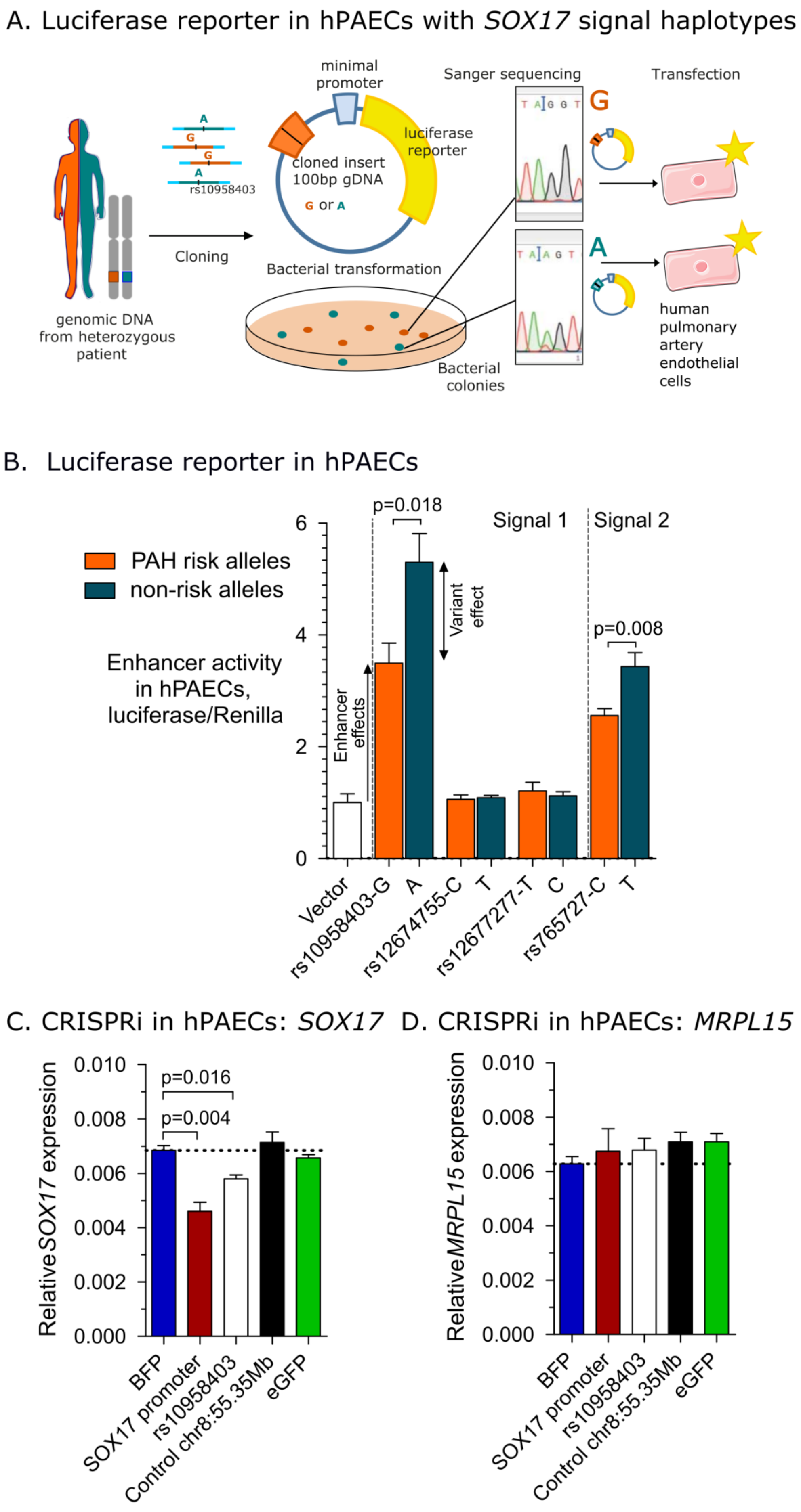


Figure 3 - *In silico* analysis of SOX17 locus. Hi-C data from human umbilical vein endothelial cells indicates regions of DNA found in close proximity in the 3D structure. The GWAS locus position is indicated by a black box, overlapping a TAD indicated in blue which contains only *SOX17.* Mapping of *SOX17* locus variants associated with PAH with public epigenomic data is underneath Hi-C data. The credible set indicates positions of variants 99% likely to contain the causal variants. Transcription factor binding sites as determined by ChIP-Seq experiments of 161 factors from ENCODE with Factorbook Motifs are shown; H indicates binding site in HeLa-S3 cervix adenocarcinoma cells, U indicates binding site in human umbilical vein endothelial cells (HUVEC). Auxiliary hidden markov models (HMM), which summarize epigenomic data to predict the functional status of genomic regions in different tissues/cells, are shown. Epigenomic data in endothelial cells (EC) including HUVEC, human pulmonary artery ECs (hPAECs) and endothelial progenitor cells (EPC), also known as blood outgrowth ECs (BOEC), indicate areas likely to contain active regulatory regions and promoters. Markers include histone 3 lysine 4 monomethylation (H3K4Me1, often found in enhancers) and trimethylation (H3K4Me3 strongly observed in promoters) and lysine 27 acetylation (often found in active regulatory regions). The blue areas indicate where epigenomic data suggest a putative enhancer region, some overlapped by variants associated with PAH. These regions were cloned for the luciferase reporter experiments (results in Figure 4B).



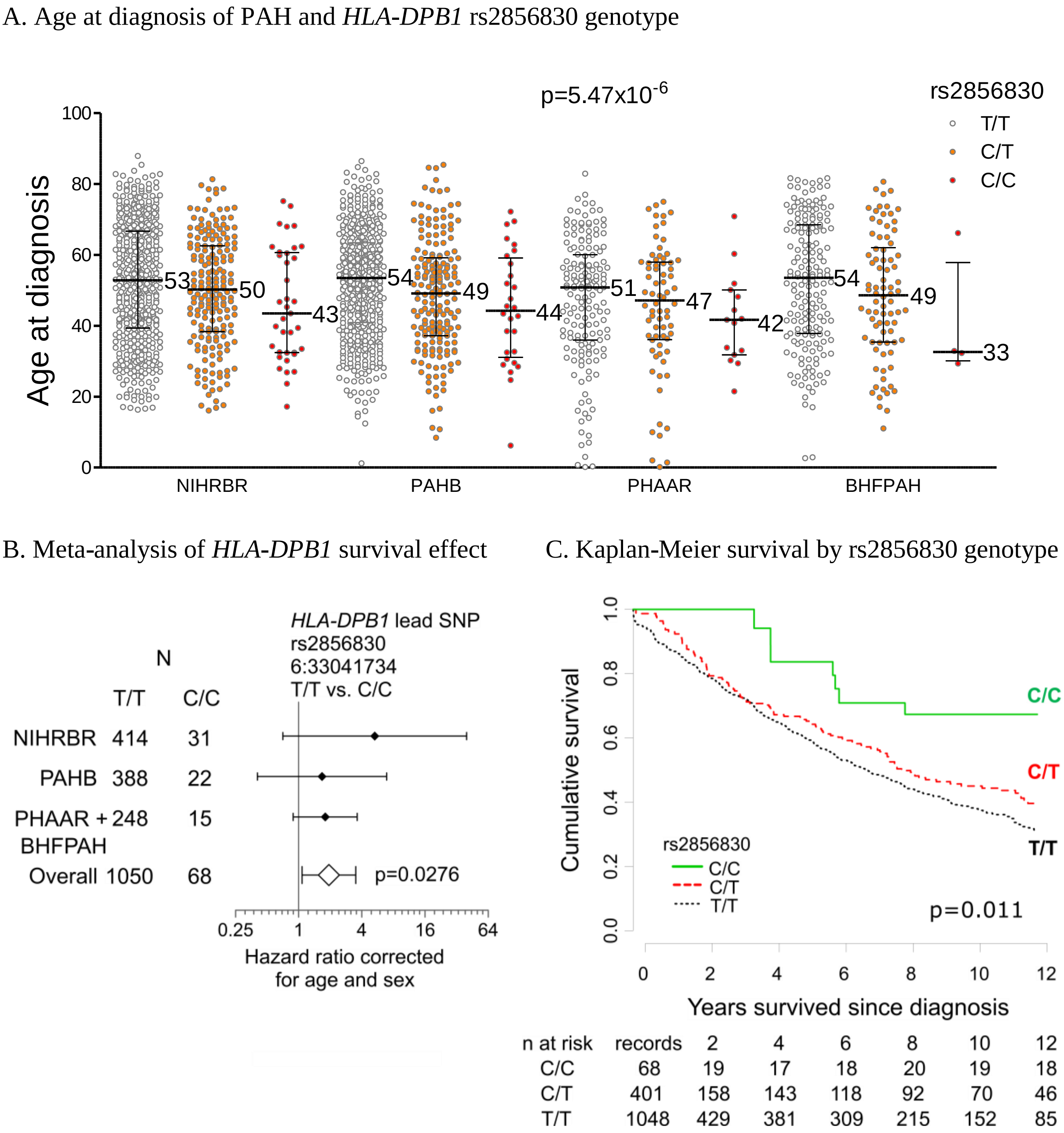
A. Luciferase reporter in hPAECs with *SOX17* signal haplotypes

B. Luciferase reporter in hPAECs

C. CRISPRi in hPAECs: *SOX17*

D. CRISPRi in hPAECs: *MRPL15*

Figure 4: A. Cartoon describing process for haplotype specific reporter construct derivation. 100bp gDNA inserts containing *SOX17* SNPs are isolated from blood outgrowth endothelial cells derived from a PAH patient heterozygous for the *SOX17* SNPs. Colonies of transformed bacteria can be sequenced to determine allele present in product. Transfection of luciferase reporter constructs containing inserts into human pulmonary artery endothelial cells allows for determination of luciferase activity. **B.** Luciferase reporter assay results. Luciferase/Renilla ratios relative to the empty vector demonstrate haplotype-dependent enhancement of promoter activity. Enhancer effects were tested by one way analysis of variance followed by Dunnett's post-hoc tests - rs10958403-G/A and rs765727-C/T were both $p<0.0001$ significant versus empty vector, variant effects of these 2 SNPs were tested by t-test. Mean±SEM of n=5 experiments. **C.** Relative expression of *Sox17:beta-actin* in hPAECs upon CRISPR-mediated repression of the near *SOX17* GWAS locus. Mean±SEM of n=4 measurements in a representative experiment. 3 further experiments showed consistent results. BFP, blue fluorescent protein; eGFP, enhanced green fluorescent protein; and control, which refers to a region between the enhancer region and the *SOX17* gene that is negative for regulatory markers, are used as negative controls. The *SOX17* promoter was targeted as a positive control of repression. Significance shown vs. BFP by Dunnett's post-hoc analysis. **D.** Relative expression of *Mrpl15:beta-actin* in hPAECs upon CRISPR-mediated repression of the GWAS locus.



A. Age at diagnosis of PAH and *HLA-DPB1* rs2856830 genotype

B. Meta-analysis of *HLA-DPB1* survival effect

C. Kaplan-Meier survival by rs2856830 genotype

n at risk	records	2	4	6	8	10	12
C/C	68	19	17	18	20	19	18
C/T	401	158	143	118	92	70	46
T/T	1048	429	381	309	215	152	85

Figure 5 – Clinical impact of *HLA-DPB1* rs2856830. A. Age at diagnosis by genotype in four cohorts of PAH patients. Bars indicate medians and interquartile range, numbers given are median values in subgroups. p-value shown is from linear regression model correcting for cohort differences. **B.** Forest plot showing hazard ratios for the rs2856830T/T vs C/C genotypes, corrected for age and sex in Cox regression survival analyses in each PAH cohort, individually and with meta-analysis results. **C.** Kaplan-Meier survival plot in PAH patients divided into groups based on the genotype of *HLA-DPA1/DPB1* SNP rs2856830 in all cohorts. N at risk indicates numbers at risk in each time period, which increases as truncated patients are recruited into the study after diagnosis and decreases as patient follow-up ends. Significance from log rank test is given.

Table legends

Table 1 - Novel loci associated with PAH in sequenced and genotyped cohorts. Odds ratios are for association between effect allele and PAH. gnomAD is the Genome Aggregation Database, which provides information including allele frequencies in different populations.

Table 2 - Associations of *HLA-DPB1* alleles with the lead SNP rs2856830. Orange indicates alleles and residues depleted in PAH cases and green indicates those enriched in PAH cases.

Figure legends

Figure 1: **Flowchart showing study design.** Four studies of European-ancestry individuals were included; one NIHRBR included rare disease patients and relatives for whom whole genome sequencing was performed. PAH patients were compared with non-PAH patients and their relatives in one discovery GWAS. Three studies, PAHB, PHAAR and BHFPAH, included PAH patients and non-PAH controls from the US, France and a mixture of European countries, respectively, for whom genome-wide array data were acquired. PAH patients were compared with non-PAH controls in each study and the results were meta-analysed in another discovery GWAS. Genome-wide significant hits from each GWAS were selected for cross-validation. Finally, all four studies were meta-analysed to provide overall associations, and conditional analysis correcting for most significant variants at each locus were used to resolve signals for multiple associations.

Figure 2 – A meta-analysis of all cohorts and regional plots of novel loci. The regional plots indicate variant location and linkage disequilibrium (LD) structure at the *HLA-DPA1/DPB1* and *SOX17* loci, respectively. At the *SOX17* locus, several variants associated with PAH are in very weak or no LD (r^2<0.2) with lead SNP rs10103692. We refer to these variants as *SOX17* signal 1 and the most significant, rs13266183 is indicated. The variants coloured as in LD with rs10103692 comprise signal 2.

Figure 3 - In silico analysis of *SOX17* locus. A. Hi-C data from human umbilical vein endothelial cells indicates regions of DNA found in close proximity in the 3D structure. The GWAS locus position is indicated by a black box, overlapping a TAD indicated in blue which contains only SOX17. **B.** Mapping of *SOX17* locus variants associated with PAH with public epigenomic data. The credible set indicates positions of variants 99% likely to contain the causal variants. Transcription factor binding sites as determined by ChIP-Seq experiments of 161 factors from ENCODE with Factorbook Motifs are shown; H indicates binding site in HeLa-S3 cervix adenocarcinoma cells, U indicates binding site in human umbilical vein endothelial cells (HUVEC). Auxiliary hidden markov models (HMM), which summarize epigenomic data to predict the functional status of genomic regions in different tissues/cells, are shown. Epigenomic data in endothelial cells (EC) including HUVEC, human pulmonary artery ECs (hPAECs) and endothelial progenitor cells (EPC), also known as blood outgrowth ECs (BOEC), indicate areas likely to contain active regulatory regions and promoters. Markers include histone 3 lysine 4 monomethylation (H3K4Me1, often found in enhancers) and trimethylation (H3K4Me3 strongly observed in promoters) and lysine 27 acetylation (often found in active regulatory regions). The blue areas indicate where epigenomic data suggest a putative enhancer region, some overlapped by variants associated with PAH. These regions were cloned for the luciferase reporter experiments (results in Figure 4B).

Figure 4: A. Cartoon describing process for haplotype specific reporter construct derivation. 100bp gDNA inserts containing *SOX17* SNPs are isolated from blood outgrowth endothelial cells derived

from a PAH patient heterozygous for the *SOX17* SNPs. Colonies of transformed bacteria can be sequenced to determine allele present in product. Transfection of luciferase reporter constructs containing inserts into human pulmonary artery endothelial cells allows for determination of luciferase activity. **B.** Luciferase reporter assay results. Luciferase/Renilla ratios relative to the empty vector demonstrate haplotype-dependent enhancement of promoter activity. Enhancer effects were tested by one way analysis of variance followed by Dunnett's post-hoc tests - rs10958403-G/A and rs765727-C/T were both $p<0.0001$ significant versus empty vector, variant effects of these 2 SNPs were tested by t-test. Mean±SEM of n=5 experiments. **C.** Relative expression of Sox17:beta-actin in hPAECs upon CRISPR-mediated repression of the near *SOX17* GWAS locus. Mean±SEM of n=4 measurements in a representative experiment. 3 further experiments showed consistent results. BFP, blue fluorescent protein; eGFP, enhanced green fluorescent protein; and control, which refers to a region between the enhancer region and the *SOX17* gene that is negative for regulatory markers, are used as negative controls. The *SOX17* promoter was targeted as a positive control of repression. Significance shown vs. BFP by Dunnett's post-hoc analysis. **D.** Relative expression of Mrpl15:beta-actin in hPAECs upon CRISPR-mediated repression of the GWAS locus.

Figure 5 – Clinical impact of *HLA-DPB1* rs2856830. A. Age at diagnosis by genotype in four cohorts of PAH patients. Bars indicate medians and interquartile range, numbers given are median values in subgroups. p-value shown is from linear regression model correcting for cohort differences. **B.** Forest plot showing hazard ratios for the rs2856830T/T vs C/C genotypes, corrected for age and sex in Cox regression survival analyses in each PAH cohort, individually and with meta-analysis results. **C.** Kaplan-Meier survival plot in PAH patients divided into groups based on the genotype of HLA-DPA1/DPB1 SNP rs2856830 in all cohorts. N at risk indicates numbers at risk in each time period, which increases as truncated patients are recruited into the study after diagnosis and decreases as patient follow-up ends. Significance from log rank test is given.