Shared and Distinct Genomics of Chronic Thromboembolic Pulmonary Hypertension and Pulmonary Embolism

Dr James Liley*, PhD, Durham University, Durham, UK
Dr Michael Newnham*, PhD, Institute of applied health research, Birmingham, UK
Dr Marta Bleda*, PhD, Dept of Medicine, University of Cambridge, UK
Dr Katherine Bunclark, MD ChB, Royal Papworth Hospital, Cambridge, UK
Dr William Auger, MD, University of California San Diego, US
Dr Joan Albert Barbera, PhD, Hospital Clinic-IDIBAPS-CIBERES, University of Barcelona, Spain
Prof. Harm Bogaard, PhD, Amsterdam UMC, Netherlands
Prof. Marion Delcroix, PhD UZ Leuven, Belgium
Dr Timothy M. Fernandes, MD, University of California San Diego
Prof. Luke Howard, PhD, Hammersmith Hospital, London, UK
Mr David Jenkins, MS, Royal Papworth Hospital, Cambridge, UK
Prof. Irene Lang, PhD, AKH-Vienna, Medical University of Vienna, Austria
Dr Eckhard Mayer, PhD, Kerckhoff Clinic, Bad Nauheim Germany
Dr Chris Rhodes, PhD, Imperial College London, London, UK
Prof. Michael Simpson, PhD, King's College London, UK
Dr Laura Southgate, PhD, St George's, University of London, UK
Prof. Richard Trembath, FRCP, King's College London, UK
Dr John Wharton, PhD, Imperial College London, London, UK
Prof. Martin R Wilkins, MD DSc , Imperial College London, London, UK
Dr Stefan Gräf, PhD, Dept of Medicine, University of Cambridge, UK
Prof. Nicholas Morrell, PhD, Dept of Medicine, University of Cambridge, UK
Dr Joanna Pepke Zaba^, PhD, Royal Papworth Hospital, Cambridge, UK
Dr Mark Toshner^†, MD, Dept of Medicine, University of Cambridge, UK

*: Joint first author
^: Joint senior author
†: Corresponding author (ph: +44 1223 638000; em: mrt34@medschl.cam.ac.uk)

Contributors statement: Author contributions were as follows (using CRediT taxonomy; http://credit.niso.org/):

JL: Data Curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing original, Writing review and editing
MN: Data Curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing review and editing

MB: Data Curation, Methodology, Software, Resources
KB: Data Curation, writing review and editing
WA: Resources
JAB: Resources
HB: Resources
MD: Resources
TF: Resources
SG: Data Curation, Investigation, Software
LH: Resources
DJ: Resources
IL: Resources
EM: Resources
CR: Data Curation, Investigation, Software
MS: Resources
LS: Resources
RT: Conceptualization, Resources, Writing review and editing
JW: Resources
MW: Resources
NM: Funding acquisition, Project administration, Supervision
JPZ: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Visualization
MT: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Visualization, Writing original, Writing review and editing.

Direct data access and verification were performed by JL and MN.

Declaration of grants: This study was supported by the NIHR cardiorespiratory BRC and an unrestricted grant from Bayer Pharmaceuticals.

Descriptor number: 9.35 (Pulmonary Hypertension: Clinical-Diagnosis/Pathogenesis/Outcome)

At a Glance Commentary
Current Scientific Knowledge on the Subject: Chronic Thromboembolic Pulmonary Hypertension (CTEPH) is a respiratory illness characterised by many potential pathophysiological processes including formation and non-resolution of thrombus, dysregulated inflammation, angiogenesis, vasculopathy and right heart failure. There are no previous investigations of its genetic causes on a genome-wide scale and by understanding the genetic contributors to disease it is likely we will better understand the pathophysiology.

What This Study Adds to the Field: We conduct the first genome-wide association study on CTEPH. We find several regional genetic associations, and partial sharing of genetic associations

with pulmonary embolism. We do not find evidence of shared genetics with idiopathic pulmonary arterial hypertension.

Some of the results of these studies have been previously reported in the form of a preprint (medRxiv, 5 Jun 2023, www.medrxiv.org/content/10.1101/2023.05.30.23290666v1).

This article has an online data supplement, which is accessible at the Supplements tab.

Abstract

Rationale

Chronic Thromboembolic Pulmonary Hypertension involves formation and non-resolution of thrombus, dysregulated inflammation, angiogenesis and the development of a small vessel vasculopathy.

Objectives

We aimed to establish the genetic basis of chronic thromboembolic pulmonary hypertension to gain insight into its pathophysiological contributors.

Methods

We conducted a genome-wide association study on 1907 European cases and 10363 European controls. We co-analysed our results with existing results from genome-wide association studies on deep vein thrombosis, pulmonary embolism and idiopathic pulmonary arterial hypertension.

Measurements and Main Results

Our primary association study revealed genetic associations at the ABO, FGG, F11, MYH7B, and HLA-DRA loci. Through our co-analysis we demonstrate further associations with chronic thromboembolic pulmonary hypertension at the F2, TSPAN15, SLC44A2 and F5 loci but find no statistically significant associations shared with idiopathic pulmonary arterial hypertension.

Conclusions

Chronic thromboembolic pulmonary hypertension is a partially heritable polygenic disease, with related though distinct genetic associations to pulmonary embolism and to deep vein thrombosis.

(162 words)

Key words: Genome-wide association study, Pulmonary Arterial Hypertension, Venous Thromboembolism

Introduction

Chronic thromboembolic pulmonary hypertension (CTEPH) is characterised by the organisation and fibrosis of thromboembolic material leading to the obstruction of proximal pulmonary arteries which, together with a secondary small-vessel vasculopathy, results in pulmonary hypertension and subsequent right heart failure.

CTEPH is conventionally considered to result from a process of disordered thrombus resolution following one or more episodes of acute pulmonary embolism (PE) (1). The pathobiology of thrombus non-resolution following acute PE however remains poorly understood but likely arises from complex interactions between mediators of the coagulation cascade, angiogenesis, platelet function and inflammation in association with host factors. Large volume acute PEs, idiopathic presentation, and PE recurrence are associated with a risk for CTEPH development (2). Inefficient anticoagulation may also trigger thrombus formation (3). These factors however do not serve to explain the development of CTEPH in most patients. Furthermore, up to 25 % of CTEPH patients do not have a history of antecedent PE. The ability to identify abnormalities in coagulation/fibrinolysis pathways in CTEPH patients is compounded by their treatment with therapeutic anticoagulation and lack of a good animal model of CTEPH.

Genetic studies in CTEPH have the potential to inform our understanding of disease pathophysiology, but have thus far been hampered by the challenge of assembling cases in rare diseases. A European prospective registry found an increased CTEPH risk in non-O blood groups, in a similar pattern to DVT and PE (4), indicating a genetic association with the disease at this

locus. This differential risk with ABO is also seen in overall risk of PE and other clotting disorders. To our knowledge, no other genetic associations with CTEPH have been confirmed at genome-wide significance ($P< 5 \times 10\text{-}8$).

The genetic basis of a comparator disease, Idiopathic Pulmonary Arterial Hypertension (IPAH) has been much more systematically explored. Heterozygous germline mutations in *BMPR2* are found in 10 – 20 % of individuals with IPAH alongside rarer sequence variants including *SMAD9*, *ACVRL1*, *ENG*, *KCNK3* and *TBX4* (5). A more recent GWAS study has also identified common variants contributing to IPAH aetiology and clinical course (6).

An improved understanding of the genetic basis of CTEPH has the potential to not only inform disease aetiopathogenesis but in quantification of CTEPH risk, preventative strategies and treatment options. An evaluation of CTEPH genome-wide associations is therefore warranted. Co-analysis with existing GWAS in PE and DVT aims to improve both discovery and the interpretation of results in comparison to other venous thromboembolic phenotypes. Given well-known genetic drivers to the development of IPAH and its shared pathobiological features of vascular remodelling, inflammation and dysregulated angiogenesis with CTEPH, genetic associations between CTEPH and IPAH were also explored.

Methods

Study samples and participants

The study was approved by the regional ethics committee (REC no. 08/H0802/32 and 08/H0304/56). All study participants provided written informed consent from their respective institutions.

GWAS on CTEPH

We conducted a two-stage design: a discovery study including only UK samples, and a replication stage using non-UK cases and a mixture of non-UK and UK controls.

CTEPH was diagnosed in accordance with international guidelines (7). All patients were diagnosed through internationally accredited specialist centres with multimodal imaging and invasive haemodynamics. The UK cases, in addition to review at nationally designated tertiary centre MDTs are additionally reviewed at the national CTEPH MDT where all cases are discussed by a multi-disciplinary team of surgeons, cardiologists, radiologists and PH specialists. Demographics of CTEPH samples are reported in Table S1 (Supplementary Methods). Consistent with historical published cohorts (4), PEA treatment was most common (62.4%), balloon pulmonary angioplasty (BPA) (2.4%), both PEA/BPA (1.6%) and medical therapy only (33.6%). Controls were sourced randomly from the population (without requiring absence of thromboembolic phenotypes). Samples in the discovery phase were genotyped on one of four platforms: the Illumina HumanOmniExpress Exome-8 v1.2 BeadChip (1555 cases, 1693 controls); the Illumina HumanOmniExpressExome-8 v1.6 BeadChip (372 cases, 12 controls); the Affymetrix Axiom Genome-Wide CEU 1 Array (541 cases, 5984 controls, including re-genotyping of 1533 controls genotyped on the Illumina HumanOmniExpressExome-8 v1.2 BeadChip) and the Affymetrix UK Biobank Axiom array (6717 controls).

We performed sample- and SNP- wise quality control on our dataset (8) and excluded cases of non-European ancestry using principal components generated using the 1000 genomes project. We imputed all genotypes to whole-genome cover using the Haplotype Reference Consortium panel on the Sanger imputation server (9,10), separating samples by genotyping platform, and we included SNPs with an INFO score of at least 0.5 across all genotyping platforms used in the study. The INFO score is a measure of imputation accuracy, interpretable as a proportion: a score of 1 indicates full knowledge of the SNP in all samples, 0 indicates no knowledge of the SNP, and other values indicate knowledge of the SNP equivalent to full knowledge in that proportion of samples (11). Full details of quality control procedures are given in the Supplementary Methods.

We separated the discovery cohort into two groups by genotyping platform (Affymetrix or Illumina) and analysed each separately. In each cohort, we used a logistic regression with ten principal component covariates to generate association statistics, and corrected results for residual genomic inflation (12). Since each analysis involved separate samples, we combined results across platforms using a routine p-value meta-analysis using Fisher's method accounting for effect directions.

Co-analysis with DVT and PE

In order to enhance our power to detect CTEPH associations, we co-analysed our p-values from the CTEPH meta-analysis with p-values derived from GWAS on self-reported PE and DVT drawn from the UK Biobank (13) (GWAS round 2; self-reported DVT (code 20002_1094) and self-reported PE (code 20002_1093)). Details of the co-analysis are given in the Supplementary

Methods. In short, the output of each co-analysis is a set of p-values for CTEPH 'adjusted' for the overall genetic similarity between CTEPH and the second disease (14), which we call 'V-values'. We also performed an analysis using results from DVT in place of results from PE, but found the results from the two analyses were very similar, so we focus principally on the analysis of PE.

CTEPH GWAS associations

Noting that our replication cohort was analysed at genome-wide SNPs, we defined genetic associations at three tiers of significance, all of which generally correspond to a genome-wide significance of overall p-value $< 5 \times 10^{-8}$ with varying levels of evidence in the discovery and replication sub-cohorts. The first tier required $P < 5 \times 10^{-6}$ in the combined discovery cohort, $P < 5 \times 10^{-3}$ in the replication cohort, and $P < 5 \times 10^{-8}$ in the combined meta-analysis, with consistent directions of effect across the two sub-analyses in the discovery study and in the replication study. The second tier, designed to ensure nominal association in each cohort and overall genome-wide significance, required a nominal association of $P < 5 \times 10^{-2}$ in discovery and replication cohorts and $P < 5 \times 10^{-8}$ in the overall meta-analysis, again with consistent directions of effect. The 'adjusted' p-values allowed a comparison of evidence for association using cFDR in a similar way to a comparison using meta-analysed p-values, and hence we defined a third tier of association requiring a p-value of 5×10^{-8} in either the overall meta-analysis or the 'adjusted' sets of p-values derived from leverage of the CTEPH summary statistics on summary statistics for PE, along with consistent directions of effect in discovery and replication cohorts. All p-value thresholds used in 'tier' definitions were chosen prior to observing the data.

There was a distribution of cases and controls across genotyping batches which could enable confounding batch effects, and differing sources of cases and controls in the replication cohort necessitated across-platform comparisons and imperfect geographical matching resulting in high inflation in association statistics. We also noted recent work indicating that blood-bank sourced control samples may have differing distributions of ABO blood groups to the general population, potentially biasing association statistics at that locus. In the Supplementary material, we analyse allele frequencies across batches and cohorts directly, and thus demonstrate that these confounding effects are unlikely to drive our positive associations.

The study design is outlined in Figure 1.

Genetic overlap with IPAH and PE

As a cause of pulmonary arterial hypertension, we considered the possibility that CTEPH shares pathology with idiopathic pulmonary hypertension (IPAH). We firstly assessed whether our findings had any associations in common with a recent GWAS on IPAH (6). To assess for genome-scale similarity in genetic basis between IPAH and CTEPH, we used linkage disequilibrium-score regression (LDSC) (15) to estimate genetic correlation ρ_g between the two traits, which measures the degree of shared genetic basis. We also estimated genetic correlation between IPAH and PE (using the summary statistics for PE used in the co-analysis with CTEPH) for comparison.

Diseases with identical genetic bases have genetic correlation 1, and diseases with completely independent genetic bases have genetic correlation 0. If IPAH and CTEPH each occurred as a consequence of some identical underlying cause, we would expect them to have genetic correlation 1, whereas if they were caused by completely independent pathological processes, the genetic correlation would be 0 (and likewise for PE and CTEPH).

Comparison of DVT, PE, and CTEPH

We would expect to see a slight difference in observed effect sizes between CTEPH, PE, and DVT at any given variant due to random variation across studies. For each of our CTEPH-associated variants, we assessed whether the observed effect sizes in CTEPH and PE were consistent with random variation if CTEPH and PE/DVT had identical underlying genetic causes. Specifically, we considered a null hypothesis that the two diseases have identical effect sizes for all SNPs, and assessed the probability of seeing large differences in effect sizes between CTEPH and PE. Our approach is detailed in the Supplementary Methods, section 'Differential effect sizes between CTEPH, DVT and PE'.

Results

GWAS on CTEPH

After quality control (see Figure 1), our dataset consisted of 1146 cases and 5498 controls in the discovery cohort, and 761 cases and 4865 controls in the replication cohort. A total of 4655481 SNPs passed quality control and were included in the final analysis. At tier 2 significance, the study had approximately 80 % power to detect an odds ratio of 1.3 for a SNP of minor allele frequency (MAF) 0.25, or an odds ratio of 1.7 for a SNP of MAF 0.05. Further details of power for tier 1 and 2 significance are shown in Supplementary Figures 1,2,3. Minimal detectable effect sizes at tier 3 significance are more complex; see Supplementary Methods.

We computed genomic inflation factors (λ) which measure the overall distribution of p-values, and should be close to 1. These were 0.95 in the discovery cohort and 1.21 in the replication cohort ($\lambda_{1000} = 1.16$; see Supplementary Methods (16)), suggesting that p-values in the replication cohort were overall lower than expected. We were not able to reduce inflation in the replication cohort by inclusion of further covariates or by use of linear mixed models, and concluded that the degree of inflation was inevitable given the imperfect geographical matching between cases and controls in the replication dataset. We corrected P-values in the replication cohort for this residual inflation (12), thereby avoiding false positives arising from this inflation.

A Manhattan plot of meta-analysed p-values is shown in Figure 2. Manhattan plots for the discovery and replication cohorts alone are shown in Supplementary Figures 4 and 5. Two regional associations (*FGG* and *ABO*) were found at tier 1 significance, and a further association (*MYH7B*) at tier 2 significance. Two further regions (*F11* and *HLA-DRA*) reached tier 3 significance on the basis of meta-analysis p-value. Results for all SNPs reaching genome-wide significance are shown in Table 1.

Co-analysis with DVT and PE

The co-analyses with PE demonstrated four further associations at tier 3 genome-wide significance (*F2*, *TSPAN15*, *SLC44A2* and *F5/NME7*). Plots of z-scores from the three analysis showed evidence of widespread sharing of associations with DVT and PE, but differential effect sizes between phenotypes (Figures 3, 4a, 4b).

CTEPH GWAS associations

FGG and ABO (tier 1)

We found an association with peak SNP rs7659024, around 4kb downstream of the *FGG* gene. The *FGG* gene codes for the gamma chain of the fibrinogen protein, a precursor for fibrin, the principal non-cellular component of blood clots. Polymorphisms in *FGG* are well-known to be associated with DVT (17). The variant is also 9kb downstream of the *FGA* gene, which codes for the alpha chain of the fibrinogen complex. The strongest association (by p-value) in CTEPH was rs687289 in the *ABO* gene, which determines ABO blood group. This locus is also known to be associated with DVT (17). Patients with non-O blood groups are at higher risk of CTEPH (18).

HLA-DRA, TSPAN15, F2, SLC44A2, F11 (tier 2/3)

An association (variant rs17202899) was found at tier 2 significance in the *HLA-DRA* gene, which to our knowledge has not been shown to be associated with DVT or PE. The variant is strongly associated with multiple autoimmune conditions, including type 1 diabetes (19), systemic lupus erythamatosis (20), and multiple sclerosis (21). Variant rs78677622, on chromosome 10, is an intron variant 10kb upstream of *TSPAN15*, which is known to be associated with DVT (17). Variant rs149903077 on chromosome 11 is an intron variant in the *DGKZ* gene, but is likely to correspond to an association of CTEPH with the *F2* gene, from which it is 390kb upstream.

Variant rs2288904 on chromosome 19 is a missense variant in the *SLC44A2* gene, variants in which are associated with DVT (17). Variant rs2289252 on chromosome 4 is an intron in *F11*, which codes for coagulation factor 11, variants in which are DVT-associated (17). Variant rs745849 on chromosome 20 is an intron in the *MYH7B* gene, for which there are nearby associations with DVT (17), though the variant itself does not reach genome-wide significance for either DVT or PE. The variant is associated with human height and ease of tanning (13). Finally, we found an association at rs796548658 on chromosome 1 at tier 3 significance. Although the peak variant is an intron in the *NME7* gene, it is likely to represent an association of CTEPH with the *F5* gene, which is strongly associated with DVT (17). This association is notable for the relatively small effect size in CTEPH.

Genetic overlap with IPAH and PE

We did not find genetic evidence of shared pathology between CTEPH and IPAH. No shared genome-wide associations are evident between our findings and a recent GWAS on IPAH (6). The

observed genetic correlation between IPAH and CTEPH was not significantly different from 0 (est. ρ_g -0.37, standard error 0.38; p-value 0.3 against H^0: ρ_g =0), but was significantly different from 1. The genetic correlation between CTEPH and self-reported PE was significantly above zero, indicating shared genetic architecture (est. ρ_g 1.07, standard error 0.44; p-value 0.014 against H^0: ρ_g =0) but not significantly different from 1, indicating that identical genetic architecture could not be ruled out with this analysis. We concluded that, on the basis of genetic correlation, CTEPH is more similar to PE than to IPAH.

Comparison of DVT, PE and CTEPH

We found a substantial difference in observed effect sizes of variants in the *F5* gene between DVT, PE and CTEPH. We also noted that the *HLA-DRA* and *MYH7B* variants are not known to be associated with DVT or PE.

For both the *F5* and *HLA-DRA* regions, the probability of observing effect sizes at least as different as those seen under a null hypothesis of identical true effect sizes between CTEPH and DVT or PE was < 0.05, using a Bonferroni correction over all variants reaching genome-wide significance for either disease. This is shown in Figure 2.

If the observed odds ratio of the peak SNP for *F5* in DVT (or SNPs in close linkage disequilibrium) were equal to the true odds ratio in CTEPH, our study would have had > 99 % power to detect an association at tier 1 significance. Likewise, if the observed odds ratio for the *HLA-DRA* association

found in our study corresponded to the true effect size in DVT, then the study on DVT would have > 99 % power to detect the association.

We conclude that the effect size of causal variants in *F5* and *HLA-DRA* in CTEPH is different to the effect of those variants in DVT and in PE. We cannot conclude that the effect size of the causal variant in *MYH7B* differs between CTEPH and DVT or PE.

Discussion

We report the first GWAS in CTEPH, comprising a multinational study on a cohort with sufficient power to find common-variant associations of reasonable size. In general, the associations we find are consistent with a shared genetic associations of venous thromboembolism, although we identify important differences in genetic architecture to PE and DVT. CTEPH is a partially heritable polygenic disease: it does not develop randomly amongst patients with pulmonary emboli, nor is development of CTEPH governed entirely by environmental triggers: if this were the case, all genetic associations for both diseases would have identical size (and variants in *F5* and *HLA-DRA* do not). Historical debate has for decades posited that the similarity in pathophysiology, presence of thrombus in some cases of IPAH and absence of index PE in up to a quarter of cases of CTEPH suggests that CTEPH is not simply the consequence of disordered thrombus fibrinolysis but instead a potential overlap of distal cases of CTEPH and IPAH (22). Our work supports evidence that CTEPH and IPAH are distinct and that despite similar vascular remodelling, inflammation and involvement of dysregulated angiogenesis, the underlying aetiologies are different. This is

consistent with work examining CTEPH cohort demographics and phenotypes (23). Genetic associations of underlying susceptibility to vascular remodelling or pulmonary hypertension do not appear to be major drivers of CTEPH in this study.

The smaller effect sizes of variants in *F5* in CTEPH may be an example of index-event bias (24), a phenomenon in which the effect of a risk factor is underestimated due to the dominance of other factors. Specifically, the large effect of the *F5* Leiden variant in causing thromboembolic disease may paradoxically mean that patients with PE carrying a *F5* Leiden variant have a *lower* burden of other genetic and environmental risk factors for CTEPH, and are hence less likely to develop CTEPH following PE than those without the variant. This could also account for the apparently smaller relative effect of *F5* variants in PE than in DVT seen in Figures 4a, 4b.

To our knowledge, *HLA-DRA* has not previously been shown to be associated with either DVT or PE, though variants in the locus have been associated with a range of immune-related phenotypes (25), likely reflecting a role in the processing and presentation of Major Histocompatibility Complex molecules. Increased CTEPH risk has long been linked with underlying autoimmune and haematological disorders (4). In addition, a variety of inflammatory cytokines are elevated in CTEPH and correlate with pulmonary artery inflammatory cell infiltration and CTEPH severity (26).

An important shortcoming of our work is imperfect geographic matching between cases and controls in the replication cohort, resulting in a degree of inflation in summary statistics. This is unavoidable with our current dataset. To manage this, we forcibly rescaled χ^2 statistics to remove the inflation (see Supplementary methods). We have also not adjusted for age and sex though the

need to do this in a disease with an incidence of less than 1.0% has previously been demonstrated to be minimal (27). We are also not powered to perform sub-analyses of patients with and without known VTE and future work should consider this. Reassuringly, our overall findings are not unexpected. A further shortcoming of our work is its restriction to individuals of western or central European ancestry, and further investigation into the genetic architecture of CTEPH in other ethnicities is warranted. Our study additionally cannot shed any light on the contribution of rare genetic variants to pathophysiology.

In summary we provide the first large scale GWAS in this rare disease and we demonstrate for the first time the genetic architecture of a complex condition leveraged against comparator datasets. These analyses establish the primacy of dysregulated thrombosis/fibrinolysis in aetiology and extend our understanding of the possible contribution of additional pathophysiological mechanisms including inflammation. CTEPH did not share any genetic associations with IPAH further confirming that despite significant shared pathophysiology these conditions have divergent aetiology.

Acknowledgements

We would like to acknowledge the UK tertiary pulmonary hypertension network and the patients who enabled this work. This work was supported by the UK National Institute for Health Research Cardiorespiratory Biomedical Research Council and an unrestricted grant from Bayer Pharmaceuticals. CJR is supported by BHF Basic Science Research fellowships (FS/15/59/31839 & FS/SBSRF/21/31025) and Academy of Medical Sciences Springboard fellowship (SBF004\1095).

Role of the funding sources

Funders had no role in study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit this paper for publication

Table 1. Genome-wide significant regions for CTEPH. Positions shown are GRCh37 and minor allele frequency (MAF) is in controls. Overall odds ratios are estimated from meta-analysis p-values and overall sample sizes. P and P(PE) refer to CTEPH meta-analysis p-values and p-values for for PE (derived from a separate GWAS) respectively. V shows v-values, which can be thought of as p-values for CTEPH adjusted to account for overall genetic similarity with PE.

Chr.	BP	RSID	MAF	OR	P	P(PE)	V	Tier	Gene
9	136137106	rs687289	0.340	1.80	6.8×10^{-27}	5.4×10^{-35}	6.8×10^{-27}	1	*ABO*
4	155520930	rs7659024	0.250	1.60	9.0×10^{-17}	4.7×10^{-14}	3.5×10^{-23}	1	*FGG*
4	187207381	rs2289252	0.400	1.30	7.7×10^{-9}	3.3×10^{-16}	3.0×10^{-15}	3	*F11*
20	33572178	rs745849	0.430	0.76	3.4×10^{-8}	7.5×10^{-2}	9.7×10^{-8}	2	*MYH7B*
6	32434481	rs17202899	0.100	1.60	4.7×10^{-8}	5.5×10^{-1}	9.2×10^{-7}	3	*HLA-DRA*
11	46349696	rs149903077	0.013	3.30	7.9×10^{-8}	1.6×10^{-12}	3.0×10^{-14}	3	*F2*
10	71196698	rs78677622	0.130	0.70	1.5×10^{-6}	8.2×10^{-12}	6.0×10^{-13}	3	*TSPAN15*
19	10742170	rs2288904	0.210	0.76	8.3×10^{-6}	2.2×10^{-7}	9.3×10^{-11}	3	*SLC44A2*
1	169272453	rs796548658	0.039	1.60	1.5×10^{-4}	3.9×10^{-18}	7.2×10^{-10}	3	*F5/NME7*

Figure Legends

Figure 1. Flow chart for study design. PCA: excluded due to inferred ancestry based on PCA. Rel/dup: excluded due to being closely related to another sample, or a duplicate of another sample. Heterozygosity: excluded due to abnormal heterozygosity rate. Missingness: excluded due to high missingness in genotype, or otherwise unusable genotype. Full details are given in the supplementary methods. In short, we recruited cases and controls from a variety of centres around the UK and Europe, including existing controls. Our discovery cohort consisted of UK and USA samples; our replicatioh cohort of European samples. Exclusions were applied sequentially: for instance, some samples which would have been excluded for abnormal heterozygosity rate in the replication cohort were already excluded by PCA.

Figure 2: Manhattan plot of $-\log_{10}(p)$-values derived from meta-analysis of discovery and replication cohorts. Points higher up correspond to variants more strongly associated with CTEPH. Variants reaching genome-wide significance ($P_{CTEPH} < 5 \times 10^{-8}$) are marked in black, and variants discovered using co-analysis with PE are marked in blue, both labelled with the likely associated gene. The black horizontal line denotes genome-wide significance ($p = 5 \times 10^{-8}$). Values of $-\log_{10}(p)$ larger than 16 are truncated to 16.

Figure 3. Back-to-back Manhattan plots for CTEPH and PE. The distance from the middle line corresponds to $-\log_{10}(p)$ values; points further from the middle line correspond to variants more strongly associated with CTEPH (upwards) or PE (downwards). Values of $-\log_{10}(p)$ larger than 16 are truncated to 16. Peak variants as in Table 1 are marked with the likely corresponding gene. There is substantial sharing between associations with CTEPH and with PE. Genome-wide associations (p<5 x 10-8) are marked in red. Additional associations discovered through leverage (cFDR) are marked in blue.

Figure 4a, 4b: Z-scores for CTEPH against Z-scores for DVT (left) and PE (right). Each point corresponds to a SNP, with colour and shape corresponding to chromosome as per the legend. Z-score pairs close to the origin are excluded. Points higher up correspond to variants more associated with DVT/PE, and points further to the right correspond to variants more associated with CTEPH. Potential genes (F11, F5, HLA-DRA, etc.) are labelled for some SNPs. The area to the right of the dotted black line is a rejection region based on a CTEPH genome-wide significance threshold of $P_{CTEPH} < 5 \times 10^{-8}$. The area to the right of the solid black line is a rejection region based on the levered analysis using conditional false discovery rates, equivalent to a V-value $< 5 \times 10^{-8}$. The solid red line shows the expected position of Z-score pairs if SNP effect sizes for CTEPH and DVT/PE were identical. If effect sizes were identical for all SNPs, the probability of any of the points corresponding to the ≈200 SNPs reaching genome-wide significance for CTEPH or DVT/PE falling outside the dashed red lines is < 0.05. We see that peak SNPs for *F5* and *HLA-DRA* fall outside the dashed lines in both plots.

References

1. Kim NH, Delcroix M, Jenkins DP, Channick R, Dartevelle P, Jansa P, et al. Chronic thromboembolic pulmonary hypertension. Journal of the American College of Cardiology. 2013;62(25S):D92–9.

2. Ende-Verhaar YM, Cannegieter SC, Noordegraaf AV, Delcroix M, Pruszczyk P, Mairuhu AT, et al. Incidence of chronic thromboembolic pulmonary hypertension after acute pulmonary embolism: a contemporary view of the published literature. European Respiratory Journal. 2017;49(2).

3. Yang S, Yang Y, Zhai Z, Kuang T, Gong J, Zhang S, et al. Incidence and risk factors of chronic thromboembolic pulmonary hypertension in patients after acute pulmonary embolism. Journal of thoracic disease. 2015;7(11):1927.

4. Pepke-Zaba J, Delcroix M, Lang I, Mayer E, Jansa P, Ambroz D, et al. Chronic thromboembolic pulmonary hypertension (CTEPH) results from an international prospective registry. Circulation. 2011;124(18):1973–81.

5. Morrell NW, Aldred MA, Chung WK, Elliott CG, Nichols WC, Soubrier F, et al. Genetics and genomics of pulmonary arterial hypertension. European Respiratory Journal. 2019;53(1).

6. Rhodes CJ, Batai K, Bleda M, Haimel M, Southgate L, Germain M, et al. Genetic determinants of risk in pulmonary arterial hypertension: international genome-wide association studies and meta-analysis. The Lancet Respiratory Medicine. 2019;7(3):227–38.

7. Humbert M, Kovacs G, Hoeper MM, Badagliacca R, Berger RMF, Brida M, et al. 2022 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension. Eur Respir J. 2023 Jan;61(1):2200879.

8. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nature protocols. 2010;5(9):1564.

9. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nature genetics. 2016;48(11):1443.

10. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nature genetics. 2016;48(10):1279.

11. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010 Jul;11(7):499–511.

12. Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. Theoretical Population Biology. 2001 Nov;60:155–66.

13. Neale BM. UK Biobank GWAS results. [cited 2019 May 1]. UK Biobank GWAS results. Available from: http://www.nealelab.is/uk-biobank

14. Liley J, Wallace C. Accurate error control in high dimensional association testing using conditional false discovery rates. Under second review; PDF on request. 2018;414318.

15. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. Nature genetics. 2015;47(11):1236–41.

16. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, et al. Assessing the impact of population stratification on genetic association studies. Nature genetics. 2004;36(4):388.

17. Germain M, Chasman DI, De Haan H, Tang W, Lindström S, Weng LC, et al. Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism. The American Journal of Human Genetics. 2015;96(4):532–42.

18. Bonderman D, Lang IM. Risk factors for chronic thromboembolic pulmonary hypertension. In: Textbook of Pulmonary Vascular Disease. Springer; 2011. p. 1253–9.

19. Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner KM, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. Nature. 2023 Jan;613(7944):508–18.

20. Bentham J, Morris DL, Graham DSC, Pinder CL, Tombleson P, Behrens TW, et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat Genet. 2015 Dec;47(12):1457–64.

21. Andlauer TFM, Buck D, Antony G, Bayas A, Bechmann L, Berthele A, et al. Novel multiple sclerosis susceptibility loci implicated in epigenetic regulation. Sci Adv. 2016 Jun;2(6):e1501678.

22. Egermayer P, Peacock AJ. Is pulmonary embolism a common cause of chronic pulmonary hypertension? Limitations of the embolic hypothesis. European Respiratory Journal. 2000;15(3):440–8.

23. Suntharalingam J, Machado RD, Sharples LD, Toshner MR, Sheares KK, Hughes RJ, et al. Demographic features, BMPR2 status and outcomes in distal chronic thromboembolic pulmonary hypertension. Thorax. 2007;62(7):617–22.

24. Dudbridge F, Allen RJ, Sheehan NA, Schmidt AF, Lee JC, Jenkins RG, et al. Adjustment for index event bias in genome-wide association studies of subsequent events. Nature communications. 2019;10(1):1561.

25. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14000 cases of seven common diseases and 3000 shared controls. Nature. 2007 Jun;447(7145):661–78.

26. Zabini D, Heinemann A, Foris V, Nagaraj C, Nierlich P, Bálint Z, et al. Comprehensive analysis of inflammatory markers in chronic thromboembolic pulmonary hypertension patients. European Respiratory Journal. 2014 Oct 1;44(4):951–62.

27. Galiè N, Humbert M, Vachiery JL, Gibbs S, Lang I, Torbicki A, 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). European Respiratory Journal. 2015 Oct 1;46(4):903–75.

28. Smith K, Lyons P, Peters J, Alberici F, Liley J, Coulson R, et al. Genome-wide association study of eosinophilic granulomatosis with polyangiitis reveals genomic loci stratified by ANCA status. 2019;

29. Consortium 1000 Genomes Project. A global reference for human genetic variation. Vol. 526, Nature. Nature Publishing Group; 2015. p. 68.

30. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012;28(24):3326–8.

31. Lo PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Chasman DI, et al. Efficient Bayesian mixed model analysis increases association power in large cohorts. Nature Genetics. 2015 Feb;47(3):284–90.

32. Andreasson OA, Harbo HF, Wang Y, Thompson WK, Schork AJ, Mattingsdal M, et al. Genetic pleiotropy between multiple sclerosis and schizophrenia but not bipolar disorder: differential involvement of immune-related gene loci. Molecular psychiatry. 2014;1–8.

33. Liley J, Wallace C. A Pleiotropy-Informed Bayesian False Discovery Rate adapted to a Shared Control Design Finds New Disease Associations From GWAS Summary Statistics. PLOS Genetics. 2015;

34. Pe'er I, Yelensky R, Altshuler D, Daly MJ. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genetic Epidemiology. 2008;32(4):381–5.

Discovery

Cases

- Papworth: 925
- Imperial: 229
- Sheffield: 59
- San Diego: 160

Controls

- UK Blood Service: 2997
- 1958 UK birth cohort: 2987

1183 | 197 | 1533 | 4451

Genotyping (Illumina) | Genotyping (Affymetrix)

Total: 1373 cases, 5984 controls

Sample Quality control		
Exclusions	**Case**	**Control**
Missingness	25	3
PCA	131	394
Rel/dup	65	78
Heterozygosity	6	11
Total	227	486

Total: 1146 cases, 5498 controls

Replication

Cases

- Belgium: 187
- Austria: 234
- Netherlands: 50
- Germany: 561
- Spain: 62

Controls

- UK: 5551
- Italy: 387
- Poland: 167
- Spain: 122
- Ireland: 15
- Germany: 437
- Sweden: 364
- Czechia: 149
- France: 71

747 | 347 | 7263

Genotyping (Illumina) | Genotyping (Affymetrix) | Genotyping (Replication)

Total: 1094 cases, 7263 controls

Sample Quality control		
Exclusions	**Case**	**Control**
Missingness	123	29
EGPA case	0	546
PCA	203	1696
Rel/dup	7	127
Heterozygosity	0	0
Total	333	2398

Total: 761 cases, 4865 controls

SNP quality control: MAF, HWE, missingness, differential missingness

Phasing and imputation

SNP quality control 2: INFO score, between-batch differences, between-control differences.

Case-control analysis: Affymetrix
171 cases, 4006 controls

Case-control analysis: Illumina
969 cases, 1492 controls

Case-control analysis Replication
761 cases, 4865 controls

Meta-analysis: discovery

Correction for inflation

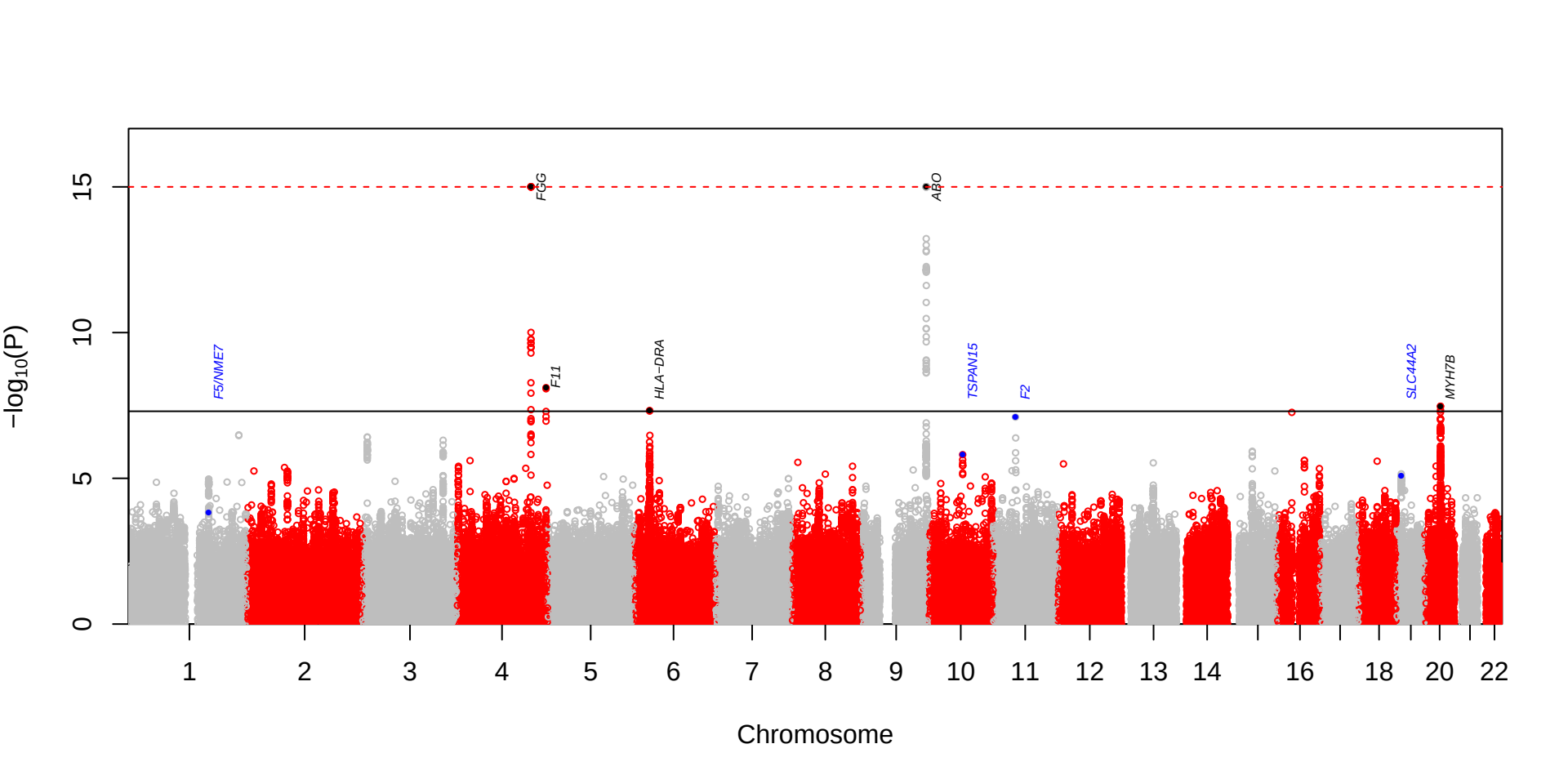
Correction for inflation

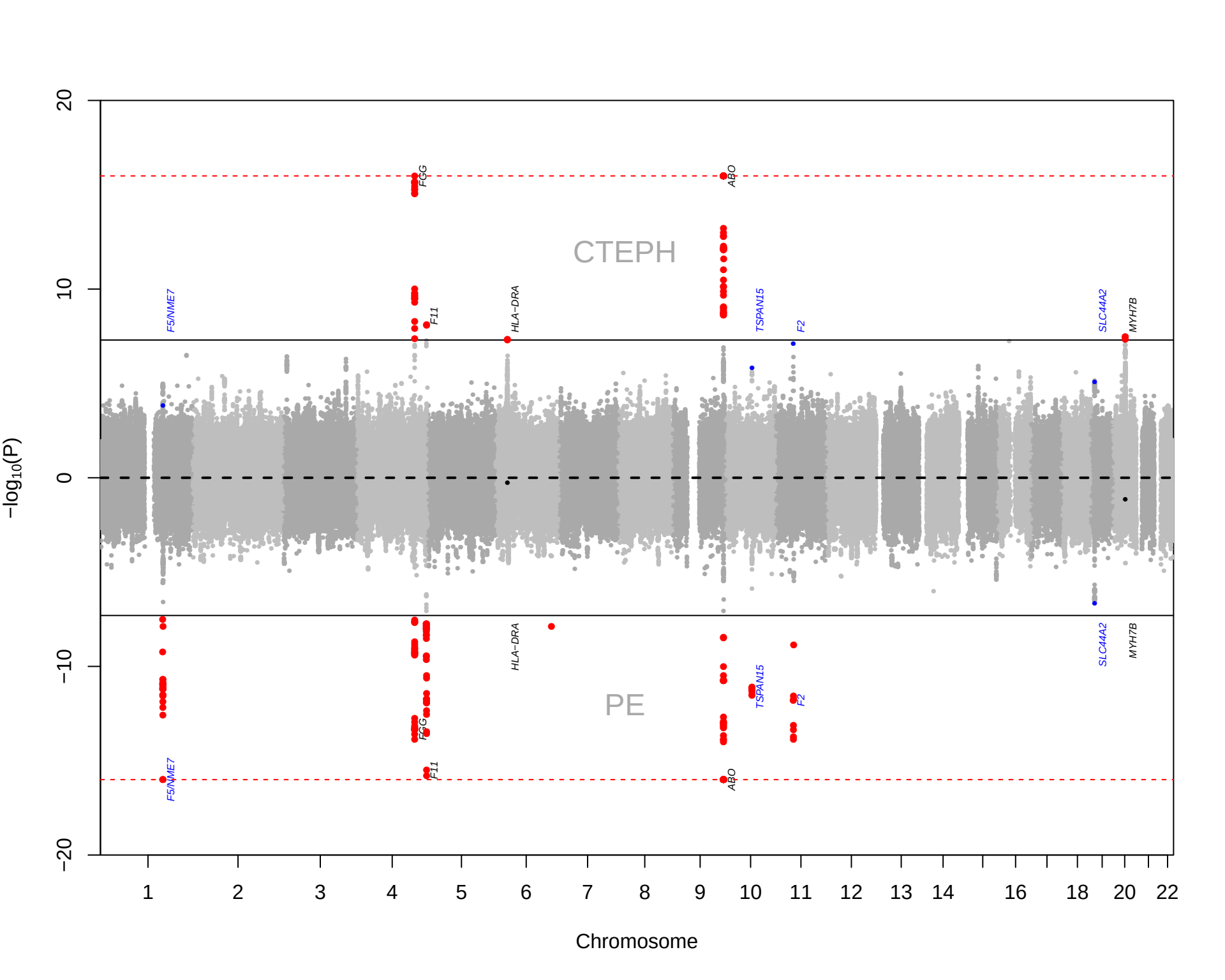
Determination of case-control findings and meta-analysis

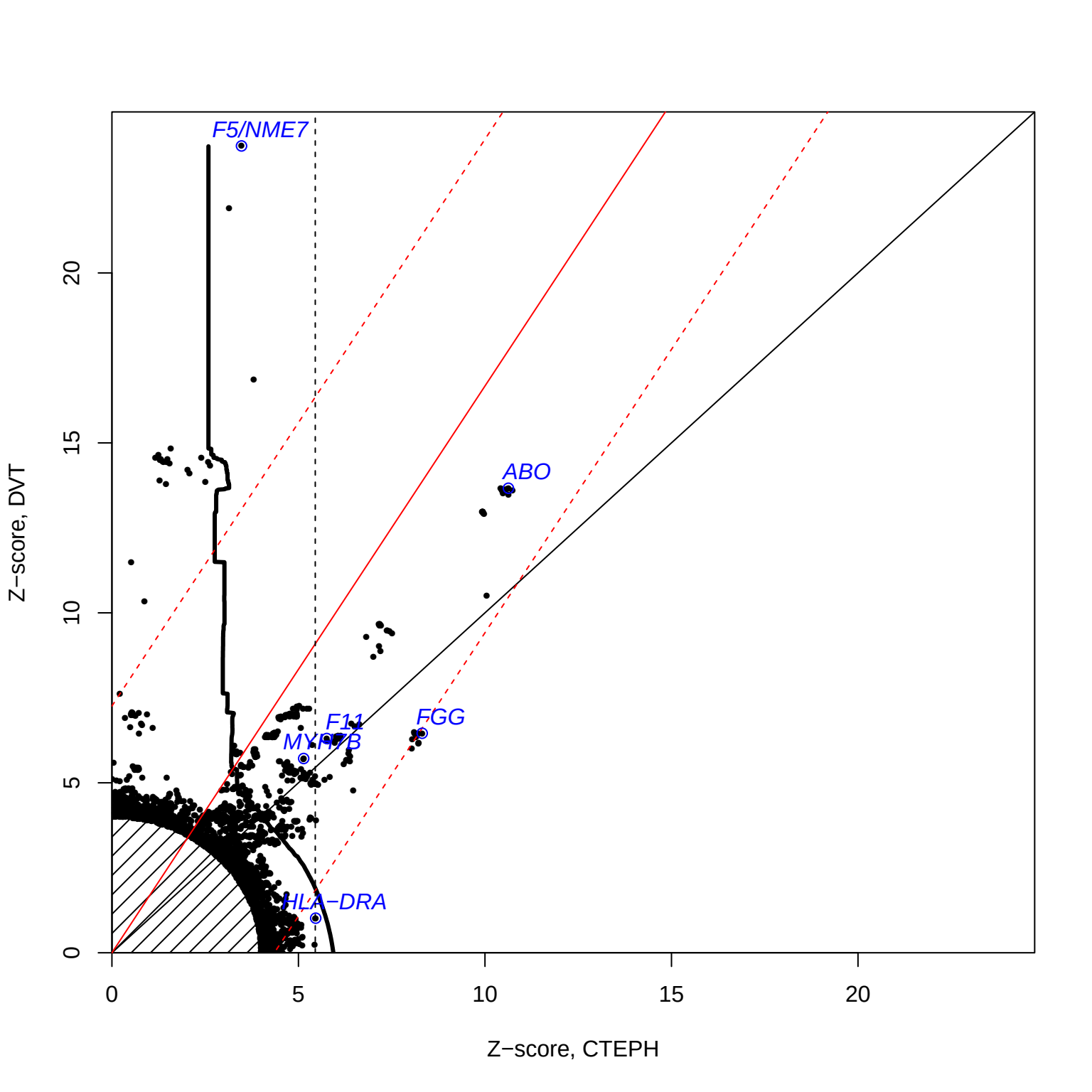
GWAS summary statistics for IPAH and VTE

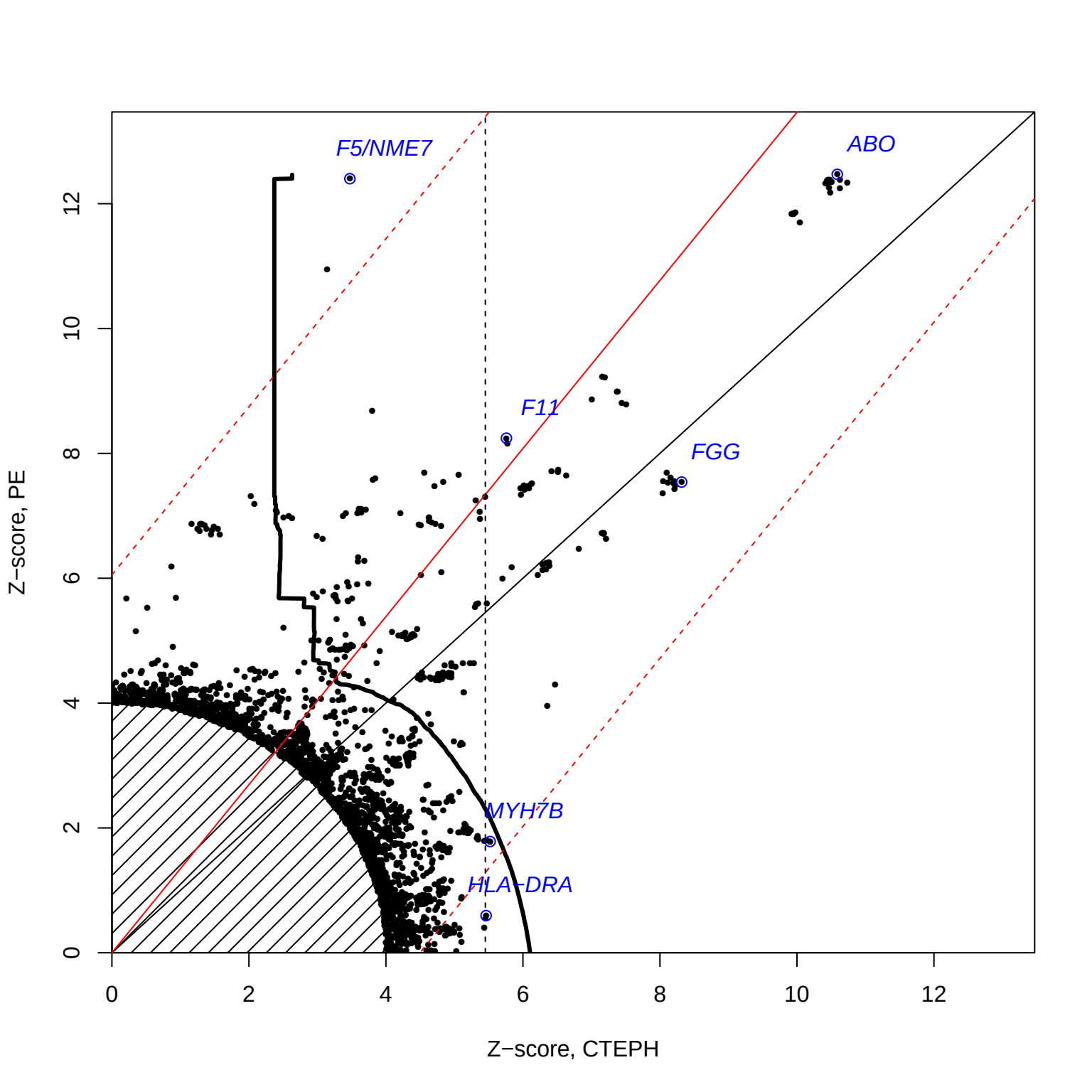
Co-analysis with related traits

Final results


−log10(P)
0
5
10
15
FGG
ABO
F5/NME7
F11
HLA−DRA
TSPAN15
F2
SLC44A2
MYH7B
1
2
3
4
5
6
7
8
9
10
11
12
13
14
16
18
20
22
Chromosome


CTEPH
PE
FGG
ABO
F5/NME7
F11
HLA−DRA
TSPAN15
F2
SLC44A2
MYH7B
20
10
0
−10
−20
1
2
3
4
5
6
7
8
9
10
11
12
13
14
16
18
20
22
Chromosome
−log10(P)


F5/NME7
ABO
F11
MYH7B
FGG
HLA−DRA
Z−score, DVT
Z−score, CTEPH
0
5
10
15
20


F5/NME7
ABO
F11
FGG
MYH7B
HLA−DRA
Z−score, PE
Z−score, CTEPH
0
2
4
6
8
10
12