Genetic susceptibility to pneumonia: A GWAS meta-analysis between UK Biobank and FinnGen

Adrian I. Campos[1,2,*], Pik Fang Kho[1], Karla X. Vazquez-Prada[3,4], Luis M. García-Marín[1], Nicholas G. Martin[1], Gabriel Cuéllar-Partida[1,‡], Miguel E. Rentería[1,2]

1. Department of Genetics & Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane QLD Australia
2. Faculty of Medicine, The University of Queensland, Brisbane QLD Australia
3. Australian Institute for Bioengineering and Nanotechnology, the University of Queensland, Brisbane, Queensland, Australia
4. School of Pharmacy, Pharmacy Australia Centre of Excellence, the University of Queensland, Woolloongabba, Queensland, Australia.
‡Current address: 23andMe Inc, Sunnyvale, CA, USA.
*Correspondence: Adrian I. Campos (adrian.campos@qimrberghofer.edu.au)

ABSTRACT

Rationale Pneumonia is a respiratory condition with complex aetiology. Host genetic variation is thought to contribute to individual differences in susceptibility and symptom manifestation.

Methods We analysed pneumonia data from the UK Biobank (14,780 cases and 439,096 controls) and FinnGen (9,980 cases and 86,519 controls). We perform genome-wide association study (GWAS) meta-analysis, gene-based test, colocalisation, genetic correlation, latent causal variable and polygenic prediction in an independent Australian sample (N=5,595) to draw insights into the genetic aetiology of pneumonia risk.

Results We identify two independent loci on chromosome 15 (lead SNPs rs2009746 and rs76474922) to be associated with pneumonia($p<5e-8$). Gene-based tests revealed eighteen genes in chromosomes 15,16 and 9, including *IL127, PBX3, APOBR* and smoking related genes *CHRNA3/5*, associated with pneumonia. Evidence of *HYKK* and *PBX3* involvement in pneumonia risk was supported by eQTL colocalisation analysis. We observed genetic correlations between pneumonia and cardiorespiratory, psychiatric and inflammatory related traits. Latent causal variable analysis suggests a strong genetic causal relationship cardiovascular health phenotypes and pneumonia risk. Polygenic risk scores (PRS) for pneumonia significantly predicted self-reported pneumonia history in an independent Australian sample, albeit with a small effect size (OR=1.11 95%CI=[1.04-1.19], $p<0.05$). Sensitivity analyses suggested the associations in chromosome 15 are mediated by smoking history, but the association of genes in chromosome 16 and 9, and polygenic prediction were robust to adjustment for smoking.

Conclusions Altogether, our results highlight common genetic variants, genes and potential pathways that contribute to individual differences in susceptibility to pneumonia, and advance our understanding of the genetic factors underlying heterogeneity in respiratory medical outcomes.

Keywords: Pneumonia, genome-wide association study (GWAS), respiratory infection, host response genetics, polygenic risk scores (PRS), UK Biobank, FinnGen.

INTRODUCTION

Pneumonia is an inflammatory condition of the lungs that usually stems from an infection. It is characterised by alveolar filling with fluid, microorganisms and immune response cells, preventing the lungs from working properly.[1] Diagnosis is confirmed with chest radiography showing abnormalities, and other pieces of evidence such as laboratory tests identifying the causal pathogen and increases in antibody count[2]. Pneumonia is associated with increased morbidity and mortality;[3] in fact, mortality estimates range between five and 14% for hospitalised patients. Risk factors for pneumonia include smoking,[4] alcoholism,[5] heart disease, and advanced age[6]. Furthermore, mortality amongst pneumonia cases is associated with factors such as hypertension and smoking.[7] Nonetheless, individuals considered '*at low risk*' of pneumonia can still develop the condition, which highlights its complexity and clinical heterogeneity.

Since the emergence of the 2020 COVID-19 pandemic, there has been an increase in pneumonia incidence and mortality.[8] Its relatively high infectivity and mortality even among *low-risk* groups calls for the investigation of genetic mechanisms underlying pathogenesis and prognosis. A recent study on 2633 British twins (728 complete pairs, 537 monozygotic and 191 dizygotic, 86.9% female) investigated the susceptibility to infection by SARS-CoV-2.[9] The researchers used a symptom-based algorithm to predict true infection in participants tested for SARS-CoV-2 and estimated heritability for symptoms including fever = 0.41 (95% CI 0.12-0.70); anosmia 0.47 (0.27-0.67) and delirium 0.49 (0.24-0.75). Overall predicted heritability of COVID-19 status was 0.50 (0.29-0.70), suggesting that symptomatic infection with SARS-CoV-2 is under host genetic influence to some extent, and reflecting inter-individual variation in the host immune response. Thus, host-specific genetic susceptibility is an emerging area of research interest[10] as it could facilitate the systematic stratification of patients by genetic risk and aid in the design of more efficient treatments.[11]

In fact, evidence from other infectious diseases points to an important role for host genetics in influencing the development of symptomatic infection.[12] Twin studies have shown higher concordance rates of tuberculosis, leprosy, poliomyelitis and hepatitis B in identical versus non-identical twins, suggesting a genetic component in susceptibility to these infectious diseases.[12] Moreover, clinical trials for drugs targeting genes with evidence of disease association are more likely to lead to useful therapies.[13,14] Thus, identification of genes and pathways that confer increased susceptibility to pneumonia could reveal new therapeutic targets and inform the design of prevention and treatment strategies.

Here, we report a GWAS meta-analysis of pneumonia history in adults using data from two large datasets, the UK Biobank and FinnGen. We identify genetic variants and genes associated with

pneumonia risk, an essential step for understanding inter-individual differences in susceptibility. We characterise the genetic aetiology of pneumonia by assessing its genetic correlations and genetic evidence for causality against ~1,500 traits with publicly available GWAS data. Finally, we demonstrate the external validity of our findings by performing polygenic prediction of self-reported pneumonia in an independent Australian sample.

METHODS

Samples and phenotypic information

For this study, we meta-analysed GWAS for pneumonia in two independent samples: the UK Biobank and FinnGen. For UK Biobank, we conducted a GWAS of pneumonia using individual-level genetic and phenotypic data from the UK Biobank. International Classification of Diseases (ICD10) codes are used to store information on participants' health conditions. Raw ICD10 data were extracted from the UK Biobank under Application Number 25331. In this study, we excluded participants of non-European ancestry to avoid potential genetic associations emerging from population stratification. Participants with a history of pneumonia were defined as those presenting any ICD10 code related to infectious pneumonia (N=14,780) (see **Supplementary Table 1**). For FinnGen, we leveraged publicly available summary statistics on the phenotype *ICD10-J10 pneumonia* which comprised 9,980 cases and 86,519 controls. Information on sample phenotyping, genotyping and GWAS in the FinnGen sample is available elsewhere.[15]

Pneumonia GWAS in the UK Biobank

GWAS was performed using BOLT-LMM, which implements a linear mixed model association analysis and fits a genetic relationship matrix as a random effect to account for cryptic relatedness and population stratification. Age, sex, genotyping array and the first 20 genetic principal components were adjusted for in the analysis. We used a stringent quality control procedure corresponding to minor allele frequency (MAF ≥ 0.01) and imputation quality (INFO≥0.60).

GWAS meta-analysis

A z-score meta analysis of pneumonia summary statistics was conducted between the UK Biobank and Finngen samples using METAL v(2011-03-25). The final meta-analysis comprised 24,760 cases and 525,615 controls. Only variants passing quality control in both cohorts were included in the meta-analysis. Furthermore, variants with inconsistent allele frequencies in both cohorts (difference >0.15) were removed. The final number of variants meta-analysed and included in this study was 7,831,927. Independent genetic signals were identified by clumping (r2<0.05, and 1Mb window) using CTG-VL (beta 0.1)[16]. A sensitivity analysis was performed by adjusting the GWAS results using multi-trait conditional and joint analysis (mtCOJO) to simultaneously adjust for two smoking phenotypes: smoking history and cigarettes per day.

Gene-based analysis

Gene-based analysis was conducted on both the main and smoking adjusted GWAS using the "set-based association analysis for human complex traits" fastBAT method[17] available on CTG-VL (https://genoma.io). fastBAT performs a set-based enrichment analysis based on the GWAS summary statistics while accounting for linkage disequilibrium (LD) between SNPs. We tested the association between 24,443 genes and pneumonia using this method. Statistical significance was defined using Benjamini-Hochberg False Discovery Rate (FDR) < 5% for multiple testing correction. Genes identified as statistically significant were further assessed for eQTL colocalisation with pneumonia.

Colocalisation and eQTL

To assess the co-occurrence of signals in GWAS data and cis- expression quantitative trait loci (eQTL) data, we performed a summary-based colocalisation analysis. We integrated our GWAS data and cis-eQTL data from lung tissue and whole blood in GTEx V7. We used GWAS and eQTL summary statistics of SNPs within 1Mb window around each fastBAT-identified gene to estimate the posterior probability that GWAS signals co-occur with eQTL signals while accounting for LD structure. This method estimates the posterior probabilities for five different scenarios: no association with either trait (PP0), association with the disease only (PP1), association with gene expression only (PP2), associations with both traits but distinct SNPs (PP3) and associations with both traits in same SNPs (PP4). A threshold of PP4/(PP3+PP4) > 0.8 was considered as evidence for co-occurrence of GWAS signals and eQTL signals at the region of interest. Colocalisation analysis was performed using the *COLOC* package in R.

Heritability and genetic correlations

We used LD-score regression (LDSC) to estimate the SNP-based heritability (h_{SNP}^2) for pneumonia on the liability scale assuming prevalence estimates of UK Biobank (3.3%) as both sample and population prevalence. Genetic correlations (rG) between pneumonia and 1,522 phenotypes were estimated using bivariate LDSC regression in CTG-VL based on a common set of HapMap3 variants. Benjamini-Hochberg FDR at 5% was used to assess statistical significance.

Genetic Causal Proportion

To assess whether significant genetic correlations observed could be explained by an underlying causal relationship between traits, we used the Latent Causal Variable (LCV) method[18] as implemented in CTG-VL. LCV uses GWAS summary statistics to estimate the genetic causal proportion (GCP) between two traits. The GCP's absolute value ranges from 0 (no genetic causality) to 1 (full genetic causality). In our study, a high GCP value (GCP > 0.60) indicates that pneumonia is likely to affect the trait of interest. In contrast, a robust negative value (GCP < -0.60) provides evidence that the trait of interest is likely to affect pneumonia. For traits of interest (deep vein

thrombosis, LDL and cholesterol) with significant evidence of a causal effect on pneumonia, generalised summary data-based Mendelian Randomisation (GSMR) was used as a secondary assessment of the existence of a causal relationship.

Target sample and polygenic risk scoring

To assess the external validity of the GWAS, we performed polygenic based prediction on an independent target sample of 5,595 unrelated Australian Adults from the Australian Genetics of Depression Study (AGDS) with complete data.[19] Pneumonia cases were identified through self-reported medical history in AGDS. PRS analysis was further adjusted for smoking by: i) additionally including smoking history as a covariate and ii) performing PRS calculation using the summary statistics adjusted for smoking history and cigarettes per day. Smoking history was assessed with the item: "*Have you smoked more than 100 cigarrettes in your lifetime*?". We employed a recently developed method, SBayesR, to obtain the conditional effects of the studied variants, thus avoiding inflation arising from using correlated SNPs due to LD. Pneumonia polygenic risk scores (PRS) were calculated using PLINK 1·9. in the AGDS sample. Briefly, a PRS is calculated by multiplying the effect size of a given risk allele (obtained from the discovery GWAS summary statistics) by the imputed number of risk alleles (using dosage probabilities) present in each individual. Then, obtaining a weighted average across all loci. To assess the association between pneumonia PRS and self-reported pneumonia history in AGDS, we used a logistic regression model (python *statsmodels*). Pneumonia PRS was the predictive variable of interest, with age, sex and the first 20 genetic ancestry principal components included as covariates.

RESULTS

Prevalence of pneumonia and sample demographics

The prevalence of lifetime pneumonia in the UK Biobank was 3.3%. Sex was associated with pneumonia, where females were less likely to have experienced the condition (Female OR = 0.713 95%C.I.=[0.69-0.737]). Furthermore, participants with a history of pneumonia were on average older than controls (OR = 1.06 95%C.I.= [1.06-1.07]). Smoking history was also associated with an increased pneumonia risk (OR=1.74 95%C.I.=[1.68-1.68] see **Table 1**).

Pneumonia GWAS

Our GWAS meta-analysis identified two independent genome-wide significant variants on 15q15.1 (index SNPs rs2009746 and rs76474922; p<5e-8; **Figure 1a**). The significant locus was located in a gene-rich region near *IREB2, CHRNA3/5* and *HYKK* (**Supplementary Figure 1**). In addition, eighteen independent loci showed suggestive association with pneumonia (**Table 2**). The amount of variance on the liability of pneumonia explained by this GWAS in the UK Biobank, also called the SNP heritability of the trait, the whole meta-analysis was estimated at 0.03 (s.e.=0.006) using LDSC regression. A sensitivity analysis using mtCOJO to adjust for smoking history and cigarettes per day revealed the hits on chromosome 15, but not other signals, to be mediated by smoking. A near

genome-wide signal in chromosome 3, near the gene *SUCNR1,* became significant after conditioning on smoking phenotypes (**Figure 1b**) Notably, the genetic correlation between the unconditional and smoking conditional GWAS was high (rg=0.9371, S.E.= 0.015).

Gene-based analysis and colocalisation

We performed gene-based association testing followed by colocalisation analysis to identify genes likely associated with pneumonia. fastBAT analysis revealed eighteen genes, in chromosomes 9,15 and 16, to be associated with pneumonia risk (**Supplementary Table 2**). Sensitivity gene-based tests suggested the association of genes in chromosome 15, but not those in chromosomes 9 and 16, to be mediated by smoking (**Figure 2a**). Two genes, *HYKK* and *PBX3*, showed evidence of colocalisation in lung tissue (**Table 3**), but not whole blood. *EIF3C* showed suggestive evidence of colocalisation in the lung, and strong evidence of colocalisation in whole blood (**Supplementary Table 3**). While *IL27*, *CHRNA3* and *CHRNA5* have eQTL signals in the vicinity of pneumonia hits, our analysis suggests that the relationship between their expression and pneumonia is better explained by two neighboring independent causal variants (**Figure 2b**).

LD-score genetic correlations

Across 1,522 traits studied, 552 traits displayed a genetic overlap with pneumonia at FDR < 5%. Traits with the strongest evidence of a genetic correlation with pneumonia included chronic obstructive pulmonary disease (COPD), “Wheeze or whistling in the chest in last year”, blood clot in the leg and myocardial infarction (**Figure 3**). Lifestyle factors such as current smoking showed a positive genetic correlation with pneumonia, indicating that variants that increase smoking behaviour also increase pneumonia risk. Genetic correlation between alcohol intake and pneumonia was conflicting, as the variable “*Alcohol usually taken with meals*” and “*Alcohol drinker status: Current*” had a negative genetic correlation with pneumonia. In contrast, the variable “Alcohol drinker status: Previous” displayed a positive genetic correlation with pneumonia. Traits related to mood or psychiatric disorders (such as depression and irritability), lifestyle variables (such as cycling to work and educational attainment), and biomarkers (such as immune cell count and C Reactive Protein [CRP]), among others, also showed significant genetic correlations with pneumonia (**Figure 3**).

Genetic Causal Proportions

To assess whether the genetic correlations observed could be explained by a causal relationship, we performed a latent causal variable analysis. Forty four of the 552 traits with a significant (FDR < 5%) genetic overlap with pneumonia showed evidence of a causal association (**see methods**). LCV provided genetic evidence on several traits causally associated with pneumonia, including deep vein thrombosis (DVT), LDL (decreased), cholesterol (decreased) among other traits closely related to cardiovascular health, such as heart failure, arrhythmias and fibrillation. Evidence for DVT, hypertension, LDL and the cholesterol causal associations were further assessed using GSMR. This

analysis showed a consistent result for DVT and hypertension, but no evidence of causality for LDL or cholesterol (**Supplementary Figure 2**). Traits highlighted as potential consequences of pneumonia included long-standing illness, lower forced vital capacity, anhedonia, pain, and taking omeprazole and co-codamol (**Figure 4** and **Supplementary Data 1**).

Polygenic prediction of pneumonia

We performed polygenic prediction of pneumonia on the AGDS sample to assess the validity of our pneumonia GWAS. The prevalence of self-reported pneumonia history (~2000 cases, ~20%) in the AGDS sample was higher than pneumonia diagnosis in the UK Biobank (~15k cases ~3%) and FinnGen (~10k cases ~10%)**.** Furthermore, the AGDS sample had a different age and sex composition from the UK Biobank (**Table 4).** We assessed whether PRS derived from the pneumonia GWAS were associated with pneumonia in the AGDS cohort using a multivariate logistic regression (**see Methods**) and identified a statistically significant, but small in effect, association between pneumonia PRS and self-reported pneumonia OR=1.06 (95%CI=[1.01,1.12]; p=0.02) per standard deviation increase of pneumonia PRS.

Sensitivity analyses

The genome-wide significant locus overlaps, and is in LD, with a set of well established smoking-associated variants including rs16969968.[20] To assess whether the genetic associations for pneumonia are mediated by smoking, we performed several sensitivity analyses. A conditional association test showed that our top hit (rs2009746) evidence of association was reduced after adjusting for three independent smoking associated variants ($p_{rs2009746}$=0.002; **Supplementary Table 4**). Nonetheless, an mtCOJO analysis suggested the associations between pneumonia and genes in chromosomes 16 and 9 to be independent from smoking (**Figure1** and **Figure2**). Finally, the association between $pneumonia_{PRS}$ and self-reported pneumonia remained statistically significant after adjustment for smoking history both on the genetic and phenotypic level (**Supplementary Table 5**).

DISCUSSION

Our findings highlighted eighteen genes, across chromosomes 6, 15 and 16 to be associated with pneumonia risk. We identified genes involved in general gene regulation (*PBX3*, *EIF3C*), iron regulation (*IREB2*), nicotine signaling (*CHRNA3/5*) and inflammatory processes (*IL27, APOBR*). Here, we integrated eQTL data with our GWAS results and performed colocalisation analysis to identify which genes have more robust evidence of association with pneumonia. Our analyses suggested *HYKK* and *PBX3* gene expression to colocalise with pneumonia. Notably, *PBX3* encodes a transcription factor whose deficiency has been linked to respiratory failure in mice.[22] *HYKK* is an enzyme involved in lysine catabolism and was recently linked to nicotine metabolism.[23] While our colocalisation analysis would suggest *HYKK*, *PBX3* and potentially *EIF3C* are associated with

pneumonia through differential gene expression, other genes identified could be associated through mechanisms such as impairment or gain of function.

Genetic variants in 15q25.1 have been extensively linked with smoking.[24] This complex region has also been previously associated with COPD[25] and lung cancer[26], and contains several compelling genes associated with nicotine addiction (*CHRNA3, CHRNB4, CHRNA5*, *HYKK*) and iron regulation (*IREB2*). We performed a sensitivity analysis and showed that 15q25.1 was not associated with pneumonia after adjusting for smoking history and cigarettes per day. Nonetheless, genes in other regions remained associated with pneumonia after adjusting for smoking. This is consistent with the observed high genetic correlation between the smoking-adjusted and unadjusted summary statistics. Moreover, polygenic prediction was also robust to adjustment for smoking history. Future efforts could leverage analyses such as pairwise GWAS or genomicSEM to further deconvolute the effects of smoking and respiratory disease. We consider this beyond the scope of the present study.

We discovered genetic correlations between pneumonia and biomarkers such as immune cell counts, cystatin C and sodium in urine. Consistently, Cystatin C and CRP levels have been linked to community-acquired pneumonia (CAP).[27,28] Furthermore, lifestyle factors such as smoking, and lower socioeconomic status (as measured by the Townsend deprivation index) were genetically correlated with pneumonia. Finally, traits requiring healthy respiratory function such as *cycling to work* and *maximum workload during a fitness test* displayed a negative genetic correlation with pneumonia.

A genetic correlation between two traits could reflect causality between traits, or horizontal pleiotropy (genes acting on both traits independently of each other). Here, we performed LCV analyses to identify traits causally associated with pneumonia. Our results suggest that deep vein thrombosis (DVT) may causally increase risk of pneumonia. This result was further confirmed using GSMR. Previous studies have noted an association between these two diseases.[29] Most studies suggest or assume that pneumonia causes DVT due to immobilization, hypoxia and inflammation. Hypoxia is one of the strongest predictors of pneumonia[29] and has been shown to increase the incidence of thrombosis through the downregulation of protein S, a natural anticoagulant.[30] Furthermore, tissue factor, along with coagulation related pathways, are known to be upregulated upon inflammation.[31] Future studies should focus on further understanding of the intricate relationship between cardiovascular and respiratory diseases.

LCV also highlighted the involvement of cholesterol levels and LDL in decreasing the risk for pneumonia. Nonetheless, these results did not replicate in our GSMR analyses. Cholesterol is essential for cellular integrity and metabolism, and its dysregulation has been linked to a variety of diseases, including cardiovascular and pulmonary disease.[32] Previous studies show that LDL and

HDL trafficking influences multiple cell types in the lung.[33] Class A scavenger receptors on alveolar macrophages uptake HDL as a source of vitamin E,[34] which is an antioxidant that plays an essential role in the clearance of oxidized lipids that would otherwise result in cytotoxic and pro-inflammatory responses.[35] Furthermore, cholesterol plays an essential role in protecting and covering the alveoli which prevents several pathological conditions.[36] Thus, total cholesterol might protect from developing pneumonia through the relationship between cholesterol and immune homeostasis in the lung. Nevertheless, low levels of LDL have been associated with better lung function,[37] and low HDL levels have been proposed as a poor prognosis marker for CAP.[38] Moreover, a recent proteomic study in patients with sepsis secondary to pneumonia were found to have an impairment in lipid metabolism (lower total cholesterol, LDL cholesterol, as well as major apolipoprotein of LDL, ApoB).[39] This is consistent with our gene based tests identifying the ApoB receptor (*APOBR*) as a potential pneumonia risk mediating gene. Overall our findings and the literature suggest that a dyslipidemic state, rather than specific levels of LDL influence pneumonia risk.

Some limitations of the present study must be acknowledged. We excluded participants of non-European ancestry to avoid biases due to population stratification. This limits the generalisability of our findings to populations of non-European ancestry. Furthermore, our results suggest that the genetic risk for pneumonia is highly complex, and several variants remain to be identified by more powered studies. Further evidence of this is the low polygenic prediction in an independent sample, which is still far from other traits where clinical relevance is starting to be considered. We replicated LCV findings using GSMR. Nonetheless, we could not attempt to replicate any of the causal associations where pneumonia was the exposure because our pneumonia GWAS was underpowered to be accurately used as an exposure. Finally, experimental approaches along with powered analyses considering not only smoking history but also smoking exposure and quantitative smoking measures are needed to claim, beyond any doubt, 15q25.1 to be associated with pneumonia over and above smoking.

In summary, pneumonia GWAS meta-analysis identified a region in 15q25.1 which has been previously linked to smoking, lung cancer and COPD. Gene-based tests association identified eighteen genes implicated in pneumonia risk in chromosomes 9, 15 and 16. Sensitivity analyses suggested the locus in chromosome 15 to be driven by smoking, but other associations were robust to adjustment for smoking related traits. eQTL colocalisation analysis in lung tissue suggested *HYKK, PBX3* and potentially *EIF3C* expression to colocalise with pneumonia. We identified traits with a significant genetic correlation and highlighted potential causally associated traits, including DVT and lipid homeostasis. Finally, validation of our GWAS was obtained by polygenic prediction of self-reported history of pneumonia in an independent sample. Polygenic prediction was robust to adjustment for smoking history, suggesting some independence of our GWAS signals from smoking

history. Increasing statistical power could help identify additional genetic targets which will, in turn, enable the development of new therapeutics and patient risk stratification based on genetic risk.

ACKNOWLEDGMENTS

This research was conducted using data from the UK Biobank resource under application number 25331. We want to acknowledge the participants and investigators of the FinnGen study. Data collection for the Australian Genetics of Depression Study was possible, thanks to funding from the Australian National Health & Medical Research Council (NHMRC) to N.G.M. (GNT1086683). A.I.C. and K.X.V.P. are both supported by UQ Research Training Scholarships from The University of Queensland (UQ). P.F.K. is supported by an Australian Government Research Training Program Scholarship from Queensland University of Technology (QUT). M.E.R. thanks support of the NHMRC and Australian Research Council (GNT1102821).

AUTHOR CONTRIBUTIONS

AIC conceived the study. AIC and PFK performed the analyses with aid and input from NGM, GCP and MER. KXVP and LGM helped interpreting the results. NGM designed and directed the AGDS study. All authors collaboratively wrote the manuscript.

DATA AVAILABILITY

The full GWAS summary statistics for this study will be made available through the NHGRI-EBI GWAS Catalogue (https://www.ebi.ac.uk/gwas/downloads/summary-statistics) upon publication. Individual level data for UK Biobank participants are available to eligible researchers through the UK Biobank (www.biobank.ac.uk). Results for the GWAS downstream analyses have been made available in CTG-VIEW (https://view.genoma.io). Code used for this study is available upon request.

REFERENCES

1. NICE. *Pneumonia in adults*. (January 09, 2016).
2. Szalados, J. E. Pneumonia in Adults. *Critical Care* 157–175 (2005) doi:10.1016/b978-0-323-02262-0.50019-1.
3. Le Jeune I and Macfarlane JT and Read RC and Roberts HJ and Levy ML et al, L. W. S. A. B. S. V. A. G. R. C. A. H. A. T. A. J. C. A. BTS Guidelines for the Management of Community Acquired Pneumonia in Adults: Update 2009. *Thorax* **64**, (2009).
4. Farr, B. M., Bartlett, C. L., Wadsworth, J. & Miller, D. L. Risk factors for community-acquired

pneumonia diagnosed upon hospital admission. British Thoracic Society Pneumonia Study Group. *Respir. Med.* **94**, 954–963 (2000).

5. Ruiz, M. *et al.* Severe Community-acquired Pneumonia. *American Journal of Respiratory and Critical Care Medicine* vol. 160 923–929 (1999).
6. Koivula, I., Sten, M. & Makela, P. H. Risk factors for pneumonia in the elderly. *Am. J. Med.* **96**, 313–320 (1994).
7. Guo, L. *et al.* Clinical Features Predicting Mortality Risk in Patients With Viral Pneumonia: The MuLBSTA Score. *Front. Microbiol.* **10**, 2752 (2019).
8. Shi, H. *et al.* Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. *The Lancet Infectious Diseases* vol. 20 425–434 (2020).
9. Williams, F. M. K. *et al.* Self-reported symptoms of covid-19 including symptoms most predictive of SARS-CoV-2 infection, are heritable. *Genetic and Genomic Medicine* (2020).
10. Tanigawa, Y. & Rivas, M. Initial Review and Analysis of COVID-19 Host Genetics and Associated Phenotypes. *LIFE SCIENCES* (2020).
11. Salnikova, L. E., Smelaya, T. V., Vesnina, I. N., Golubev, A. M. & Moroz, V. V. Genetic susceptibility to nosocomial pneumonia, acute respiratory distress syndrome and poor outcome in patients at risk of critical illness. *Inflammation* **37**, 295–305 (2014).
12. Cooke, G. S. & Hill, A. V. Genetics of susceptibility to human infectious disease. *Nat. Rev. Genet.* **2**, 967–977 (2001).
13. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).
14. King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet.* **15**, e1008489 (2019).
15. FinnGen. FinnGen Documentation of R2 release. https://finngen.gitbook.io/finngen-documentation/-LvQ4yR2YFUM5eFTjieO/ (2020).
16. Cuellar-Partida, G. *et al.* Complex-Traits Genetics Virtual Lab: A community-driven web platform for post-GWAS analyses. *Genetics* 194 (2019).
17. Bakshi, A. *et al.* Fast set-based association analysis using summary data from GWAS

identifies novel gene loci for human complex traits. *Sci. Rep.* **6**, 32894 (2016).

18. O'Connor, L. J. & Price, A. L. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nat. Genet.* **50**, 1728–1734 (2018).
19. Byrne, E. M. *et al.* The Australian Genetics of Depression Study: Study Description and Sample Characteristics. *Genetics* e937 (2019).
20. Saccone, N. L. *et al.* Multiple independent loci at chromosome 15q25.1 affect smoking quantity: a meta-analysis and comparison with lung cancer and COPD. *PLoS Genet.* **6**, (2010).
21. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237–244 (2019).
22. Rhee, J. W. *et al.* Pbx3 Deficiency Results in Central Hypoventilation. *Am. J. Pathol.* **165**, 1343–1350 (2004).
23. Buchwald, J. *et al.* Genome-wide association meta-analysis of nicotine metabolism and cigarette consumption measures in smokers of European descent. *Mol. Psychiatry* (2020) doi:10.1038/s41380-020-0702-z.
24. Bierut, L. & Cesarini, D. How Genetic and Other Biological Factors Interact with Smoking Decisions. *Big Data* **3**, 198–202 (2015).
25. Hardin, M. *et al.* CHRNA3/5, IREB2, and ADCY2 are associated with severe chronic obstructive pulmonary disease in Poland. *Am. J. Respir. Cell Mol. Biol.* **47**, 203–208 (2012).
26. McKay, J. D. *et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat. Genet.* **49**, 1126–1132 (2017).
27. Holloway, A. J., Yu, J., Arulanandam, B. P., Hoskinson, S. M. & Eaves-Pyles, T. Cystatins 9 and C as a Novel Immunotherapy Treatment That Protects against Multidrug-Resistant New Delhi Metallo-Beta-Lactamase-1-Producing Klebsiella pneumoniae. *Antimicrob. Agents Chemother.* **62**, (2018).
28. García Vázquez, E. *et al.* C-reactive protein levels in community-acquired pneumonia. *Eur. Respir. J.* **21**, 702–705 (2003).
29. Rae, N., Finch, S. & Chalmers, J. D. Cardiovascular disease as a complication of community-

acquired pneumonia. *Curr. Opin. Pulm. Med.* **22**, 212–218 (2016).

30. Pilli, V. S. *et al.* Hypoxia downregulates protein S expression. *Blood* **132**, 452–455 (2018).
31. Esmon, C. T. Inflammation and thrombosis. *J. Thromb. Haemost.* **1**, 1343–1348 (2003).
32. Ravnskov, U. High cholesterol may protect against infections and atherosclerosis. *QJM* **96**, 927–934 (2003).
33. Gowdy, K. M. & Fessler, M. B. Emerging roles for cholesterol and lipoproteins in lung disease. *Pulm. Pharmacol. Ther.* **26**, 430–437 (2013).
34. Kolleck, I. *et al.* HDL is the major source of vitamin E for type II pneumocytes. *Free Radical Biology and Medicine* **27**, 882–890 (1999).
35. Fessler, M. B. A New Frontier in Immunometabolism. Cholesterol in Lung Health and Disease. *Ann. Am. Thorac. Soc.* **14**, S399–S405 (2017).
36. Andersson, J. M., Grey, C., Larsson, M., Ferreira, T. M. & Sparr, E. Effect of cholesterol on the molecular structure and transitions in a clinical-grade lung surfactant extract. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E3592–E3601 (2017).
37. Barochia, A. V. *et al.* Serum apolipoprotein A-I and large high-density lipoprotein particles are positively correlated with FEV1 in atopic asthma. *Am. J. Respir. Crit. Care Med.* **191**, 990–1000 (2015).
38. Chien, Y.-F., Chen, C.-Y., Hsu, C.-L., Chen, K.-Y. & Yu, C.-J. Decreased serum level of lipoprotein cholesterol is a poor prognostic factor for patients with severe community-acquired pneumonia that required intensive care unit admission. *J. Crit. Care* **30**, 506–510 (2015).
39. Sharma, N. K. *et al.* Lipid metabolism impairment in patients with sepsis secondary to hospital acquired pneumonia, a proteomic analysis. *Clin. Proteomics* **16**, 29 (2019).

FIGURES

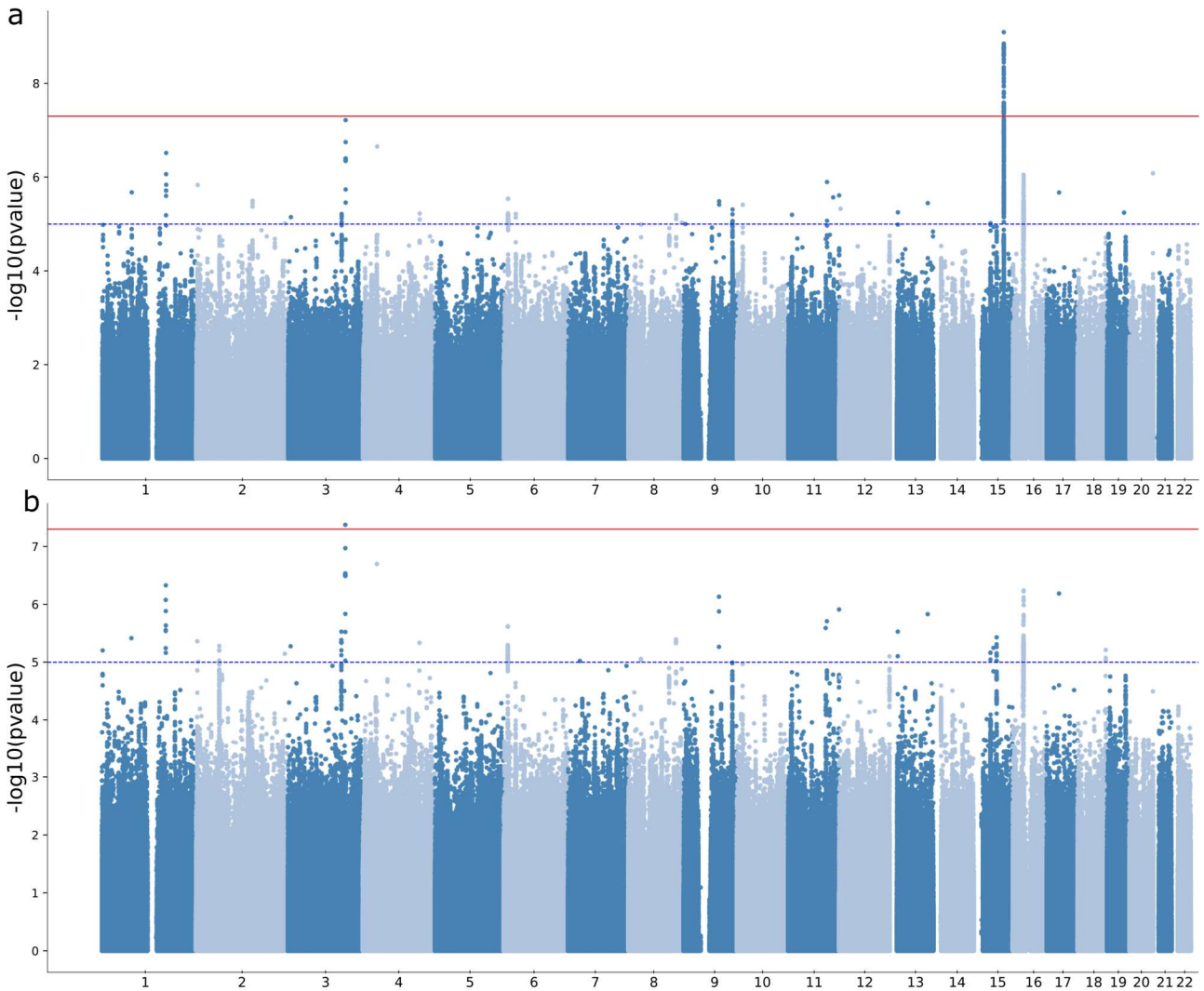

Figure 1. Pneumonia GWAS meta-analysis and gene based association tests

a) Manhattan plot shows the results of the genome-wide association study meta-analysis. Each dot represents a genetic variant. The x-axis is the genomic location ordered by chromosome. The y-axis represents the statistical evidence of the association (-log10 transformed p-value). The solid red and dashed blue lines represent the genome-wide and suggestive association significance thresholds. **b)** Manhattan plot shows the results of a sensitivity analysis using mtCOJO to condition on smoking history and cigarettes per day. Note the hit on chromosome 15 is no longer significant after this adjustment, while other signals remain largely unchanged.

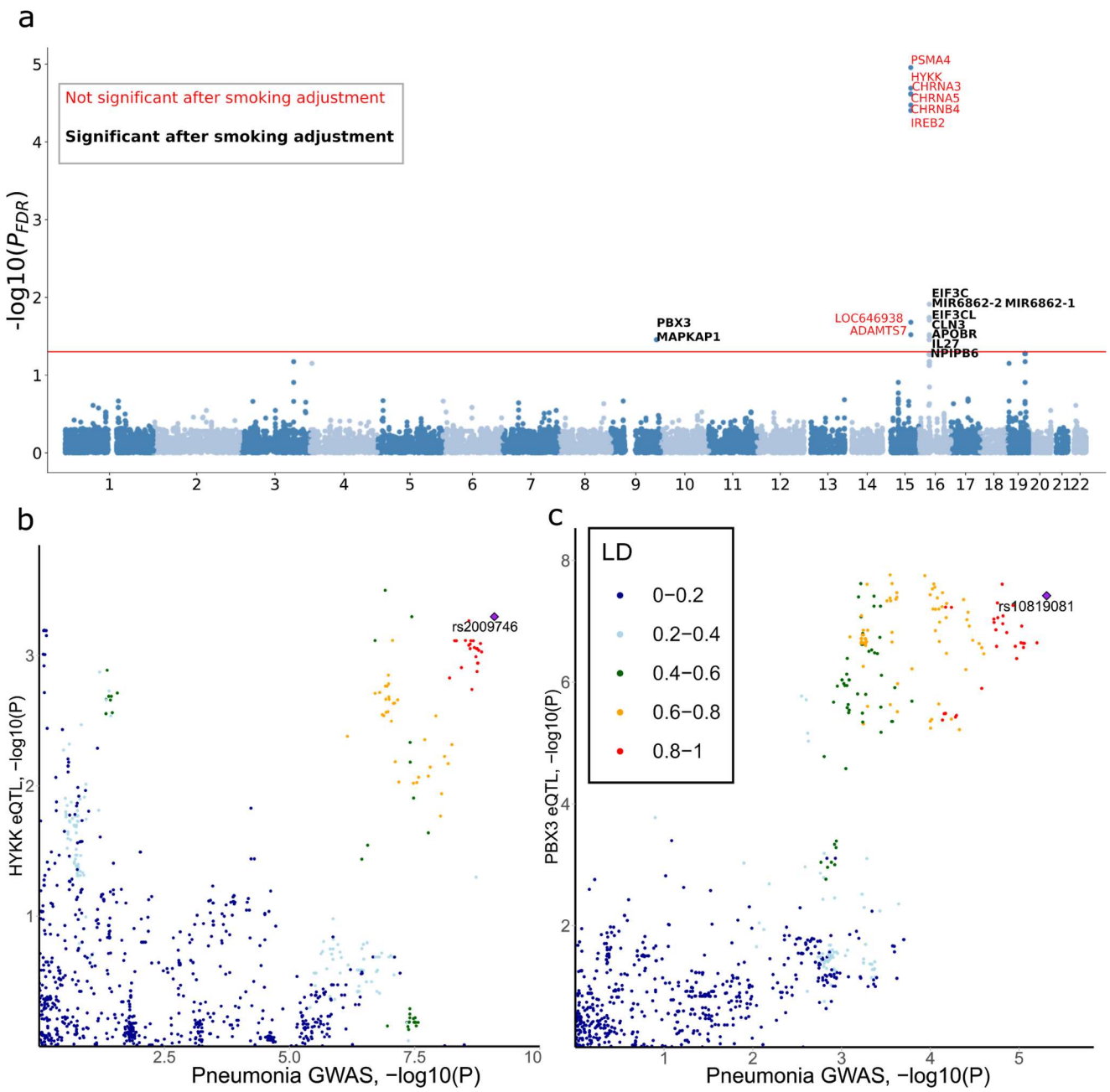

Figure 2. HYKK (CHR 15) and PBX3 (CHR9) eQTLs colocalise with pneumonia

a) Gene based test association results. Each dot represents a gene and its position on the y-axis corresponds to the p-value for association with pneumonia adjusted for multiple testing. Genes in bold (black) were robust to adjustment for smoking phenotypes, whereas genes in non-bold (red) font were not. Genes above the red line are significantly associated with pneumonia, and were assessed for eQTL colocalisation. b and c show colocalisation plots assessing shared signals between lung eQTLs and the pneumonia meta-analysis. Each dot represents a genetic variant. The x axis represents the evidence for association (-log10pvalue) between that variant and pneumonia. The y-axis represents the evidence for association between that variant and expression of the gene of interest. Colocalisation happens when there is a high level of co-occurrence between GWAS signals and eQTL signals. Two independent signals driving each trait would show two signals along the x and y axis respectively. Results shown only for HYKK (a) and PBX3 (b) as these genes showed evidence of colocalisation.

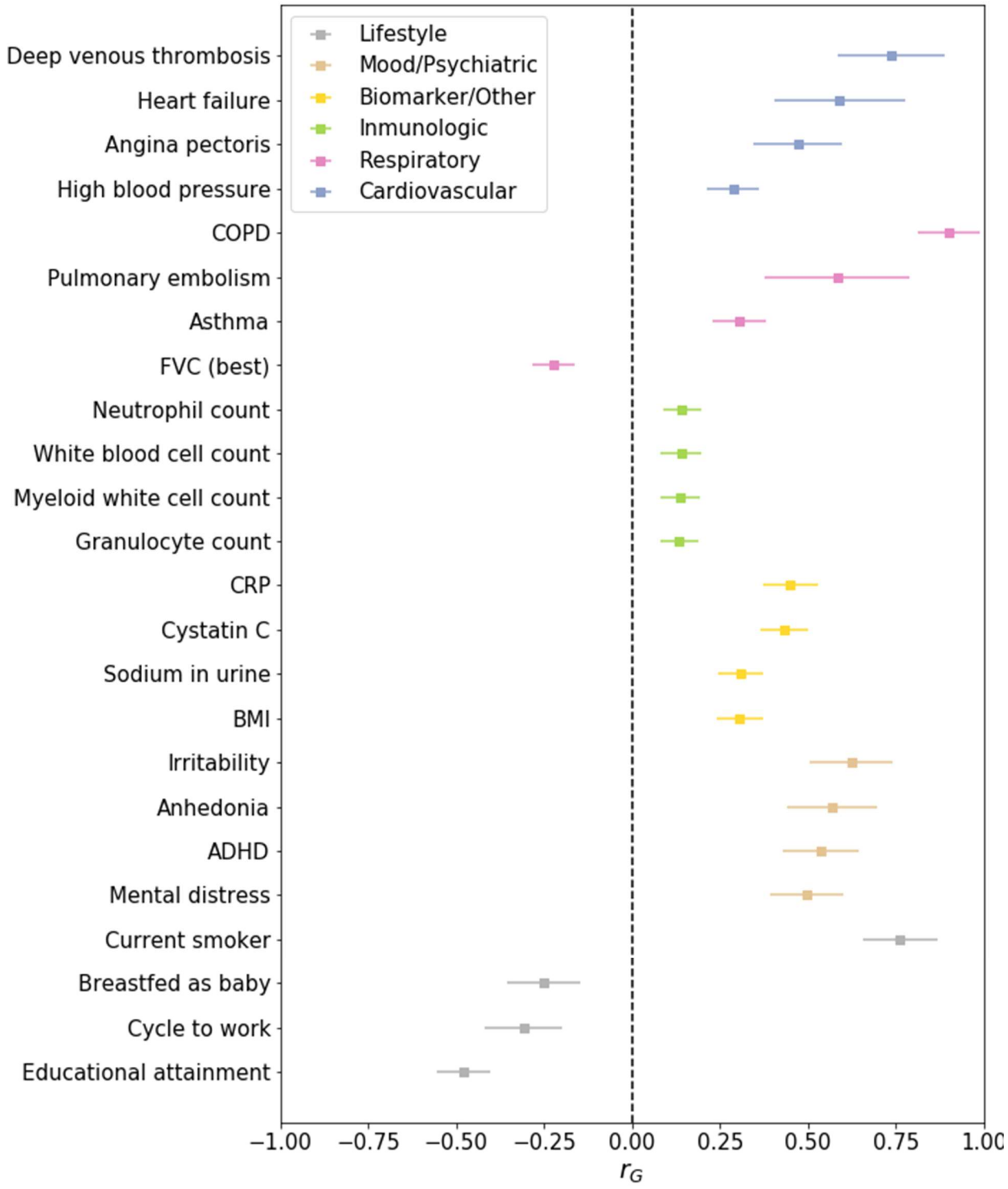

Figure 3 Pneumonia is genetically correlated with respiratory, circulatory, metabolic and lifestyle traits

Forest plot showing genetic correlations (rG) between pneumonia and traits of interest. Genetic correlations were estimated using bi-variate LD-score regression. All of the results shown are statistically significant. Due to space restrictions, the full results are available as Supplementary Data 1. Error bars represent standard errors of the genetic correlations.

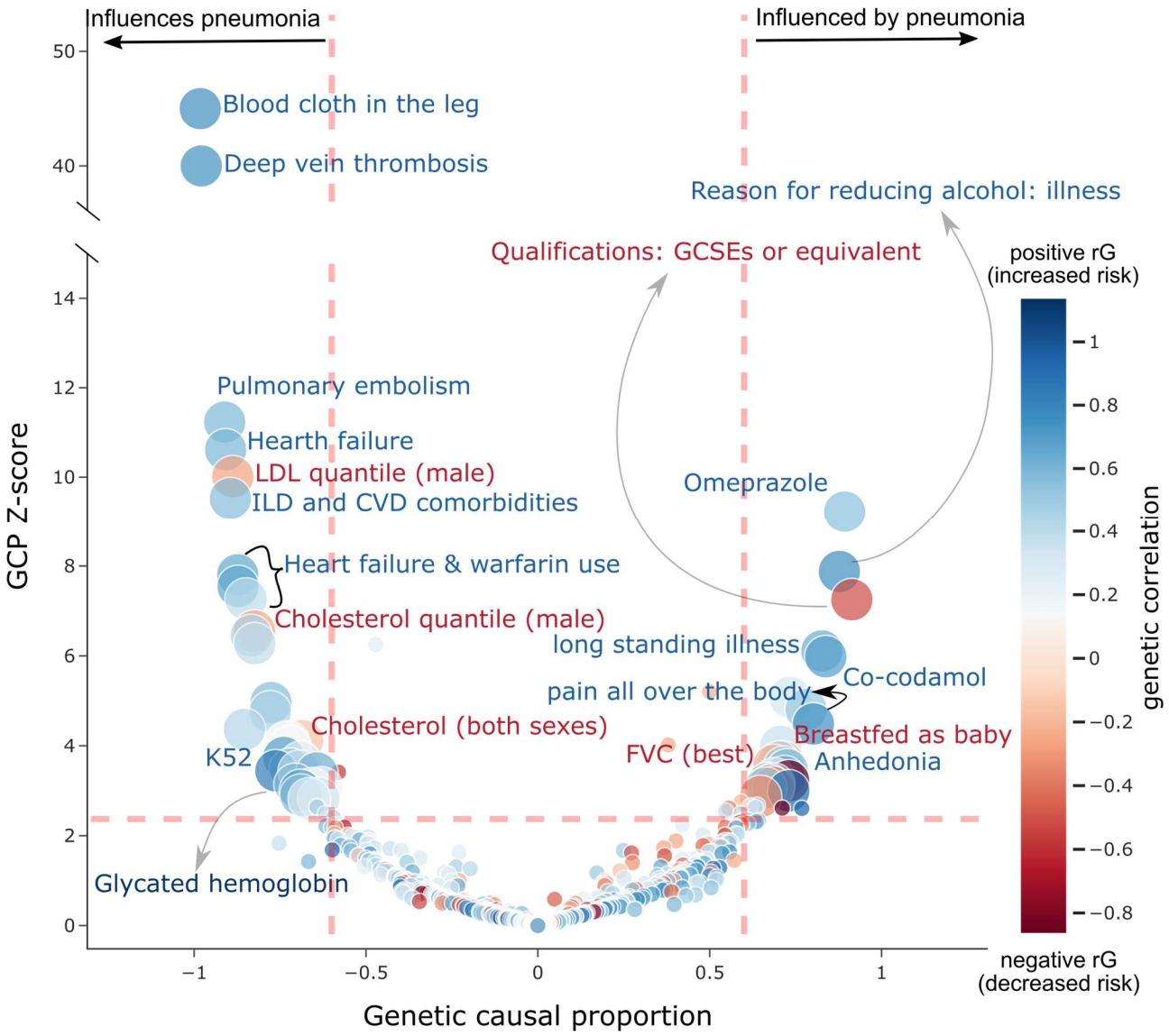

Figure 4 Pneumonia causal association analysis

Causal architecture plot showing the results of a phenome-wide latent causal variable analysis assessing the evidence for a causal association between pneumonia and other traits (see Methods). Each point represents a trait that showed a significant genetic correlation with pneumonia. The x-axis represents the genetic causal proportion; high values indicate evidence for a causal association between pneumonia and the trait of interest. Positive values indicate that pneumonia is likely to act as a risk factor for the trait (i.e. it *causes* the other trait). In contrast, negative values would highlight risk factors for pneumonia. Traits are coloured based on their genetic correlation with pneumonia and indicate the direction of the causal association (i.e. increasing risk or decreasing risk). Trait or trait category labels with a colour indicating the direction of the causal association have been added.

TABLES

Table 1. GWAS UK Biobank sample composition

	Cases	Controls	OR (95% C.I.)
Sample size	14780 (3.3%)	439096 (96.7%)	NA
Female N(%)	6490 (44%)	240059 (55%)	0.713 (0.69-0.737)
Age mean(sd)	60.4 (7.2)	56.7 (8.0)	1.06 (1.06-1.07)
Smoking history N(%)	9143 (62%)	198667 (45%)	1.74 (1.68-1.68)

Data for participants of European ancestry included in the GWAS

Table 2. Pneumonia GWAS meta analysis and sensitivity results

CHR	SNP	BP	A1	A2	FREQ	BETA	SE	P	P adj ever smoked	P adj ever smoked and cigs per day
15	**rs2009746**	78754102	A	G	0.67	-0.012	0.002	**8.08E-10**	**1.36E-10**	4.42E-03
15	**rs76474922**	78884553	A	C	0.91	0.02	0.003	**3.16E-09**	**1.85E-09**	4.53E-04
3	**rs11708673**	151584294	A	T	0.18	-0.014	0.002	6.06E-08	5.59E-08	**4.21E-08**
4	rs144242331	37127462	A	G	0.02	0.036	0.007	2.21E-07	2.23E-07	2.00E-07
1	rs1894692	169467654	A	G	0.98	-0.034	0.007	3.05E-07	1.84E-07	4.69E-07
20	rs3810478	62191475	T	G	0.65	0.01	0.002	8.29E-07	7.11E-07	3.20E-05
16	rs4787458	28531287	A	G	0.62	-0.01	0.002	8.95E-07	NA	NA
11	rs470263	102649856	T	C	0.64	-0.01	0.002	1.27E-06	1.67E-06	1.96E-06
2	rs9309718	3497661	A	G	0.26	0.01	0.002	1.47E-06	1.42E-06	4.35E-06
1	rs34517439	78450517	A	C	0.12	0.014	0.003	2.11E-06	1.90E-06	3.84E-06
17	rs62057446	32291020	T	C	0.94	0.019	0.004	2.13E-06	1.93E-06	6.49E-07
11	rs1154905	134775317	A	C	0.35	-0.009	0.002	2.44E-06	3.23E-06	1.22E-06
11	rs11606719	118782474	C	G	0.97	0.028	0.006	2.69E-06	3.11E-06	1.64E-05
6	rs200243764	11484783	G	GA	0.96	0.022	0.005	2.89E-06	2.89E-06	2.40E-06
2	rs62169465	148532638	T	C	0.23	-0.01	0.002	3.19E-06	3.31E-06	3.24E-05
9	rs150438131	93170623	A	G	0.01	0.043	0.009	3.27E-06	3.31E-06	7.38E-07
13	rs76713055	100226814	A	G	0.98	-0.033	0.007	3.57E-06	2.92E-06	1.48E-06
10	rs138075843	15339390	T	C	0.03	-0.026	0.006	3.89E-06	6.20E-06	1.07E-05
12	rs79345814	3507480	A	T	0.02	0.029	0.006	4.71E-06	6.52E-06	1.84E-05
9	rs10819081	128629174	A	C	0.38	-0.009	0.002	4.89E-06	9.21E-06	1.01E-05

Showing all SNPs with at least suggestive evidence of association with pneumonia (p<1e-5). *SNPs with genome-wide significant evidence of association (p<5e-8) are in bold. CHR- chromosome; BP- base pair position; SNP - variant identifier; A1 - effect allele; A2- non-effect allele; FREQ - effect allele frequency; BETA- Effect allele effect size; SE - effect size standard error; P - p value.

Table 3. Colocalization of lung eQTLs with pneumonia GWAS loci

Gene	COLOC Posterior Probability						
	PP0	PP1	PP2	PP3	PP4	PP3+PP4	PP4/(PP3+PP4)
EIF3C	0.023	0.028	0.151	0.183	0.614	0.797	0.771
APOBR	0.433	0.525	0.015	0.018	0.009	0.027	0.341
EIF3CL	0.411	0.499	0.038	0.046	0.005	0.052	0.102
NPIPB6	0.419	0.508	0.031	0.038	0.004	0.042	0.099
IL27	1.26E-07	1.53E-07	0.452	0.548	3.99E-04	0.548	0.001
CLN3	0.428	0.519	0.023	0.027	0.002	0.030	0.073
PSMA4	0.002	0.918	1.67E-04	0.067	0.012	0.079	0.156
CHRNA5	1.63E-13	6.52E-11	0.002	0.998	1.26E-06	0.998	1.26E-06
IREB2	0.002	0.900	2.12E-04	0.085	0.013	0.098	0.129
HYKK	0.001	0.439	2.46E-04	0.098	0.461	0.560	**0.824**
CHRNA3	3.53E-04	0.141	0.002	0.753	0.103	0.856	0.120
PBX3	3.36E-04	4.12E-05	0.538	0.066	0.396	0.461	**0.858**
ADAMTS7	0.002	0.788	0.001	0.206	0.004	0.209	0.017
MAPKAP1	0.626	0.077	0.257	0.031	0.009	0.040	0.216

PP0 - no association with gene expression and pneumonia risk; PP1- association with pneumonia GWAS only; PP2 - association with gene expression only; PP3 - association with gene expression and pneumonia GWAS, but two distinct SNP; PP4 - association with gene expression and pneumonia GWAS, shared SNP; Genes not shown could not be assessed due to lack of expression in the relevant tissue or lack of eQTL data. Genes that showed evidence of colocalization are in bold.

Table 4. Target sample (AGDS) composition and demographics

	Cases	Controls
Sample size	1206 (21%)	4389 (78%)
Female (%)*	919 (76%)	3179 (72%)
Age (sd)*	48.6 (14.6)	42.5 (14.6)
Light smokers*	955 (79%)	3058 (69%)
Pneumonia PRS (sd)*	0.14 (1.02)	0.04 (0.98)

Data for unrelated participants of European ancestry used for the replication and PRS. *P<0.05 two sample t-test.