Genetic determinants of risk in pulmonary arterial hypertension: international genome-wide association studies and meta-analysis


Summary

Background Rare genetic variants cause pulmonary arterial hypertension, but the contribution of common genetic variation to disease risk and natural history is poorly characterised. We tested for genome-wide association for pulmonary arterial hypertension in large international cohorts and assessed the contribution of associated regions to outcomes.

Methods We did two separate genome-wide association studies (GWAS) and a meta-analysis of pulmonary arterial hypertension. These GWAS used data from four international case-control studies across 11744 individuals with European ancestry (including 2085 patients). One GWAS used genotypes from 5895 whole-genome sequences and the other GWAS used genotyping array data from an additional 5849 individuals. Cross-validation of loci reaching genome-wide significance was conducted for genome-wide associations and meta-analysis. Additional studies were needed to confirm the association between HLA typing or rs2856830 genotyping and survival, and to determine whether HLA typing or rs2856830 genotyping improves risk stratification in clinical practice or trials.

Findings A locus near SOX17 (rs10103692, odds ratio 1·80 [95% CI 1·55–2·08], p=5·13×10⁻¹⁵) and a second locus in HLA-DPA1 and HLA-DPB1 (collectively referred to as HLA-DPA1/DPB1 here; rs2856830, 1·56 [1·42–1·71], p=7·65×10⁻²⁰) within the class II MHC region were associated with pulmonary arterial hypertension. The SOX17 locus had two independent signals associated with pulmonary arterial hypertension (rs13266183, 1·36 [1·25–1·48], p=7·6×10⁻¹⁵; and rs10103692). Functional and epigenomic data indicate that the risk variants near SOX17 alter gene regulation via an enhancer active in endothelial cells. Pulmonary arterial hypertension risk variants determined haplotype-specific enhancer activity, and CRISPR-mediated inhibition of the enhancer reduced SOX17 expression. The HLA-DPA1/DPB1 rs2856830 genotype was strongly associated with survival. Median survival from diagnosis in patients with pulmonary arterial hypertension with the C/C homozygous genotype was double (13–50 years [95% CI 12·07 to >13·50]) that of those with the T/T genotype (6·97 years [6·02–8·05]), despite similar baseline disease severity.

Interpretation This is the first study to report that common genetic variation at loci in an enhancer near SOX17 and in HLA-DPA1/DPB1 is associated with pulmonary arterial hypertension. Impairment of SOX17 function might be more common in pulmonary arterial hypertension than suggested by rare mutations in SOX17. Further studies are needed to confirm the association between HLA typing or rs2856830 genotyping and survival, and to determine whether HLA typing or rs2856830 genotyping improves risk stratification in clinical practice or trials.

Funding UK NIHR, BHF, UK MRC, Dinosaur Trust, NIH/NHLBI, ERS, EMBO, Wellcome Trust, EU, AHA, ACCinPharm, 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.

Copyright © 2018 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.
Pulmonary arterial hypertension refers to an uncommon but devastating disorder characterised by oblitative pulmonary vascular remodelling, leading to a progressive increase in pulmonary vascular resistance and right heart failure. Annual mortality for idiopathic and heritable pulmonary arterial hypertension remains around 10%, despite the use of modern therapies.\(^1,2\) The high mortality partly reflects the limited effect of available licensed treatments on 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 characterised heterogeneity in the causes of pulmonary arterial hypertension.

Recent gene sequencing studies\(^4,5\) have revealed rare mutations in several genes, including BMPR2, genes encoding potassium channels, and most recently the transcription factor SOX17.\(^6,7\) Rare genetic variation is associated with both the risk of developing pulmonary arterial hypertension in large international cohorts\(^8\).\(^-\(^10\) We did two genome-wide association studies (GWAS)\(^11\), rare genetic analysis including those attributable to common variation, primarily in genes associated with transforming growth factor-\(\beta\) family members, including BMPR2, but also in the transcription factor SOX17, is known to cause pulmonary arterial hypertension. However, little is known about the contribution of common variation to this disorder. Additionally, both pulmonary vascular endothelial dysfunction and altered immune and inflammatory signalling are observed in pulmonary arterial hypertension, but the underlying genetic mechanisms are poorly characterised.

### Methods

**Pulmonary arterial hypertension cohorts and genotyping**

We did two genome-wide association studies and a meta-analysis on pulmonary arterial hypertension. Pulmonary arterial hypertension was defined by haemodynamic criteria according to international guidelines.\(^7\) Unrelated individuals with idiopathic, heritable, or anorexigen-associated pulmonary arterial hypertension were included. Individuals with evidence of other known causes of pulmonary arterial hypertension were excluded; therefore, no patients were known to have pulmonary arterial hypertension associated with clinically diagnosable autoimmune diseases (appendix pp 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, opt-out policy (appendix p 2).

Given the rarity of pulmonary arterial hypertension, four studies were used for the analyses. In the UK National Institute for Health Research BioResource (NIHRBR) Rare Diseases study, whole-genome sequencing (Illumina, San Diego, CA, USA; mean depth of around 35x; appendix p 2) was done in 5895 individuals of European descent, each with a rare disorder from 16 categories or their unaffected relatives, and 847 had pulmonary arterial hypertension (appendix p 4). 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 (appendix p 6).

Three studies used genome-wide genotyping arrays: the US National Biological Sample and Data Repository for Pulmonary Arterial Hypertension (also known as...
PAH Biobank (PAHB) study included 694 individuals with pulmonary arterial hypertension and 1560 controls ascertained for a large pharmacogenomic study at Vanderbilt University (Nashville, TN, USA); the Pulmonary Hypertension Allele-Associated Risk (PHAAR) study included 269 individuals with pulmonary arterial hypertension and 1068 population-based controls from France; and the British Heart Foundation Pulmonary Arterial Hypertension (BHFPAH) study consisted of 275 individuals with pulmonary arterial hypertension and 1983 population-based controls from several European countries (appendix p 12). All genotyping studies were imputed (appendix p 12), and single-nucleotide polymorphisms (SNPs) with good imputation quality ($r^2=0.3$) were taken for testing. Individuals from NIHRBR, PHAAR, and BHFPAH were tested for relatedness to prevent inclusion of the same or related individuals across studies. Other quality-control steps are detailed in the appendix (pp 4, 5, 12).

**Association analyses**

We used logistic regression to test single-marker variants for genetic association with a diagnosis of pulmonary arterial hypertension assuming a log-additive genetic model and adjusting for sex, read length chemistry (NIHRBR only), and population structure using the first four (NIHRBR and PHAAR), three (PAHB), or ten (BHFPAH) principal components. We calculated the genomic inflation factor, which was verified to be between 1 and 1.05 for each study.

We used two independent sets for discovery: whole-genome sequencing data from NIHRBR (n=5895, including 847 pulmonary arterial hypertension cases); and meta-analysis of genotyping studies PAHB, PHAAR, and BHFPAH (n=5849, including 1238 pulmonary arterial hypertension cases). We cross-validated findings and confirmed loci in a meta-analysis of all four studies using the inverse variance-weighted fixed-effects approach (which maximises power for discovery studies), implemented in the GWAMA software tool. Random-effects meta-analysis was subsequently applied to estimate generalisability of the results to different populations.

We did a conditional analysis including the lead variant in each locus as a covariate to test for independent distinct genetic variants (SNPs) covering the genome, to test whether an allele of a genetic variant is associated with a disease or trait, or the levels of a continuous trait of interest.

**Common variant**

An SNP for which the frequency of the less frequent allele is at least 5% in a given population. Typically, a common variant has subtle biological effects, as opposed to rare variants (also known as mutations), which can cause diseases or extreme phenotypes.

**Genetic locus**

A position on the genome defined by the chromosome number and the genetic distance in centimorgans (cM) or physical distance in base pairs (bps) on the chromosome. A locus can refer to a gene or a non-coding region of varying length (eg, from one hundred to millions of bps).

**Credible set**

A set of variants that is statistically likely (eg, with 95% probability) to contain the causal variant for the disease or trait of interest at a genetic locus.

**Panel: Key terms**

**Genome-wide association study**

A genetic analysis approach, typically using millions of common variants (eg, single-nucleotide polymorphisms [SNPs]) covering the genome, to test whether an allele of a genetic variant is associated with a disease or trait, or the levels of a continuous trait of interest.

Enbridge, Factorbook Motifs, and Blueprint). The locus was investigated using CRISPR-mediated repression in human pulmonary artery endothelial cells (hpPAECs; 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 20 bp guide RNA (appendix pp 6–7). Cells were harvested following blasticidin selection, and the expression of SOX17 as well as neighbouring MRPL15 and TMEM68 was 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 was isolated from endothelial progenitor cells (also known as blood outgrowth endothelial cells) derived from a patient with pulmonary arterial hypertension who was heterozygous for the lead SNP at SOX17 and used to clone 100 bp putative enhancer regions containing the SOX17 pulmonary arterial hypertension 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 genomic DNA inserts differing only by the allele of the SNP of interest. Reporter plasmids were transfected into hpPAECs by electroporation, and luciferase activity was measured to quantify the enhancer function of the inserts with the relevant haplotype.

**Association analyses**

We used logistic regression to test single-marker variants for genetic association with a diagnosis of pulmonary arterial hypertension assuming a log-additive genetic model and adjusting for sex, read length chemistry (NIHRBR only), and population structure using the first four (NIHRBR and PHAAR), three (PAHB), or ten (BHFPAH) principal components. We calculated the genomic inflation factor, which was verified to be between 1 and 1.05 for each study.

We used two independent sets for discovery: whole-genome sequencing data from NIHRBR (n=5895, including 847 pulmonary arterial hypertension cases); and meta-analysis of genotyping studies PAHB, PHAAR, and BHFPAH (n=5849, including 1238 pulmonary arterial hypertension cases). We cross-validated findings and confirmed loci in a meta-analysis of all four studies using the inverse variance-weighted fixed-effects approach (which maximises power for discovery studies), implemented in the GWAMA software tool. Random-effects meta-analysis was subsequently applied to estimate generalisability of the results to different populations.

We did a conditional analysis including the lead variant in each locus as a covariate to test for independent distinct signals reaching $p<5\times10^{-8}$.

We used LDlink to assess linkage disequilibrium of variants in all European populations from the 1000 Genomes Project. Credible sets of variants considered 99% likely to include the functional causal variants were calculated by summing ranked posterior probabilities (appendix p 6).

**Annotation and functional assessment of the locus near SOX17**

The locus near SOX17 was assessed against publicly available functional annotation datasets (including ENCODE, Factorbook Motifs, and Blueprint). The locus was investigated using CRISPR-mediated repression in human pulmonary artery endothelial cells (hpPAECs; 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 20 bp guide RNA (appendix pp 6–7). Cells were harvested following blasticidin selection, and the expression of SOX17 as well as neighbouring MRPL15 and TMEM68 was 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 was isolated from endothelial progenitor cells (also known as blood outgrowth endothelial cells) derived from a patient with pulmonary arterial hypertension who was heterozygous for the lead SNP at SOX17 and used to clone 100 bp putative enhancer regions containing the SOX17 pulmonary arterial hypertension 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 genomic DNA inserts differing only by the allele of the SNP of interest. Reporter plasmids were transfected into hpPAECs by electroporation, and luciferase activity was measured to quantify the enhancer function of the inserts with the relevant haplotype.

**Panel: Key terms**

**Genome-wide association study**

A genetic analysis approach, typically using millions of common variants (eg, single-nucleotide polymorphisms [SNPs]) covering the genome, to test whether an allele of a genetic variant is associated with a disease or trait, or the levels of a continuous trait of interest.

**Common variant**

An SNP for which the frequency of the less frequent allele is at least 5% in a given population. Typically, a common variant has subtle biological effects, as opposed to rare variants (also known as mutations), which can cause diseases or extreme phenotypes.

**Genetic locus**

A position on the genome defined by the chromosome number and the genetic distance in centimorgans (cM) or physical distance in base pairs (bps) on the chromosome. A locus can refer to a gene or a non-coding region of varying length (eg, from one hundred to millions of bps).

**Credible set**

A set of variants that is statistically likely (eg, with 95% probability) to contain the causal variant for the disease or trait of interest at a genetic locus.
Articles

Statistical analysis

Loci associated with pulmonary arterial hypertension were tested for associations with clinical variables (appendix pp 17–18). All-cause mortality was the primary endpoint in survival analyses using Kaplan-Meier estimates and Cox regression in the survival package in R, version 3.3.0. Survival was calculated from diagnosis to date of death or censoring (Oct 31, 2016, for NIHRBR; Aug 1, 2017, for PAHB; Sept 27, 2017, for PHAAR; Oct 12, 2017, for BHFPAH), with left truncation using date of genetic consent, and patients were censored at lung or heart-and-lung transplantation. Age and sex were covariates to correct for their association with prognosis. NIHRBR and PAHB analysed separately, and PHAAR and BHFPAH were combined before analysis because of their smaller sample sizes. Cox regression results from these three analyses were then meta-analysed using the default random-effects model restricted maximum-likelihood estimator method implemented in the metafor package in R, version 3.3.0. All cohorts were combined for Kaplan-Meier analysis. We did sensitivity analyses excluding pathogenic BMPR2 variant carriers, all pathogenic rare variant carriers, and patients diagnosed in previous decades who might have been exposed to different treatments.

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. HLA alleles and amino acids were tested for association with the novel lead variants or case-control status by χ² test with false discovery rate correction.

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. CJR, KB, MB, MHa, LSo, MG, MWP, CH, AA, Kbh, JHK, MK, AU, LH, JA, EMS, and SGr had access to raw data for analyses. The corresponding authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

In two separate GWAS discovery analyses (figure 1), we identified two loci associated with pulmonary arterial hypertension reaching genome-wide significance (p<5×10⁻⁸; table 1; appendix p 23). One locus was 100–200 kb upstream of the transcription factor SOX17. A second locus was within HLA-DPA1/DPB1, which encodes the MHC class II DP α and β chains. Both the SOX17 and HLA-DPA1/DPB1 loci reached genome-wide significance in the discovery analyses; our cross-validation strategy confirmed that the same alleles were more frequent in pulmonary arterial hypertension than in other disease or population controls in both analyses (table 1). The genome-wide meta-analysis of all four studies confirmed their associations with pulmonary arterial hypertension (rs2856830, odds ratio 1.56–2.75 with PAH).

Figure 1: Study design

HLA-DPA1 and HLA-DPB1 are collectively referred to as HLA-DPA1/DPB1 in this Article. BHFPAH=British Heart Foundation Pulmonary Arterial Hypertension study. GWAS=genome-wide association study. NIHRBR=National Institute for Health Research BioResource study. PAH=pulmonary arterial hypertension. PAHB=PAH Biobank study. PHAAR=Pulmonary Hypertension Allele-Associated Risk study.
To study the effects of the pulmonary arterial hypertension risk variants on the putative enhancers defined by the epigenomic signals, we developed reporter constructs containing 100 bp of the regions containing either the risk allele or non-risk alleles at each of the four SNPs using genomic DNA from a patient heterozygous for both rs10958403 and rs765727. Of these constructs, constructs containing rs12674755 or rs12677277 exhibited enhancer activity (Appendix p 14). One of these elements (around hg19 chr8:55-270 Mb) contains a cluster of three of four credible variants from SOX17 signal 1 (Figure 3). Another (around hg19 chr8:55-252 Mb) contains one credible variant from SOX17 signal 2. Of these variants, rs10958403 in signal 1 and rs765727 in signal 2 overlap a Dnase I hypersensitivity signal, which indicates accessible chromatin (allowing binding of transcription factors), detected in hPAECs (Figure 3).

To search for evidence of regulatory elements in relevant tissues at SOX17 signal 1 and signal 2, we examined publicly available epigenomic data (including histone modifications; Figure 3; Appendix p 24). We identified several putative enhancer elements active in both lung tissue and endothelial cells (Figure 3). One of these elements (around hg19 chr8:55-270 Mb) contains a cluster of three of four credible variants from SOX17 signal 1 (Figure 3). Another (around hg19 chr8:55-252 Mb) contains one credible variant from SOX17 signal 2. Of these variants, rs10958403 in signal 1 and rs765727 in signal 2 overlap a Dnase I hypersensitivity signal, which indicates accessible chromatin (allowing binding of transcription factors), detected in hPAECs (Figure 3).

To search for evidence of regulatory elements in relevant tissues at SOX17 signal 1 and signal 2, we examined publicly available epigenomic data (including histone modifications; Figure 3; Appendix p 24). We identified several putative enhancer elements active in both lung tissue and endothelial cells (Figure 3). One of these elements (around hg19 chr8:55-270 Mb) contains a cluster of three of four credible variants from SOX17 signal 1 (Figure 3). Another (around hg19 chr8:55-252 Mb) contains one credible variant from SOX17 signal 2. Of these variants, rs10958403 in signal 1 and rs765727 in signal 2 overlap a Dnase I hypersensitivity signal, which indicates accessible chromatin (allowing binding of transcription factors), detected in hPAECs (Figure 3).

To study the effects of the pulmonary arterial hypertension risk variants on the putative enhancers defined by the epigenomic signals, we developed reporter constructs containing 100 bp of the regions containing either the risk allele or non-risk alleles at each of the four 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 threefold and sixfold induction of luciferase:Renilla ratio, p<0.0001), whereas constructs containing rs12674755 or rs12677277 had no effect compared with the empty vector control (Figure 4). We also observed haplotype-specific activity with the active constructs, which differed only by the alleles at pulmonary arterial hypertension-associated risk variants rs10958403 or rs765727 (both p<0.05; Figure 4).
DNA folding patterns determined by Hi-C data from lung tissue and endothelial cells (human umbilical vein endothelial cells [figure 3] and human microvascular endothelial cells [data not shown]) indicate that the SOX17 pulmonary arterial hypertension locus resides in a defined topologically associated domain in which the only gene found, and thus likely target of any regulatory elements in this region, is SOX17.

CRISPR-mediated inhibition of the SOX17 signal 1 region in hPAECs resulted in selective downregulation of SOX17 expression but not the expression of neighbouring genes MRPL15 and TMEM68, suggesting that the enhancers in this locus specifically regulate SOX17 (figure 4; appendix p 26).

We investigated whether the HLA-DPA1/DPB1 and SOX17 variants affect clinical outcomes in pulmonary arterial hypertension, 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 patients with pulmonary arterial hypertension with the C/C homozygous genotype was double (13·50 years [95% CI 12·07 to >13·50]) that of those with the T/T genotype (6·97 years [6·02–8·05]). Cox regression survival analyses showed that the rs2856830 T/T genotype conferred an increased annual risk of death in pulmonary arterial hypertension (hazard ratio [HR] 1·94 [95% CI 1·08–3·51]; figure 5).

Sensitivity analyses excluding pathogenic BMPR2 variant carriers, all pathogenic rare variant carriers, and patients diagnosed in previous decades who might have...
been exposed to different treatment regimens gave results similar to the main analyses (appendix p 27).

We tested both loci for association with other clinical variables, including disease severity measures and comorbidities (appendix pp 17–18). The C allele at HLA-DPA1/DPB1 lead SNP rs2856830 was associated with younger age at diagnosis (figure 5), with C/C homozygotes presenting a decade earlier than T/T homozygotes (appendix p 17). The rs2856830 genotype was not associated with vasoresponder status.

The HLA-DPA1/DPB1 locus included a missense variant rs1042140 in HLA-DPB1 reaching genome-wide significance (table 1) in partial linkage disequilibrium ($r^2=0.45$ with lead rs2856830 in Europeans). The SNP rs1042140 determines a glutamic acid (Glu⁶⁹) 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 two digits represent related groups of similar alleles (e.g., DPB1*02), and four digits represent specific proteins with distinct amino acid sequences (e.g., DPB1*02:01). We found that the pulmonary arterial hypertension-enriched C allele of rs2856830 was associated with HLA-DPB1*02:01/02:02/16:01 (all $p<1\times10^{-9}$ after false discovery rate correction; table 2; appendix p 19), which

Figure 3: In-silico analysis of SOX17 locus

Hi-C data from human umbilical vein endothelial cells (hUVECs) indicate regions of DNA found in close proximity in the three-dimensional structure. The genomic region containing the significant variants identified by the genome-wide association study (GWAS) analysis is indicated by a black box, overlapping a topologically associated domain (TAD) indicated in blue, which contains only SOX17. Mapping of SOX17 locus variants associated with pulmonary arterial hypertension with public epigenomic data is underneath Hi-C data. The credible set indicates positions of variants 99% likely to contain the causal variants. Auxiliary hidden Markov models, which summarise epigenomic data to predict the functional status of genomic regions in different tissues or cells, are shown. Epigenomic data in endothelial cells including hUVECs, human pulmonary artery endothelial cells (hPAECs), and endothelial progenitor cells (EPCs), indicate areas likely to contain active regulatory regions and promoters. Markers include histone H3 lysine 4 monomethylation (H3K4Me1; often found in enhancers) and trimethylation (H3K4Me3; strongly observed in promoters) and H3 lysine 27 acetylation (H3K27Ac; often found in active regulatory regions). The blue vertical blocks indicate where epigenomic data suggest a putative enhancer region, some overlapped by variants associated with pulmonary arterial hypertension. These regions were cloned for the luciferase reporter experiments (figure 4B). DHSs=DNase hypersensitivity sites.
all contain the Glu\(^69\) residue. The most numerous DPB1*02:01 and DPB1*04:01 alleles were associated with survival in patients with pulmonary arterial hypertension (HR 0·70 [95% CI 0·49–1·00] for DPB1*02:01 and 1·33 [1·04–1·70] for DPB1*04:01; appendix p 21).

The risk alleles at both signals within the SOX17 locus are common (risk allele frequencies are 74% for rs13266183-C and 92% for rs9298503-C), such that 1230 (59%) of 2085 patients with pulmonary arterial hypertension were homozygous for the risk allele at both SOX17 SNPs, compared with 4443 (46%) of 9659 controls.

The alleles at HLA-DPB1 associated with the poorest outcomes are also common (risk allele frequency of 86% for rs2856830-T), such that 1432 (69%) of 2085 patients with PAH were homozygous for the risk allele at this locus, compared with 4443 (46%) of 9659 controls.

**Figure 4:** In-vitro analysis of SOX17 locus

(A) Process for haplotype-specific reporter construct derivation. 100 bp genomic DNA inserts containing SOX17 single-nucleotide polymorphisms (SNPs) are isolated from endothelial progenitor cells derived from a patient with pulmonary arterial hypertension (PAH) heterozygous for the SOX17 SNPs. Colonies of transformed bacteria can be sequenced to determine alleles present in the product. Transfection of luciferase reporter constructs into human pulmonary artery endothelial cells (hPAECs) 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 two SNPs were tested by \(t\) test. The mean (SEM; error bars) of five experiments is shown. (C) Relative expression of SOX17:ACTB in hPAECs on CRISPR-mediated repression of the near SOX17 genome-wide association study (GWAS) locus. The mean (SEM; error bars) of four measurements in a representative experiment is shown. Three further experiments showed consistent results. Blue fluorescent protein (BFP), enhanced green fluorescent protein (eGFP), 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 versus BFP by Dunnett’s post-hoc analysis. (D) Relative expression of MRPL15:ACTB in hPAECs on CRISPR-mediated repression of the GWAS locus.
2085 patients with pulmonary arterial hypertension had the T/T genotype associated with the poorest outcomes and 1975 (95%) of 2085 patients had at least one T allele.

**Discussion**

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

Both in-silico and experimental analyses of the common variants upstream of the SOX17 gene suggest that they affect susceptibility to pulmonary arterial hypertension through regulation of SOX17 expression. We provided direct evidence that inhibition of SOX17 signal 1 reduced SOX17 expression, and luciferase activity experiments showed a functional variant in both signals. Combined with the in-silico data, it seems highly likely that signal 2 would also be targeting SOX17 because no other gene is present in the topologically associated domain and promoter capture HiC data show that this area associates...
with the SOX17 promoter. We have recently reported enrichment and familial segregation in pulmonary arterial hypertension of causal rare deleterious variation in SOX17, implicating this gene in the pathogenesis of pulmonary arterial hypertension. SOX17 is involved in the development of the endoderm, vascular endothelium, haemopoietic cells, and cardiomyocytes.

SOX17 also determines the endothelial fate of CD34 progenitor cells de-differentiated from fibroblasts. Deletion in the mouse leads to abnormal pulmonary vascular development, poor distal lung perfusion and biventricular hypertrophy. SOX17 is a pro-angiogenic transcription factor and interacts with well established endothelial molecular mediators: reduction of SOX17 in endothelial cells through Notch activation (which is associated with BMPR2 signalling) restricts angiogenesis. Conversely, vascular endothelial growth factor (VEGF) upregulates SOX17 and, as part of a positive-feedback loop, SOX17 promotes expression of VEGF receptor 2.

We report that HLA-DPB1 alleles are associated with pulmonary arterial hypertension and have a pivotal role in determining disease progression. The beneficial effect of the C/C genotype at rs2856830 on survival is clinically significant, extending average survival from about 7 to 14 years, despite no apparent difference in baseline disease severity by standard clinical measures, including haemodynamics and exercise capacity. Patients with the C allele at rs2856830 presented at a significantly younger age than those with the T allele, but the association of the HLA-DPB1 SNP with survival remains significant after correction for both age and sex. The parallel in pulmonary arterial hypertension is probably through its association with specific HLA-DPB1 alleles. Class II (HLA-DRB1, HLA-DQB1, and HLA-DPB1) antigen-presenting proteins have crucial roles in the adaptive immune response.

The mechanism of rs2856830 involvement in pulmonary arterial hypertension is likely to be driven by trans effects of known rare pathogenic variants in pulmonary arterial hypertension, we did sensitivity analyses that demonstrated that the associations were independent of BMPR2 and other rare pathogenic variants.

**Table 2: Associations of HLA-DPB1 alleles with the lead SNP rs2856830**

<table>
<thead>
<tr>
<th>Amino acid residues in HLA-DPB1 alleles</th>
<th>Frequencies by GWAS SNP rs2856830</th>
<th>q after FDR correction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T/T</td>
<td>T/C</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>DPB1*02:02†</td>
<td>6/7×10−³⁴</td>
<td>6/7×10−³⁴</td>
</tr>
<tr>
<td>DPB1*02:02†</td>
<td>4/7×10−³⁴</td>
<td>4/7×10−³⁴</td>
</tr>
<tr>
<td>DPB1*02:02†</td>
<td>1/7×10−³⁴</td>
<td>1/7×10−³⁴</td>
</tr>
<tr>
<td>DPB1*02:02†</td>
<td>2/7×10−³⁴</td>
<td>2/7×10−³⁴</td>
</tr>
</tbody>
</table>

Data are n/N (%) unless otherwise stated. FDR=false discovery rate. GWAS=genome-wide association study. SNP=single-nucleotide polymorphism. †Alleles and residues depleted in pulmonary arterial hypertension cases. ‡Alleles and residues enriched in pulmonary arterial hypertension cases.

The potential role of this modification in berylliosis. 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. A number of individual amino acid residues in the peptide-binding pockets of the HLA-DPB1 molecule affect its function and T-cell recognition, either by changing peptide antigen binding or the conformation of the peptide-binding groove. The HLA-DPB1 alleles 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.
Although the sequencing and array platforms used in this study might perform differently across the genome, the signals detected for each platform remained strong.

This study has some limitations. The majority of patients studied were prevalent cases, and the association with survival is only marginally significant and based on a relatively small sample size, from multiple studies with different ascertainment criteria. Thus, confirmation of the survival analysis in an independent sample of patients recruited at diagnosis would greatly increase confidence in this finding. Some variants displayed heterogeneity of effects between studies, which is most likely due to the limited sample size in the smaller genotyping studies. Variants in CBLN2 and other loci previously associated with pulmonary arterial hypertension were not replicated by this study, suggesting that these previous findings were either false positives or only relevant to the specific subpopulations studied.

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 to rare, deleterious mutations to the same disease. HLA-DPB1, and wider immune regulatory pathways, should be considered a priority for patient stratification and investigation of new treatments in pulmonary arterial hypertension. SOX17 is a key endothelial regulator and its dysfunction in pulmonary arterial hypertension might be more common than suggested by the occurrence of rare pathogenic variants in heritable cases.

Contributors
All authors were involved in study design, data collection, and constructive criticism of the manuscript. CJR, KB, MB, MHa, LSo, MG, MWP, JA, LH, IP, RK, SG, WCN, RCT, AAD, NWm, and MRW were involved in data interpretation and writing of the manuscript. Data were analysed by CJR, KB, MB, MHa, LSo, MG, MWP, CH, JA, AA, KBH, JHK, MK, AU, LH, EMs, and SG.

Declaration of interests
CJR reports personal fees from Actelion Pharmaceuticals. HG reports personal fees and non-financial support from Actelion, Novartis, Pfizer, and Bayer. KWh reports personal fees from Actelion, Novartis, and Pfizer; Gilead; and reports grants from NIH/NHLBI. HAG reports personal fees from Actelion, Bayer, GSK, Novartis, Pfizer, Bellerophon Pulse Technologies, and MSD; and grants from Deutsche Forshungsgemeinschaft. JSRG reports grants and personal fees from Actelion, Bayer, GSK, and MSD; personal fees from Arena, Bellerophon, Complexa, and Pfizer; and grants from United Therapeutics and Anscor. JH reports grants from NIH and the US Department of Veterans Affairs; personal fees from Janssen Pharmaceuticals, Shenzhen Rheumatic Disease Hospital (Shenzhen, China), Columbia University (NY, USA), and New York University (NY, USA); and options from Board of Directors service from Now Diagnostics. NSh reports financial support in the form of a sub-award from an NIH grant to the institution to support infrastructure. MGH is a consultant for Actelion, Bayer, Lilly, and United Therapeutics through University of Colorado (Aurora, CO, USA) contracts, and reports grant funding from the NIH and US Food and Drug Administration. DGK reports grants, personal fees, and non-financial support from Actelion, Bayer, GSK, and MSD. GK reports personal fees and non-financial support from Actelion, Bayer, GSK, MSD, Pfizer, AOP, Boehringer Ingelheim, Novartis, and Chiesi. TI reports consultancy fees from Actelion, Bayer, and Gilead. MN reports educational travel grants from MSD and GSK. HO reports grants from Bayer, Unither Pharmaceuticals, Actelion Pharmaceuticals, Roche, Boehringer Ingelheim, and Pfizer; personal fees from Gilead Sciences, EnvyPharma, and Nebu-Tec; and personal fees and non-financial support from Bayer, Unither Pharmaceuticals, Actelion Pharmaceuticals, Pfizer, Eli Lilly, Novartis, AstraZeneca, Boehringer Ingelheim, Chiesi, Menarini, MSD, and GSK. AJP reports grants and personal fees from Actelion and Bayer; personal fees from GSK and Pfizer; and grants from Gilead. JP-Z or her institution has received research or educational grants, and she has served on the advisory boards of Actelion, Merck, Bayer, and GSK. GS reports grants from NIH; grants and non-financial support from Actelion; and personal fees from Bayer and Gilead. RS is on the speakers’ bureau and has done funded clinical research or consulting for Actelion, Bayer, Gilead, Arena, GSK, and United Therapeutics. WS reports personal fees from United Therapeutics, Liquidia Technologies, and Bayer AG. MS reports grants from NIH/Cincinnati Children’s Hospital Medical Center, NIH, US National Science Foundation, Aires/Mast Therapeutics, Novartis; and personal fees from United Therapeutics, Gilead, Actelion, Bayer, St Jude Medical, Hovione, and Complexa. TT reports personal fees from Actelion and Gilead. FT reports grants from United Therapeutics, Gilead, Medtronic, GSK, and Pfizer; and GeNO; and personal fees from Actelion, Bayer Pharmaceuticals, SteadyMed, Reata, Arena, and Bellerophon. AKW reports grants from NIH, JWH’s personal fees from Actelion Pharmaceuticals. RJJW reports grants from NIH/NHLBI. SJW reports grants and personal fees from Actelion Pharmaceuticals and Bayer, and personal fees from MSD and GSK. WCN reports grants from NIH. All other authors declare no competing interests.

Acknowledgments
We gratefully acknowledge the participation of patients recruited to the UK National Institute of Health Research Biomedical Centre (NIHR BRC) Rare Diseases study. We thank the NIHR/BRC staff and coordination teams at the University of Cambridge (Cambridge, UK), and the research nurses and coordinators at the specialist pulmonary hypertension centres involved in this study. We are also grateful to Jenny Thomson and Carolyn Langman for invaluable assistance in patient recruitment for the British Heart Foundation Pulmonary Arterial Hypertension (BHFPAH) study. The UK National Cohort of Idiopathic and Heritable PAH is supported by the NIHR BRC; the BHF (SP/12/12/29836); the BHF Cambridge Centre of Cardiovascular Research Excellence; the UK Medical Research Council (MR/R005919/1); the Dinosaur Trust; BHF Programme grants to CTR (RG/08/006/25502); NWm (RG/11/11/38007), and MRW (RG/10/16/28575). We also gratefully acknowledge the participation of patients recruited to the US National Institutes of Health/National Heart, Lung, and Blood Institute (NIH/NHLBI)-sponsored National Biological Sample and Data Repository for PAH (also known as PAH Biobank). We thank our physicians, research nurses, and coordinators at the 38 pulmonary hypertension centres across the USA involved in the PAH Biobank. Vanderbilt University Medical Center’s BioVU projects are supported by numerous sources: institutional funding, private agencies, and federal grants. These include the list of investigators and enrolment centres.

For a list of investigators and enrolment centres see http://www.pahbiobank.org
NIH-funded Shared Instrumentation Grant S10RR023541; and CTSA grants UL1TR002245, UL1TR004045, and UL1TR002495. The genotyping of the VESPA samples was supported by RC2GM092626. The authors acknowledge use of BRC Core Facilities provided by financial support from the UK Department of Health via the NIHR Comprehensive Biomedical Research Centre award to Cambridge Biomedical Research Centre, Imperial College Healthcare NHS Trust, and 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 NIHR Senior Investigator. CH is a NIHR Rare Disease Translational Research Collaboration Clinical PhD Fellow. CJR is supported by a BHF Intermediate Basic Science Research fellowship (FS/15/59/31839). LH and JA are the recipients of ERS and joint ERS/EMBO Long Disease Translational Research Collaboration Clinical PhD Fellow. CJR is supported by the Wellcome Trust Institutional Strategic Support Fund (204809/Z/16/Z) awarded to St George’s, University of London (London, UK). IP is supported by the Wellcome Trust (WT209915), and the EU H2020 programme (DYNAnheart, project number 635995). Funding for the PAH Biobank is provided by NIH/NHLBI (H1L05331). WCN and MWP are supported by NIH/NHLBI (H1L05331). JHK receives support from the American Heart Association (16SDG29090005) and the American College of Clinical Pharmacy Research Institute (Futures Grant). AAD receives support from the Dutch Heart Foundation, Dutch Federation of Sciences (Hertie Program; LabEx LERMIT, ANR-14-LABX-0145; and RHU BIO-ART LUNG 2020, ANR-15-R1US-0002). MRW and H-AG receive funding from German Research Foundation (DFG) SFB1241, project A09. The popgen 2.0 network is supported by a grant from the German Ministry for Education and Research (01E1Y0103). We thank all the patients and their families who contributed to this research and the UK Pulmonary Hypertension Association for their support.

References