Genome-wide association study of chronic sputum production implicates loci involved in mucus production and infection

Citation for published version:
Packer, RJ, Shrine, N, Hall, R, Melbourne, CA, Thompson, R, Williams, AT, Paynton, ML, Guyatt, AL, Allen, RJ, Lee, PH, John, C, Campbell, A, Hayward, C, de Vries, M, Vonk, JM, Davitte, J, Hessel, E, Michalovich, D, Betts, JC, Sayers, I, Yeo, A, Hall, IP, Tobin, MD & Wain, LV 2023, 'Genome-wide association study of chronic sputum production implicates loci involved in mucus production and infection', *European Respiratory Journal*. https://doi.org/10.1183/13993003.01667-2022

Digital Object Identifier (DOI):
10.1183/13993003.01667-2022

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
European Respiratory Journal

EUROPEAN RESPIRATORY *journal*

FLAGSHIP SCIENTIFIC JOURNAL OF ERS

Early View

Original research article

Genome-wide association study of chronic sputum production implicates loci involved in mucus production and infection

Richard J. Packer, Nick Shrine, Robert Hall, Carl A. Melbourne, Rebecca Thompson, Alex T. Williams, Megan L. Paynton, Anna L. Guyatt, Richard J. Allen, Paul H. Lee, Catherine John, Archie Campbell, Caroline Hayward, Maaike de Vries, Judith M. Vonk, Jonathan Davitte, Edith Hessel, David Michalovich, Joanna C. Betts, Ian Sayers, Astrid Yeo, Ian P. Hall, Martin D Tobin, Louise V. Wain

Please cite this article as: Packer RJ, Shrine N, Hall R, *et al*. Genome-wide association study of chronic sputum production implicates loci involved in mucus production and infection. *Eur Respir J* 2023; in press (https://doi.org/10.1183/13993003.01667-2022).

Genome-wide association study of chronic sputum production implicates loci involved in mucus production and infection

Richard J Packer[1,8], Nick Shrine[1], Robert Hall[2], Carl A Melbourne[1], Rebecca Thompson[2], Alex T Williams[1], Megan L Paynton[1], Anna L Guyatt[1], Richard J Allen[1], Paul H Lee[1], Catherine John[1,8], Archie Campbell[3], Caroline Hayward[4], Maaike de Vries[5], Judith M Vonk[5], Jonathan Davitte[6], Edith Hessel[7], David Michalovich[7], Joanna C Betts[7], Ian Sayers[2], Astrid Yeo[7], Ian P Hall[2], Martin D Tobin[1,8], Louise V Wain[1,8]

1. Department of Health Sciences, University of Leicester, Leicester, UK.
2. Centre for Respiratory Research, NIHR Nottingham Biomedical Research Centre, School of Medicine, Biodiscovery Institute, University of Nottingham, Nottingham, UK.
3. Centre for Genomic and Experimental Medicine, Institute of Genetics & Cancer, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK
4. Medical Research Council Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh EH4 2XU, UK
5. University of Groningen, University Medical Center Groningen, Department of Epidemiology & Groningen Research Institute for Asthma and COPD (GRIAC), Groningen, The Netherlands.
6. GSK R&D, Collegeville, PA, USA
7. GSK R&D, Stevenage, UK
8. Leicester NIHR Biomedical Research Centre, Glenfield Hospital, Leicester, UK
Corresponding Author: richard.packer@leicester.ac.uk

Abstract

Background

Chronic sputum production impacts on quality of life and is a feature of many respiratory diseases. Identification of the genetic variants associated with chronic sputum production in a disease agnostic sample could improve understanding of its causes and identify new molecular targets for treatment.

Methods

We conducted a genome-wide association study (GWAS) of chronic sputum production in UK Biobank. Signals meeting genome-wide significance ($P<5x10^{-8}$) were investigated in additional independent studies, were fine-mapped, and putative causal genes identified by gene expression analysis. GWAS of respiratory traits were interrogated to identify whether the signals were driven by existing respiratory disease amongst the cases and variants were further investigated for wider pleiotropic effects using phenome-wide association studies (PheWAS).

Findings

From a GWAS of 9,714 cases and 48,471 controls, we identified six novel genome-wide significant signals for chronic sputum production including signals in the Human Leukocyte Antigen (HLA) locus, chromosome 11 mucin locus (containing *MUC2*, *MUC5AC* and *MUC5B*) and the *FUT2* locus. The four common variant associations were supported by independent studies with a combined sample size of up to 2,203 cases and 17,627 controls. The mucin locus signal had previously been reported for association with moderate-to-severe asthma. The HLA signal was fine-mapped to an amino-acid change of threonine to arginine (frequency 36.8%) in HLA-DRB1 (HLA-*DRB1**03:147). The signal near *FUT2* was associated with expression of several genes including *FUT2,* for which the direction of effect was tissue dependent. Our PheWAS identified a wide range of associations including blood cell traits, liver biomarkers, infections, gastrointestinal and thyroid-associated diseases, and respiratory disease.

Interpretation

Novel signals at the *FUT2* and mucin loci suggest that mucin fucosylation may be a driver of chronic sputum production even in the absence of diagnosed respiratory disease and provide genetic support for this pathway as a target for therapeutic intervention.

Introduction

Increased sputum production impacts on daily activities and quality of life - and is a shared feature of many respiratory diseases. Worldwide, 545 million people have chronic respiratory conditions, with those associated with chronic sputum production including chronic obstructive pulmonary disease (COPD), asthma, bronchiectasis, chronic bronchitis, and cystic fibrosis. Chronic respiratory disease is the third leading cause of death worldwide, with 3.91 million deaths in 2017 [1].

The determinants of chronic sputum production in disease are not completely understood [2]. Most studies of excess sputum production have been in subjects with chronic bronchitis and COPD where it has been associated with lower lung function [3, 4] and higher risk of both exacerbation and respiratory symptoms [5]. Risk factors for excess sputum production include smoking and occupational and environmental pollutants [4, 6–8]. Currently available drug treatments for those with chronic sputum production do not generally affect the rate of production of sputum, but act as mucolytics and expectorants [9–11].

Genome-wide association studies have highlighted pathways underlying a range of respiratory traits and diseases, and highlighted potentially relevant drug targets [12, 13]. Previous genome wide association studies of sputum production [14–17] and have not identified any genome-wide significant findings.

We hypothesised that identifying genetic variants that are associated with chronic sputum production in a large general population sample could improve understanding of its causes and identify new molecular targets for treatment. To test this hypothesis, we undertook a genome-wide association study (GWAS) of risk of chronic sputum production in 9,714 cases and 48,471 controls from UK Biobank and sought replication of the association signals in five additional independent studies totalling 2,203 cases and 17,627 controls. We performed phenome-wide association studies (PheWAS) and interrogation of gene expression data to characterise the association signals and determine which genes may be driving these signals.

Methods

Study population

Information about chronic sputum production was obtained from the online lifetime occupation survey that was emailed to 324,653 UK Biobank participants with existing email addresses between June and September 2015 and achieved a response rate of 38% (31% of all of those contacted provided a full completion of the questionnaire [18]). For this study, we defined cases as those who answered “yes” to the question “do you bring up phlegm/sputum/mucus daily?” (UK Biobank data-field 22504, total 121,283 participants provided a “yes” or “no” response). Controls were defined as those who answered “no” to this question. Cases and controls were further restricted to those of genetically-determined European ancestry, as previously defined [19], with available smoking data (data-field 20160). Related individuals were removed, with cases preserved over controls when excluding one of a pair (or more) of related individuals (data-field 22021, “related” defined as a KING kinship coefficient ≥ 0.0884, equivalent to second-degree relatedness or closer). For related pairs within the cases or controls, the individual with the lowest genotype missingness (data-field 22005) was retained. From all available controls, we defined a subset of controls with a similar age (data-field id 34) and sex (data-field id 31) distribution to the cases at a 1:5 ratio with the cases.

Demographics and respiratory characteristics of the case and controls were derived using the following definitions: doctor-diagnosed asthma (UK Biobank data-field 22127), moderate-to-severe asthma (as previously described [20]), doctor-diagnosed chronic bronchitis (data-field 22129), cough on most days (data-field 22502), smoking status (data-field 20160), COPD Global Initiative for Chronic Obstructive Lung Disease (GOLD) stage 1-4 and stages 2-4 (defined using baseline spirometry as previously described [19], [21]) bronchiectasis and cystic fibrosis (Supplementary Tables 1 and 2).

UK Biobank has ethical approval from North West – Haydock Research Ethics Committee (21/NW/0157). Written informed consent was provided by all participants.

Genome-wide association study of chronic sputum production

Genetic data from the v3 March 2018 UK Biobank data release, imputed to the Haplotype Reference Consortium panel r1.1 2016, was used for the genome-wide association study giving 27,317,434 variants for analysis.

Association testing was performed using logistic regression under an additive genetic model in PLINK 2.0 [22] with age, sex, array version, never/ever smoking status and the first 10 principal components of ancestry as covariates. Variants were excluded if they had an imputation quality INFO score <0.5 or a minor allele count (MAC) <20. Association signals were considered genome-wide significant at $P<5x10^{-8}$. Independent signals were initially defined using a 1Mb window (500kb each side of the sentinel variant) and then using conditional analyses implemented in GCTA-COJO [23]. All variant coordinates are for genome build GRCh37. Region plots were created using LocusZoom [24].

Replication

We sought replication in five general population cohorts which surveyed participants for chronic sputum production; Generation Scotland [25], EXCEED Study [26], LifeLines 1, LifeLines 2 and Vlagtwedde-Vlaardingen[17]. Further details are provided in the Supplementary text.

In addition, the overlap of primary care sputum codes with the chronic sputum production question (UK Biobank data-field 22504) was evaluated to identify whether primary care codes could be used to define an additional independent case-control dataset from those in UK Biobank who did not respond to the online lifetime occupation survey (Supplementary text).

Fine-mapping

We undertook Bayesian fine-mapping (29) for all genome-wide significant signals that were not in the HLA region to define 99% credible sets of variants i.e. sets of variants that are 99% probable to contain the true causal variant (assuming that it has been measured).

To fine-map signals within the HLA region (chr6:29,607,078-33,267,103 (b37)) to a specific HLA gene allele or amino acid change, we re-imputed our discovery samples using IMPUTE2 v2.3.1 with a reference panel that enabled imputation of 424 classical HLA alleles and 1,276 amino acid changes as described in [27]. We then repeated the association testing as described above.

Mapping association signals to putative causal genes

We used functional annotation and co-localisation with expression Quantitative Trait Loci (eQTL) signals to identify putative causal genes at each signal.

Annotation of the variants in each credible set was performed using SIFT [28], PolyPhen-2 and CADD, all implemented using the Ensemble GRCh37 Variant Effect Predictor (VEP) [29], alongside FATHMM [30]. Variants were annotated as deleterious if they were labelled deleterious by SIFT, probably damaging or possibly damaging by PolyPhen-2, damaging by FATHMM (specifying the "Inherited Disease" option of the "Coding Variants" method, and using the "Unweighted" prediction algorithm) or had a CADD scaled score ≥20.

We queried the sentinel variants in GTEx V8 [31] and BLUEPRINT [32] (see Supplementary Table 3 for list of tissues). We tested for colocalisation of GWAS and eQTL signals using coloc [33]; H4 >80% was used to define a shared causal variant for eQTL and GWAS signals.

Associations with other phenotypes

To investigate whether the signals of association with sputum production were driven by underlying respiratory phenotypes of the cases, a look-up for each signal was undertaken for fourteen respiratory or respiratory-related traits from GWAS results (moderate-to-severe asthma (N cases=5,135, controls=25,675) [20], lung function (Forced Expired Volume in 1 second [FEV_1], Forced Vital Capacity [FVC], FEV_1/FVC, peak expiratory flow (PEF)) (N=400,102) [19], respiratory infection (N cases=19,459, controls=101,438) [34], chronic cough (N cases=15,213, controls=94,731), chronic bronchitis (N cases=977, controls = 108,967), idiopathic pulmonary fibrosis (IPF) (N cases=2,668, controls=8,591) [35], smoking traits (Smoking age-of-onset (N=124,590), smoking cessation (N cases=141,649, controls=27321), smoking cigarettes-per-day (N=120,744), smoking initiation (N cases=170,772, controls=212,859) and asthma (N cases=23,948, controls=118,538) [36]). Smoking trait results were from the UK Biobank component of [37]; chronic cough and chronic bronchitis were defined for this study using UK Biobank data, see Supplementary text. Where the sentinel variant was not available in the look-up dataset, we utilised an alternative variant from the credible set with the highest posterior probability of being causal. A Bonferroni adjustment for 84 association tests was applied requiring a $P < 5.95\times10^{-4}$ for association to be classified as statistically significant. Imputed HLA gene allele or amino acid changes were used for signals in the HLA region.

To investigate associations of the chronic sputum-associated variants with a wider range of phenotypes, we performed PheWAS for 2,172 traits in UK Biobank (FDR<0.01, Supplementary Text) and searched the Open Targets Genetics Portal ($P<5\times10^{-8}$, version 0.4.0 (bd664ca) - accessed 16th April 2021[38]). PheWAS for imputed HLA alleles was performed using DeepPheWAS [39] (see Supplementary text).

Sensitivity analyses

To further investigate whether the effects of the variants associated with risk of chronic sputum production differ between ever and never smokers, or between individuals with and without a history of chronic respiratory disease (spirometry defined COPD GOLD1+, doctor diagnosed asthma or doctor diagnosed chronic bronchitis), we tested association of sentinel variants in ever and never smokers and those with and without evidence of chronic respiratory disease separately. We additionally evaluated whether the associations differed between males and females or by the time of year of the survey (UK Biobank data-field 22500). Finally, we evaluated whether adjusting for current smoking (UK Biobank data-field 22506) (rather than ever vs never smoker status) affected the results.

Results

A total of 10,481 participants answered "yes" to the question "Do you bring up phlegm/sputum/mucus daily?" and 110,802 answered "no" (Supplementary Table 4). After excluding those with missing genotype and essential covariate data, and those of genetically determined European ancestry, a total of 9,714 cases and 48,471 controls (Figure 1) were included in the GWAS. Ever smoking and respiratory disease were more common in the cases than in the controls (Table 1). The genomic control inflation factor (lambda) was 1.026 so no adjustments to the test statistics were applied (Supplementary Figure 1). Six independent novel signals met the genome-wide significance threshold of $P<5x10^{-8}$ (Table 2 and Supplementary Figure 2). These were four common variant signals (minor allele frequency > 5%) in or near *MUC2, FUT2,* HLA-*DRB1* and *NKX3-1,* and two intronic rare variant signals (minor allele frequency < 1%) in *OCIAD1* and *NELL1* (Figure 2).

No systematic differences were seen in effect sizes when stratifying by smoking status, by history of chronic respiratory disease, by sex, by time of year of survey or when including current smoking status as a covariate (Supplementary Table 5, Supplementary Figures 3 to 8) for the six sentinel variants. Through comparison of survey responses and linked primary care data we showed that primary care codes were not adequate proxies for the survey responses (Supplementary Text). We sought replication in five independent cohorts with a combined sample size to 1977 cases and 17,627 controls; data from all five replication cohorts were only available for the *FUT2* locus. Although none of the signals met criteria for significance in a meta-analysis of the replication cohorts, the directions of effect were consistent with the discovery results for the signals in or near *MUC2, FUT2, OCIAD1,* HLA-*DRB1* and *NKX3-1* and all except the signals at *NELL1* and HLA-*DRB1* also increased in significance when the replication and discovery results were meta-analysed (Table 2, Supplementary Table 13, and Supplementary Figure 9).

Novel associations with chronic sputum production

HLA *locus*

The HLA signal was fine-mapped to an amino-acid change of threonine to arginine (frequency 36.8%) at codon 233 of exon 5 of *HLA-DRB1* (HLA-*DRB1**03:147) that was associated with decreased risk (OR 0.91 [95% C.I. 0.88-0.94]) of chronic sputum production ($P=3.43\times10^{-9}$). The amino acid change was in linkage disequilibrium with the GWAS sentinel variant rs374248993 (linkage disequilibrium

R^2=0.74 with HLA-*DRB1**03:147) and the signal for rs374248993 was attenuated when conditioned on the amino acid change (Supplementary Figures 10 and 11).

HLA-*DRB1**03:147 was significantly associated with FEV_1, FEV_1/FVC and PEF at genome-wide significance ($P<5x10^{-8}$) (Figure 3 and Supplementary Table 6). The amino acid associated with increased risk of chronic sputum production (threonine) was associated with increased lung function; this had not been previously reported. The HLA PheWAS identified multiple significant associations for the HLA allele associated with increased risk of chronic sputum production with a wide range of quantitative traits (for example, blood cell traits, liver biomarkers) and diseases (including decreased risk of gastrointestinal and thyroid-associated diseases, and increased risk of bronchiectasis and asthma) (Supplementary Table 7).

MUC2 locus

For the mucin locus signal (rs779167905 allele T), the allele associated with risk of chronic sputum production was also significantly associated with increased risk of asthma (OR 1.06, P=0.0027) and moderate-to-severe asthma (OR 1.13, P=$6.3x10^{-7}$), increased FVC (beta 0.0087, P=$6x10^{-4}$) and decreased risk of IPF (OR 0.84, P=$7.5x10^{-6}$) (Figure 3, Supplementary Table 6). There were no associations with gene expression for rs779167905 in GTEx or BLUEPRINT. However, we have previously shown that a proxy of rs779167905 (rs11602802, r2=XX) was associated with mRNA levels of MUC5AC in bronchial epithelial brush samples collected from asthma patients, with the risk allele being associated with elevated MUC5AC expression [ref].

Genome-wide significant associations with IPF [40] and moderate-severe asthma [20] have previously been reported at this chromosome 11 locus and so we undertook a conditional analysis to identify whether the chronic sputum production signal was independent of these previous signals. Repeating the association testing for this variant conditioning on the previously reported variants (rs35705950 [40] and rs11603634 [20]) identified that the chronic sputum production GWAS signal was independent of the IPF signal (rs779167905, conditional P=$1.18x10^{-10}$) but was not independent of the previously reported moderate-to-severe asthma signal (rs779167905, conditional P=0.0039) (Supplementary Figures 12 and 13). Furthermore, the IPF association for rs779167905 (using proxy SNP rs10902094) was also attenuated when conditioned on rs35705950 (OR 0.99, P=0.784).

Our PheWAS and Open Targets Genetics Portal analysis identified that the *MUC2* locus signal (rs779167905) allele that was associated with increased risk of chronic sputum production (allele T) was associated with higher risk of asthma and asthma-related traits in other studies [41–43] and with lower risk of gall-bladder disease (Supplementary Table 7 and 8).

FUT2 locus

The *FUT2* credible set included two variants that were annotated as functional using VEP. This included a stop-gain variant in *FUT2* (rs601338, linkage disequilibrium r2 0.992 with sentinel rs492602) and a nearby missense variant (rs602662 r2 0.882 with sentinel rs492602) that resulted in a Glycine to Serine amino acid change for the allele positively correlated with the chronic sputum production risk allele (Supplementary Tables 9 and 10).

Sentinel variant rs492602 at the *FUT2* locus was associated with gene expression for *FUT2*, *NTN5, RASIP1, SEC1P* and *MAMSTR* for which there was support for co-localisation of eQTL and GWAS signals in multiple tissues from GTEx V8 (Figure 4, Supplementary Table 11). Increased risk of chronic sputum production was consistently correlated with increased expression of *NTN5* and *MAMSTR* across a range of tissues. In contrast, the direction of the *FUT2* expression signal varied by tissue with increased risk of chronic sputum production correlated with decreased expression of *FUT2* in brain tissues and with increased expression in gastrointestinal tissue. There were no associations in lung tissue and upper airway tissues were not available.

The sentinel variant for the *FUT2* region signal on chromosome 19 (rs492602) was associated with lung function measures FEV_1/FVC and PEF ($P=2.2x10^{-6}$ and $P=1.1x10^{-6}$, respectively), with the chronic sputum production risk allele (G) associated with decreased lung function (Figure 3, Supplementary Table 6).

Our PheWAS and Open Targets Genetics Portal analysis for this variant identified 141 associations spanning multiple disease areas, phenotypes, and biomarkers (Supplementary Tables 7 and 8). In summary, the allele associated with increased risk of chronic sputum production was associated with increased risk of gallstones [42, 44, 45], type 1 diabetes [46] and Crohn's disease [47–50], elevated vitamin B12 [51–54] and cholesterol and fat metabolites [41, 42, 55–59], hypertension/cardiovascular disease [42, 44, 60], excess alcohol with associated sequelae [44, 61–63], increased risk of mumps and lower risk of childhood ear infections [64]. Higher risk of chronic sputum production was also associated with higher levels of gamma glutamyl transferase, total bilirubin and aspartate amino transferase, and lower levels of alanine aminotransferase and alkaline phosphatase.

Other novel loci

Using functional annotation of variants and eQTL analysis, no putative causal genes could be assigned to the signals in or near *OCIAD1* and *NELL1*. There was a single co-localising eQTL for *SLC25A37* in the *NKX3-1* locus with increased risk of chronic sputum production associated with a reduced expression of *SLC25A37* in brain cortex (Supplementary Table 11, Supplementary Figure 14).

Discussion

We describe a GWAS of chronic sputum production to identify genome-wide significant signals and our novel findings implicate genes involved in mucin production and fucosylation, as well as the HLA class II histocompatibility antigen, HLA-DRB1. We provide functional evidence that the SNP signals we identify are associated with gene expression of *FUT2, MUC5AC* and *SLC5A37.*

Smoking is believed to be the main cause of excess sputum production, and is also associated with chronic infections, reduced lung function and susceptibility to chronic respiratory disease. Through identification of genetic association signals that are independent of smoking and history of chronic respiratory disease, our study demonstrates the value in studying a disease-relevant phenotype in a very large population that is agnostic to respiratory disease or smoking status.

The most significant signal implicated the gene *FUT2* which has been widely studied for its role in blood group antigen expression and association with gastric and respiratory infection. *FUT2* encodes fucosyltransferase 2 which mediates the transfer of fucose to the terminal galactose on glycan chains of cell surface glycoproteins and glycolipids. FUT2 creates a soluble precursor oligosaccharide FuC-alpha ((1,2)Galbeta-) called the H antigen which is an essential substrate for the final step in the soluble ABO blood group antigen synthesis pathway. The *FUT2* locus allele associated with increased risk of chronic sputum production in this study is correlated with a nonsense allele that leads to inactivated FUT2, which results in a non-secretory phenotype of ABO(H) blood group antigens [65] for homozygous carriers. This nonsense allele (rs601338 allele A) has frequencies of 25-50% in South Asian, European and African populations but is rare (<1%) in East Asian populations [66]. Candidate gene studies of this locus have identified that non-secretors (at increased risk of chronic sputum production according to our study) have a lower risk of H. Pylori infection [67], rotavirus A infection [68, 69], norovirus infection [70–72], infant (12-24 months) respiratory illness [73], asthma exacerbations [74], otitis media [75], exacerbation in non-cystic fibrosis bronchiectasis and *Pseudomonas aeruginosa* airway infection in the same group [76], some evidence of slower HIV progression [72] and a higher risk of pneumococcal and meningococcal infection [77]. The T allele of another variant in high linkage disequilibrium at this locus (rs681343, r^2=0.996 with rs492602), associated with increased risk of chronic sputum production in our study, was recently reported to be associated with increased risk of human polyomavirus 1 (BKV) virus infection, as measured by antibody response [78]. A recent GWAS of critically ill cases of COVID-19 (cases N=7491), showed that the risk allele for chronic mucus production (G) of rs492602 was protective against life threating COVID-19 (P=4.55x10^{-9}, OR 0.88, CI 0.87-0.90) [79]. However, this finding was not replicated in the latest COVID-19 Host Genetics Initiative results for a similar phenotype [80]. The differing directions of effect of this signal on different phenotypes may be explained by the SNP effects on FUT2 expression which differ across cell and tissue types. Further targeted experiments in relevant cell and tissue types would be needed to elucidate this and define the likely effects of targeting FUT2 directly or indirectly.

Epitopes that are fucosylated by FUT2 play a role in cell-cell interaction including host-microbe interaction [81, 82] and mediate interaction with intestinal microbiota, thereby influencing its composition [83–86]. Whilst there has been no direct evidence of host-pathogen binding on the FUT2 generated epitopes for non-gastrointestinal infection there is evidence that FUT2 can influence non-binding ligands such as sialic acid [87]. Sialic acid binding has been shown to be important for adenovirus binding in cell models [88] and modulating this binding has been implicated as a possible mechanism for increasing risk of mumps infection [64].

FUT2 may also be key to the function of mucins, including those encoded by genes at our other significant locus (i.e. *MUC2*, *MUC5AC*, *MUC5B*). Mucins are a major constituent of airway mucus and MUC5AC is major gel-forming mucin secreted by airway epithelial cells. FUT2 may play a key role in MUC5AC regulation leading to excess mucus production or its increased viscosity; a common characteristic observed in patients with airway obstructive diseases including asthma, bronchitis, and COPD. Analysis of oligosaccharides released from insoluble colonic mucins, largely Muc2, by mass spectrometry shows complete lack of terminal fucosylation of *O*-linked oligosaccharides in Fut2-LacZ-null mice [89]. FUT2 has also been shown to determine the *O*-glycosylation pattern of Muc5ac in mice [90]. The significant signal at *MUC2* in our analysis was not independent of the previously reported moderate-to-severe asthma signal [20] for which *MUC5AC* was implicated as the most likely causal gene using gene expression data from bronchial epithelial cells. In that study we went on to show that the signal (rs11602802 used as proxy) was associated with mRNA levels of MUC5AC in bronchial epithelial brush samples collected from asthma patients, with the risk allele being associated with elevated MUC5AC. There was also a non-significant trend for MUC5B to have a reduced mRNA level in the presence of the moderate-severe asthma risk allele. These *ex vivo* observations have recently been replicated in nasal epithelial cell brush samples in an independent cohort and extended to show this signal (rs12788104 within the credible set of *MUC2* signal) regulates MUC5AC protein levels *in vitro* using nasal epithelial cells from genotyped subjects in the air liquid interface model [91]. Although our analysis did not identify an association at the *MUC2* locus with COPD-related traits (FEV_1 and FEV_1/FVC), a recent study has also highlighted MUC5AC as a potential biomarker for COPD prognosis [92].

The particular allele that was found to explain the association signal in the HLA region (HLA-DRB1*03:147 [93], has only recently been reported and so there is limited information about functionality. Associations of this allele with other GWAS traits should be interpreted with caution given the high LD across the region. Furthermore, the association of this allele with increased sputum production and increased lung function reminds us that increased sputum production is part of the adaptive immune response to environmental insult, and approaches to target mucus production must also consider potential negative effects of reducing sputum production.

We only report overlap of chronic sputum production association signals with association signals for gene expression regulation where there is statistical support that these signals share a causal variant. In addition to a comprehensive PheWAS, we provide a deeper assessment of associations with relevant respiratory phenotypes that highlights previously unreported associations with lung function for the *HLA-DRB1* and *FUT2* signals.

As only a subset of UK Biobank participants provided answers to the sputum production question, we expected that we might be able to define a replication case control dataset from the remaining >300,000 participants using primary care data. However, evaluation of the positive predictive value of primary care codes for sputum production, when compared to the questionnaire data, was very low (see Supplementary Text). This could reflect a low utilisation of sputum codes in primary care or that participants have not reported this symptom to their General Practitioner (GP). We obtained supportive evidence for four of the signals utilising data from five general population cohorts. The limited sample size (the case sample size for replication was 23% of the size available for discovery) impacted our ability to show statistically significant replication. Furthermore, we note that, for three of the replication cohorts (LIfeLines 1 and 2 and Vlagtwedde-Vlaardingen), the sputum production question asked specifically about winter symptoms whilst the UK Biobank survey did not restrict to any specific season. However, given the strong evidence summarised above for the involvement of the probable causal genes in control of pathways relevant to mucus production, we believe the associations identified are highly likely to be real. Due to very low numbers, we were unable to evaluate the effects of these signals in individuals of non-European ancestry thereby limiting the generalisability of our findings to non-European ancestry groups. Efforts are urgently needed to improve diversity in genomics research [94] such as the planned Our Future Health initiative in the UK. In summary, the HLA, *MUC2* and *FUT2* loci show strong candidacy for a role in sputum production, with overlap with infection and related phenotypes and known mechanistic interactions between the genes at the *FUT2* and *MUC2* loci, suggesting that these signals are likely to be robust. The large number of associations of the *FUT2* locus with a broad array of phenotypes, tissue-dependent expression of *FUT2,* and association with expression of other genes in the region, may have implications for drug targeting guided by this locus. Experimental studies to characterise the specific interplay between FUT2 activity and mucin genes expressed in the airways are warranted.

Conclusion

Chronic sputum production is a phenotype characteristic of several respiratory diseases, as well as being common cause for referrals in the absence of overt disease and is of interest for pharmaceutical intervention. We report novel genetic factors which influence chronic sputum production, and these signals highlight fucosylation of mucin as a driving factor of chronic sputum production. These signals could provide insight into the molecular pathways of sputum production and represent potential future targets for drug development [95].

Data availability

Genome-wide association statistics from the case-control analysis of chronic sputum production will be made available via GWAS Catalog [to be submitted following peer-review].

Funding and Acknowledgements

L.V.W. holds a GSK / Asthma + Lung UK Chair in Respiratory Research (C17-1). M.D.T. is supported by a Wellcome Trust Investigator Award (WT202849/Z/16/Z). M.D.T. and I.P.H. hold NIHR Senior Investigator Awards. C.J. held a Medical Research Council Clinical Research Training Fellowship (MR/P00167X/1). LVW, MDT, IS and IPH report collaborative research funding from GSK to undertake the submitted work.

The research was partially supported by the NIHR Leicester Biomedical Research Centre and the NIHR Nottingham Biomedical Research Centre; the views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. This research was funded in part by the Wellcome Trust. We acknowledge the support of the Health Data Research UK BREATHE Digital Innovation Hub (UKRI Award MC_PC_19004). This research was conducted under UK Biobank application 45243. This research used the SPECTRE and ALICE High Performance Computing Facility at the University of Leicester.

Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates [CZD/16/6] and the Scottish Funding Council [HR03006] and is currently supported by the Wellcome Trust [216767/Z/19/Z]. Genotyping of the GS:SFHS samples was carried out by the Genetics Core Laboratory at the Edinburgh Clinical Research Facility, University of Edinburgh, Scotland and was funded by the Medical Research Council UK and the Wellcome Trust (Wellcome Trust Strategic Award "STratifying Resilience and Depression Longitudinally" (STRADL) Reference 104036/Z/14/Z). CH is supported by a Medical Research Council University Unit Programme grant MC_UU_00007/10 (QTL in Health and Disease).

Recruitment to the Generation Scotland CovidLife study was facilitated by SHARE - the Scottish Health Research Register and Biobank.

SHARE is supported by NHS Research Scotland, the Universities of Scotland and the Chief Scientist Office of the Scottish Government.

Competing interests

LVW, MDT, IS and IPH report collaborative research funding from GSK to undertake the submitted work. LVW, MDT, CJ, ALG, and RJP report funding from Orion Pharma outside of the submitted work. LVW reports consultancy for Galapagos. JD, EH, DM, JCB and AY were employees of GSK at the time of this study. DM is an employee of Benevolent AI; CAM is an employee of Mirador Analytics. NS, RH, RT, ATW, MLP, PHL, AC, CH, MV and JMV report no competing interests.

Table 1 Demographics, ever-smoking status, doctor-diagnosed asthma, doctor-diagnosed chronic bronchitis, cough, moderate-to-severe asthma and COPD GOLD stage 1-4 status of cases and controls included in the GWAS of chronic sputum production. *Total 6942 cases and 36321 controls with available spirometry that passed QC.

	Cases N=9714	Controls N=48471
Mean age (years)	57.7	57.7
% female	42.5	42.5
Ever smoked (%)	5306 **(54.6)**	20912 **(43.1)**
Current smoker (%)	983 **(10.2)**	1569**(3.2)**
Doctor-diagnosed chronic bronchitis (%)	407 **(4.2)**	416 **(0.86)**
Doctor-diagnosed asthma (%)	2630 **(27.1)**	5251 **(10.8)**
Cough on most days (%)	7022 **(72.3)**	3999 **(8.3)**
Moderate-to-severe asthma (%)	520 **(5.4)**	521 **(1.1)**
Self-reported chronic sinusitis (%)	181 (1.9)	1057 (2.2)
Meets spirometry criteria GOLD 1-4 (%)	1511 **(21.8)***	4766 **(13.1)***

Table 2: Novel genome-wide significant signals of association with chronic sputum production.

Chr:Position (GRCh37)	RSID	Locus (BP distance from gene)*	coded/non-coded	Coded allele frequency % (count)**		OR (95 CI)	P	Imputation quality (INFO)***	# variants in 99% credible set (highest posterior probability)
4:48854355	rs79998532	*OCIAD1* (intronic)	A/G	0.2% (233)	Discovery	2.36 (1.76-3.16)	$8.00\text{x}10^{-09}$	0.92	3 (0.86)
					Replication	3.3 (0.11-98.6)	0.49		
					Meta-analysis	2.37 (1.77-3.17)	$6.36\text{x}10^{-09}$		
6:32496534	rs374248993	*HLA-DRB1* ǂ	G/C	57% (66355)	Discovery	1.12 (1.08-1.16)	$7.30\text{x}10^{-11}$	0.87	*HLA-DRB1**03:147ǂ
					Replication	1.01 (0.84-1.21)	0.93		
					Meta-analysis	1.11 (1.08-1.15)	$1.31\text{x}10^{-10}$		
8:23480686	rs79401075	*NKX3-1* (59,765)	A/G	10% (11620)	Discovery	1.18 (1.12-1.24)	$8.90\text{x}10^{-11}$	0.98	30 (0.32)
					Replication	1.20 (0.95-1.52)	0.12		
					Meta-analysis	1.18 (1.12-1.24)	$2.65\text{x}10^{-11}$		
11:1116931	rs779167905	*MUC2* (12,513)	T/TTCTA	67% (78158)	Discovery	1.12 (1.08-1.16)	$1.20\text{x}10^{-10}$	0.98	30 (0.15)
					Replication	1.09 (0.93-1.28)	0.29		
					Meta-analysis	1.12 (1.08-1.15)	$6.99\text{x}10^{-11}$		
11:20887601	rs529240826	*NELL1* (intronic)	GC/G	0.51% (588)	Discovery	1.91 (1.52-2.4)	$2.50\text{x}10^{-08}$	0.67	2 (0.83)
					Replication	0.83 (0.38-1.78)	0.63		

					Meta-analysis	1.79 (1.44-2.22)	1.99x10^{-07}		
19:49206417	rs492602	*FUT2* (exonic)	G/A	51% (58803)	Discovery	1.11 (1.08-1.15)	3.20x10^{-11}	1	32 (0.07)
					Replication	1.06 (1-1.14)	0.07		
					Meta-analysis	1.10 (1.07-1.13)	1.21x10^{-11}		

ǂ = amino-acid change of threonine to arginine at codon 233 of exon 5 of *HLA-DRB1 (*HLA gene allele HLA-*DRB1**03:147)

BP = base pairs; OR = odds ratio; CI = confidence interval

*Start or end of nearest gene

** Values for discovery SNPs.

*** INFO score taken from discovery.

Figure captions

Figure 1 Study flow chart detailing case control selection from the UK Biobank cohort.

Figure 2 LocusZoom plots of the six sentinel signals, **a).** *OCIAD1* signal (rs79998532), **b).** *HLA-DRB5* signal (rs374248993), **c).** *NKX3-1* signal (rs79401075), **d).** *MUC2* signal (rs779167905), **e).** *NELL1* signal (rs529240826), and **f).** *FUT2* signal (rs492602)

Figure 3 Results for association of sentinel variant risk alleles with respiratory traits. Results are aligned to the risk allele for chronic sputum production, effect direction 'Increasing' can be read as increasing risk for binary traits and increasing values in quantitative traits. Chronic bronchitis and smoking age of onset, cigarettes per day and cessation phenotype lookups were omitted as no associations with P<0.05 found. *P $<5.95 \times 10^{-4}$ (Bonferroni adjustment for 84 association tests) ** P$<5 \times 10^{-8}$. Note that the IPF association for rs779167905 (using proxy SNP rs10902094) was attenuated when conditioned on rs35705950 (OR 0.99, P=0.784).

Figure 4 Results for eQTL colocalization for the *FUT2* locus using variant **rs492602.** The numbers within the grid are the posterior probability of colocalization (H4), with results aligned to the risk allele G for the **rs492602** variant. Missing numbers indicate no data was available for the respective gene and tissue.

References

1. Li X, Cao X, Guo M, Xie M, Liu X. Trends and risk factors of mortality and disability adjusted life years for chronic respiratory diseases from 1990 to 2017: systematic analysis for the Global Burden of Disease Study 2017. *BMJ* 2020; : m234.

2. Kim V, Criner GJ. Chronic Bronchitis and Chronic Obstructive Pulmonary Disease. *Am J Respir Crit Care Med* 2013; 187: 228–237.

3. Kim V, Zhao H, Boriek AM, Anzueto A, Soler X, Bhatt SP, Rennard SI, Wise R, Comellas A, Ramsdell JW, Kinney GL, Han MK, Martinez CH, Yen A, Black-Shinn J, Porszasz J, Criner GJ, Hanania NA, Sharafkhaneh A, Crapo JD, Make BJ, Silverman EK, Curtis JL, COPDGene Investigators. Persistent and Newly Developed Chronic Bronchitis Are Associated with Worse Outcomes in Chronic Obstructive Pulmonary Disease. *Ann Am Thorac Soc* 2016; 13: 1016–1025.

4. Pelkonen M, Notkola I-L, Nissinen A, Tukiainen H, Koskela H. Thirty-year cumulative incidence of chronic bronchitis and COPD in relation to 30-year pulmonary function and 40-year mortality: a follow-up in middle-aged rural men. *Chest* 2006; 130: 1129–1137.

5. Kim V, Han MK, Vance GB, Make BJ, Newell JD, Hokanson JE, Hersh CP, Stinson D, Silverman EK, Criner GJ, COPDGene Investigators. The chronic bronchitic phenotype of COPD: an analysis of the COPDGene Study. *Chest* 2011; 140: 626–633.

6. Dijkstra AE, de Jong K, Boezen HM, Kromhout H, Vermeulen R, Groen HJM, Postma DS, Vonk JM. Risk factors for chronic mucus hypersecretion in individuals with and without COPD: influence of smoking and job exposure on CMH. *Occup Environ Med* 2014; 71: 346–352.

7. Trupin L, Earnest G, San Pedro M, Balmes JR, Eisner MD, Yelin E, Katz PP, Blanc PD. The occupational burden of chronic obstructive pulmonary disease. *Eur Respir J* 2003; 22: 462–469.

8. Matheson MC, Benke G, Raven J, Sim MR, Kromhout H, Vermeulen R, Johns DP, Walters EH, Abramson MJ. Biological dust exposure in the workplace is a risk factor for chronic obstructive pulmonary disease. *Thorax* 2005; 60: 645–651.

9. Tarrant BJ, Le Maitre C, Romero L, Steward R, Button BM, Thompson BR, Holland AE. Mucoactive agents for chronic, non-cystic fibrosis lung disease: A systematic review and meta-analysis: Mucoactive agents in chronic non-CF management. *Respirology* 2017; 22: 1084–1092.

10. Rubin BK. Mucolytics, expectorants, and mucokinetic medications. *Respir Care* 2007; 52: 859–865.

11. Shen Y, Huang S, Kang J, Lin J, Lai K, Sun Y, Xiao W, Yang L, Yao W, Cai S, Huang K, Wen F. Management of airway mucus hypersecretion in chronic airway inflammatory disease: Chinese expert consensus (English edition). *Int J Chron Obstruct Pulmon Dis* 2018; 13: 399–407.

12. Wain LV, Shrine N, Artigas MS, Erzurumluoglu AM, Noyvert B, Bossini-Castillo L, Obeidat M, Henry AP, Portelli MA, Hall RJ, Billington CK, Rimington TL, Fenech AG, John C, Blake T, Jackson VE, Allen RJ, Prins BP, Understanding Society Scientific Group, Campbell A, Porteous DJ, Jarvelin M-R, Wielscher M, James AL, Hui J, Wareham NJ, Zhao JH, Wilson JF, Joshi PK, Stubbe B, et al. Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nat Genet* 2017; 49: 416–425.

13. El-Husseini ZW, Gosens R, Dekker F, Koppelman GH. The genetics of asthma and the promise of genomics-guided drug target discovery. *Lancet Respir Med* 2020; 8: 1045–1056.

14. Dijkstra AE, Boezen HM, van den Berge M, Vonk JM, Hiemstra PS, Barr RG, Burkart KM, Manichaikul A, Pottinger TD, Silverman EK, Cho MH, Crapo JD, Beaty TH, Bakke P, Gulsvik A, Lomas DA, Bossé Y, Nickle DC, Paré PD, de Koning HJ, Lammers J-W, Zanen P, Smolonska J, Wijmenga C, Brandsma C-A, Groen HJM, Postma DS, the LifeLines Cohort Study group. Dissecting the genetics of chronic mucus hypersecretion in smokers with and without COPD. *Eur Respir J* 2015; 45: 60–75.

15. Dijkstra AE, Smolonska J, van den Berge M, Wijmenga C, Zanen P, Luinge MA, Platteel M, Lammers J-W, Dahlback M, Tosh K, Hiemstra PS, Sterk PJ, Spira A, Vestbo J, Nordestgaard BG, Benn M, Nielsen SF, Dahl M, Verschuren WM, Picavet HSJ, Smit HA, Owsijewitsch M, Kauczor HU, de Koning HJ, Nizankowska-Mogilnicka E, Mejza F, Nastalek P, van Diemen CC, Cho MH, Silverman EK, et al. Susceptibility to Chronic Mucus Hypersecretion, a Genome Wide Association Study. Hartl D, editor. *PLoS ONE* 2014; 9: e91621.

16. Lee JH, Cho MH, Hersh CP, McDonald M-LN, Crapo JD, Bakke PS, Gulsvik A, Comellas AP, Wendt CH, Lomas DA, Kim V, Silverman EK, COPDGene and ECLIPSE Investigators. Genetic susceptibility for chronic bronchitis in chronic obstructive pulmonary disease. *Respir Res* 2014; 15: 113.

17. Zeng X, Vonk JM, de Jong K, Xu X, Huo X, Boezen HM. No convincing association between genetic markers and respiratory symptoms: results of a GWA study. *Respir Res* 2017; 18: 11.

18. De Matteis S, Jarvis D, Young H, Young A, Allen N, Potts J, Darnton A, Rushton L, Cullinan P. Occupational self-coding and automatic recording (OSCAR): a novel web-based tool to collect and code lifetime job histories in large population-based studies. *Scand J Work Environ Health* 2017; 43: 181–186.

19. Shrine N, Guyatt AL, Erzurumluoglu AM, Jackson VE, Hobbs BD, Melbourne CA, Batini C, Fawcett KA, Song K, Sakornsakolpat P, Li X, Boxall R, Reeve NF, Obeidat M, Zhao JH, Wielscher M, Weiss S, Kentistou KA, Cook JP, Sun BB, Zhou J, Hui J, Karrasch S, Imboden M, Harris SE, Marten J, Enroth S, Kerr SM, Surakka I, Vitart V, et al. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat Genet* 2019; 51: 481–493.

20. Shrine N, Portelli MA, John C, Soler Artigas M, Bennett N, Hall R, Lewis J, Henry AP, Billington CK, Ahmad A, Packer RJ, Shaw D, Pogson ZEK, Fogarty A, McKeever TM, Singapuri A, Heaney LG, Mansur AH, Chaudhuri R, Thomson NC, Holloway JW, Lockett GA, Howarth PH, Djukanovic R, Hankinson J, Niven R, Simpson A, Chung KF, Sterk PJ, Blakey JD, et al. Moderate-to-severe asthma in individuals of European ancestry: a genome-wide association study. *The Lancet Respiratory Medicine* 2019; 7: 20–34.

21. Global Initiative for Chronic Obstructive Lung Disease (GOLD), pocket guide to COPD diagnosis, management, and prevention. A Guide for health care professionals. [Internet]. 2019Available from: https://goldcopd.org/wp-content/uploads/2018/11/GOLD-2019-POCKET-GUIDE-FINAL_WMS.pdf.

22. Purcell, Shaun C Christopher. Plink 2.0 [Internet]. Available from: www.cog-genomics.org/plink/2.0/.

23. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics* 2011; 88: 76–82.

24. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 2010; 26: 2336–2337.

25. Smith BH, Campbell H, Blackwood D, Connell J, Connor M, Deary IJ, Dominiczak AF, Fitzpatrick B, Ford I, Jackson C, Haddow G, Kerr S, Lindsay R, McGilchrist M, Morton R, Murray G, Palmer CN, Pell JP, Ralston SH, St Clair D, Sullivan F, Watt G, Wolf R, Wright A, Porteous D, Morris AD. Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Med Genet* 2006; 7: 74.

26. John C, Reeve NF, Free RC, Williams AT, Ntalla I, Farmaki A-E, Bethea J, Barton LM, Shrine N, Batini C, Packer R, Terry S, Hargadon B, Wang Q, Melbourne CA, Adams EL, Bee CE, Harrington K, Miola J, Brunskill NJ, Brightling CE, Barwell J, Wallace SE, Hsu R, Shepherd DJ, Hollox EJ, Wain LV, Tobin MD. Cohort profile: Extended Cohort for E-health, Environment and DNA (EXCEED). *International Journal of Epidemiology* 2019; : dyz175.

27. Jia X, Han B, Onengut-Gumuscu S, Chen W-M, Concannon PJ, Rich SS, Raychaudhuri S, de Bakker PIW. Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS ONE* 2013; 8: e64683.

28. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research* 2012; 40: W452–W457.

29. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. *Genome Biol* 2016; 17: 122.

30. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, Day INM, Gaunt TR. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* 2013; 34: 57–65.

31. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* 2017; 550: 204–213.

32. Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martín D, Watt S, Yan Y, Kundu K, Ecker S, Datta A, Richardson D, Burden F, Mead D, Mann AL, Fernandez JM, Rowlston S, Wilder SP, Farrow S, Shao X, Lambourne JJ, Redensek A, Albers CA, Amstislavskiy V, Ashford S, Berentsen K, Bomba L, Bourque G, Bujold D, Busche S, et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* 2016; 167: 1398-1414.e24.

33. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, Plagnol V. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. Williams SM, editor. *PLoS Genet* 2014; 10: e1004383.

34. Williams AT, Shrine N, Naghra-van Gijzel H, Betts JC, Hessel EM, John C, Packer R, Reeve NF, Yeo AJ, Abner E, Åsvold BO, Auvinen J, Bartz TM, Bradford Y, Brumpton B, Campbell A, Cho MH, Chu S, Crosslin DR, Feng Q, Esko T, Gharib SA, Hayward C, Hebbring S, Hveem K, Jarvelin M-R, Jarvik GP, Landis SH, Larson EB, Liu J, et al. Genome-wide association study of susceptibility to hospitalised respiratory infections. *Wellcome Open Res* 2021; 6: 290.

35. Allen RJ, Guillen-Guio B, Oldham JM, Ma S-F, Dressen A, Paynton ML, Kraven LM, Obeidat M, Li X, Ng M, Braybrooke R, Molina-Molina M, Hobbs BD, Putman RK, Sakornsakolpat P, Booth HL, Fahy WA, Hart SP, Hill MR, Hirani N, Hubbard RB, McAnulty RJ, Millar AB, Navaratnam V, Oballa E, Parfrey H, Saini G, Whyte MKB, Zhang Y, Kaminski N, et al. Genome-Wide Association Study of Susceptibility to Idiopathic Pulmonary Fibrosis. *Am J Respir Crit Care Med* 2019; : rccm.201905-1017OC.

36. Demenais F, Bisgaard H, Barnes KC, Cookson WOC, Altmüller J, Ang W, Barr RG, Beaty TH, Becker AB, Beilby J, Bisgaard H, Bjornsdottir US, Bleecker E, Bønnelykke K, Boomsma DI, Bouzigon E, Brightling CE, Brossard M, Brusselle GG, Burchard E, Burkart KM, Bush A, Chan-Yeung M, Chung KF, Couto Alves A, Curtin JA, Custovic A, Daley D, de Jongste JC, Del-Rio-Navarro BE, et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat Genet* 2018; 50: 42–53.

37. Jiang Y, Li Y, Brazel DM, Chen F, Datta G, Davila-Velderrain J, McGuire D, Tian C, Zhan X, Choquet H, Docherty AR, Faul JD, Foerster JR, Fritsche LG, Gabrielsen ME, Gordon SD, Haessler J, Hottenga J-J, Huang H, Jang S-K, Jansen PR, Ling Y, Mägi R, Matoba N, McMahon G, Mulas A, Orrù V, Palviainen T, Pandit A, Reginsson GW, 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* 2019; 51: 237–244.

38. Ghoussaini M, Mountjoy E, Carmona M, Peat G, Schmidt EM, Hercules A, Fumis L, Miranda A, Carvalho-Silva D, Buniello A, Burdett T, Hayhurst J, Baker J, Ferrer J, Gonzalez-Uriarte A, Jupp S, Karim MA, Koscielny G, Machlitt-Northen S, Malangone C, Pendlington ZM, Roncaglia P, Suveges D, Wright D, Vrousgou O, Papa E, Parkinson H, MacArthur JAL, Todd JA, Barrett JC, et al. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res* 2021; 49: D1311–D1320.

39. Packer R, Williams A, Hennah W, Eisenberg M, Fawcett K, Pearson W, Guyatt A, Edris A, Hollox E, Rao B, Bratty J, Wain L, Dudbridge F, Tobin M. Deep-PheWAS: a pipeline for phenotype generation and association analysis for phenome-wide association studies [Internet]. Genetic and Genomic Medicine; 2022 MayAvailable from: http://medrxiv.org/lookup/doi/10.1101/2022.05.05.22274419.

40. Seibold MA, Wise AL, Speer MC, Steele MP, Brown KK, Loyd JE, Fingerlin TE, Zhang W, Gudmundsson G, Groshong SD, Evans CM, Garantziotis S, Adler KB, Dickey BF, du Bois RM, Yang IV, Herron A, Kervitsky D, Talbert JL, Markin C, Park J, Crews AL, Slifer SH, Auerbach S, Roy MG, Lin J, Hennessy CE, Schwarz MI, Schwartz DA. A Common *MUC5B* Promoter Polymorphism and Pulmonary Fibrosis. *N Engl J Med* 2011; 364: 1503–1512.

41. Wu Y, Byrne EM, Zheng Z, Kemper KE, Yengo L, Mallett AJ, Yang J, Visscher PM, Wray NR. Genome-wide association study of medication-use and associated disease in the UK Biobank. *Nat Commun* 2019; 10: 1891.

42. UK Biobank GWAS V2 results [Internet]. 2018 [cited 2020 Aug 3].Available from: http://www.nealelab.is/uk-biobank/.

43. Pividori M, Schoettler N, Nicolae DL, Ober C, Im HK. Shared and distinct genetic risk factors for childhood-onset and adult-onset asthma: genome-wide and transcriptome-wide studies. *Lancet Respir Med* 2019; 7: 509–522.

44. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, LeFaive J, VandeHaar P, Gagliano SA, Gifford A, Bastarache LA, Wei W-Q, Denny JC, Lin M, Hveem K, Kang HM, Abecasis GR, Willer CJ, Lee S. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* 2018; 50: 1335–1341.

45. Ferkingstad E, Oddsson A, Gretarsdottir S, Benonisdottir S, Thorleifsson G, Deaton AM, Jonsson S, Stefansson OA, Norddahl GL, Zink F, Arnadottir GA, Gunnarsson B, Halldorsson GH, Helgadottir A, Jensson BO, Kristjansson RP, Sveinbjornsson G, Sverrisson DA, Masson G, Olafsson I, Eyjolfsson GI, Sigurdardottir O, Holm H, Jonsdottir I, Olafsson S, Steingrimsdottir T, Rafnar T, Bjornsson ES, Thorsteinsdottir U, Gudbjartsson DF, et al. Genome-wide association meta-analysis yields 20 loci associated with gallstone disease. *Nat Commun* 2018; 9: 5101.

46. Onengut-Gumuscu S, Chen W-M, Burren O, Cooper NJ, Quinlan AR, Mychaleckyj JC, Farber E, Bonnie JK, Szpak M, Schofield E, Achuthan P, Guo H, Fortune MD, Stevens H, Walker NM, Ward LD, Kundaje A, Kellis M, Daly MJ, Barrett JC, Cooper JD, Deloukas P, Type 1 Diabetes Genetics Consortium, Todd JA, Wallace C, Concannon P, Rich SS. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* 2015; 47: 381–386.

47. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, Ripke S, Lee JC, Jostins L, Shah T, Abedian S, Cheon JH, Cho J, Dayani NE, Franke L, Fuyuno Y, Hart A, Juyal RC, Juyal G, Kim WH, Morris AP, Poustchi H, Newman WG, Midha V, Orchard TR, Vahedi H, Sood A, Sung JY, Malekzadeh R, Westra H-J, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* 2015; 47: 979–986.

48. de Lange KM, Moutsianas L, Lee JC, Lamb CA, Luo Y, Kennedy NA, Jostins L, Rice DL, Gutierrez-Achury J, Ji S-G, Heap G, Nimmo ER, Edwards C, Henderson P, Mowat C, Sanderson J, Satsangi J, Simmons A, Wilson DC, Tremelling M, Hart A, Mathew CG, Newman WG, Parkes M, Lees CW, Uhlig H, Hawkey C, Prescott NJ, Ahmad T, Mansfield JC, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* 2017; 49: 256–261.

49. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, Anderson CA, Bis JC, Bumpstead S, Ellinghaus D, Festen EM, Georges M, Green T, Haritunians T, Jostins L, Latiano A, Mathew CG, Montgomery GW, Prescott NJ, Raychaudhuri S, Rotter JI, Schumm P, Sharma Y, Simms LA, Taylor KD, Whiteman D, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn’s disease susceptibility loci. *Nat Genet* 2010; 42: 1118–1125.

50. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Schumm LP, Sharma Y, Anderson CA, Essers J, Mitrovic M, Ning K, Cleynen I, Theatre E, Spain SL, Raychaudhuri S, Goyette P, Wei Z, Abraham C, Achkar J-P, Ahmad T, Amininejad L, Ananthakrishnan AN, Andersen V, Andrews JM, Baidoo L, Balschun T, Bampton PA, Bitton A, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 2012; 491: 119–124.

51. Hazra A, Kraft P, Selhub J, Giovannucci EL, Thomas G, Hoover RN, Chanock SJ, Hunter DJ. Common variants of FUT2 are associated with plasma vitamin B12 levels. *Nat. Genet.* 2008; 40: 1160–1162.

52. Nongmaithem SS, Joglekar CV, Krishnaveni GV, Sahariah SA, Ahmad M, Ramachandran S, Gandhi M, Chopra H, Pandit A, Potdar RD, H D Fall C, Yajnik CS, Chandak GR. GWAS identifies population-specific new regulatory variants in FUT6 associated with plasma B12 concentrations in Indians. *Hum. Mol. Genet.* 2017; 26: 2551–2564.

53. Tanaka T, Scheet P, Giusti B, Bandinelli S, Piras MG, Usala G, Lai S, Mulas A, Corsi AM, Vestrini A, Sofi F, Gori AM, Abbate R, Guralnik J, Singleton A, Abecasis GR, Schlessinger D, Uda M, Ferrucci L. Genome-wide association study of vitamin B6, vitamin B12, folate, and homocysteine blood concentrations. *Am. J. Hum. Genet.* 2009; 84: 477–482.

54. Hazra A, Kraft P, Lazarus R, Chen C, Chanock SJ, Jacques P, Selhub J, Hunter DJ. Genome-wide significant predictors of metabolites in the one-carbon metabolism pathway. *Hum. Mol. Genet.* 2009; 18: 4677–4687.

55. Klarin D, Damrauer SM, Cho K, Sun YV, Teslovich TM, Honerlaw J, Gagnon DR, DuVall SL, Li J, Peloso GM, Chaffin M, Small AM, Huang J, Tang H, Lynch JA, Ho Y-L, Liu DJ, Emdin CA, Li AH, Huffman JE, Lee JS, Natarajan P, Chowdhury R, Saleheen D, Vujkovic M, Baras A, Pyarajan S, Di Angelantonio E, Neale BM, Naheed A, et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat Genet* 2018; 50: 1514–1523.

56. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, Ganna A, Chen J, Buchkovich ML, Mora S, Beckmann JS, Bragg-Gresham JL, Chang H-Y, Demirkan A, Den Hertog HM, Do R, Donnelly LA, Ehret GB, Esko T, Feitosa MF, Ferreira T, Fischer K, Fontanillas P, Fraser RM, Freitag DF, Gurdasani D, Heikkilä K, Hyppönen E, Isaacs A, Jackson AU, et al. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* 2013; 45: 1274–1283.

57. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, Pirruccello JP, Ripatti S, Chasman DI, Willer CJ, Johansen CT, Fouchier SW, Isaacs A, Peloso GM, Barbalic M, Ricketts SL, Bis JC, Aulchenko YS, Thorleifsson G, Feitosa MF, Chambers J, Orho-Melander M, Melander O, Johnson T, Li X, Guo X, Li M, Shin Cho Y, Jin Go M, Jin Kim Y, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 2010; 466: 707–713.

58. Weiss FU, Schurmann C, Guenther A, Ernst F, Teumer A, Mayerle J, Simon P, Völzke H, Radke D, Greinacher A, Kuehn J-P, Zenker M, Völker U, Homuth G, Lerch MM. Fucosyltransferase 2 (FUT2) non-secretor status and blood group B are associated with elevated serum lipase activity in asymptomatic subjects, and an increased risk for chronic pancreatitis: a genetic association study. *Gut* 2015; 64: 646–656.

59. Hoffmann TJ, Theusch E, Haldar T, Ranatunga DK, Jorgenson E, Medina MW, Kvale MN, Kwok P-Y, Schaefer C, Krauss RM, Iribarren C, Risch N. A large electronic-health-record-based genome-wide study of serum lipids. *Nat. Genet.* 2018; 50: 401–413.

60. Wu Y, Byrne EM, Zheng Z, Kemper KE, Yengo L, Mallett AJ, Yang J, Visscher PM, Wray NR. Genome-wide association study of medication-use and associated disease in the UK Biobank. *Nat Commun* 2019; 10: 1891.

61. Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, Datta G, Davila-Velderrain J, McGuire D, Tian C, Zhan X, 23andMe Research Team, HUNT All-In Psychiatry, Choquet H, Docherty AR, Faul JD, Foerster JR, Fritsche LG, Gabrielsen ME, Gordon SD, Haessler J, Hottenga J-J, Huang H, Jang S-K, Jansen PR, Ling Y, Mägi R, Matoba N, McMahon G, Mulas A, 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* 2019; 51: 237–244.

62. Chambers JC, Zhang W, Sehmi J, Li X, Wass MN, Van der Harst P, Holm H, Sanna S, Kavousi M, Baumeister SE, Coin LJ, Deng G, Gieger C, Heard-Costa NL, Hottenga J-J, Kühnel B, Kumar V, Lagou V, Liang L, Luan J, Vidal PM, Mateo Leach I, O'Reilly PF, Peden JF, Rahmioglu N, Soininen P, Speliotes EK, Yuan X, Thorleifsson G, Alizadeh BZ, et al. Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nat. Genet.* 2011; 43: 1131–1138.

63. Sanchez-Roige S, Palmer AA, Fontanillas P, Elson SL, 23andMe Research Team, the Substance Use Disorder Working Group of the Psychiatric Genomics Consortium, Adams MJ, Howard DM, Edenberg HJ, Davies G, Crist RC, Deary IJ, McIntosh AM, Clarke T-K. Genome-Wide Association Study Meta-Analysis of the Alcohol Use Disorders Identification Test (AUDIT) in Two Population-Based Cohorts. *Am J Psychiatry* 2019; 176: 107–118.

64. Tian C, Hromatka BS, Kiefer AK, Eriksson N, Noble SM, Tung JY, Hinds DA. Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nat Commun* 2017; 8: 599.

65. Kelly RJ, Rouquier S, Giorgi D, Lennon GG, Lowe JB. Sequence and Expression of a Candidate for the Human *Secretor* Blood Group α(1,2)Fucosyltransferase Gene (*FUT2*): HOMOZYGOSITY FOR AN ENZYME-INACTIVATING NONSENSE MUTATION COMMONLY CORRELATES WITH THE NON-SECRETOR PHENOTYPE. *J. Biol. Chem.* 1995; 270: 4640–4649.

66. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 2015; 526: 68–74.

67. Ikehara Y, Nishihara S, Yasutomi H, Kitamura T, Matsuo K, Shimizu N, Inada K, Kodera Y, Yamamura Y, Narimatsu H, Hamajima N, Tatematsu M. Polymorphisms of two fucosyltransferase genes (Lewis and Secretor genes) involving type I Lewis antigens are associated with the presence of anti-Helicobacter pylori IgG antibody. *Cancer Epidemiol. Biomarkers Prev.* 2001; 10: 971–977.

68. Imbert-Marcille B-M, Barbé L, Dupé M, Le Moullac-Vaidye B, Besse B, Peltier C, Ruvoën-Clouet N, Le Pendu J. A FUT2 Gene Common Polymorphism Determines Resistance to Rotavirus A of the P[8] Genotype. *The Journal of Infectious Diseases* 2014; 209: 1227–1230.

69. Payne DC, Currier RL, Staat MA, Sahni LC, Selvarangan R, Halasa NB, Englund JA, Weinberg GA, Boom JA, Szilagyi PG, Klein EJ, Chappell J, Harrison CJ, Davidson BS, Mijatovic-Rustempasic S, Moffatt MD, McNeal M, Wikswo M, Bowen MD, Morrow AL, Parashar UD. Epidemiologic Association Between *FUT2* Secretor Status and Severe Rotavirus Gastroenteritis in Children in the United States. *JAMA Pediatr* 2015; 169: 1040.

70. Larsson MM, Rydell GEP, Grahn A, Rodríguez-Díaz J, Åkerlind B, Hutson AM, Estes MK, Larson G, Svensson L. Antibody Prevalence and Titer to Norovirus (Genogroup II) Correlate with Secretor *(FUT2)* but Not with ABO Phenotype or Lewis *(FUT3)* Genotype. *J INFECT DIS* 2006; 194: 1422–1427.

71. Ruvoën-Clouet N, Belliot G, Le Pendu J. Noroviruses and histo-blood groups: the impact of common host genetic polymorphisms on virus transmission and evolution: Noroviruses and herd innate protection. *Rev. Med. Virol.* 2013; 23: 355–366.

72. Carlsson B, Kindberg E, Buesa J, Rydell GE, Lidón MF, Montava R, Mallouh RA, Grahn A, Rodríguez-Díaz J, Bellido J, Arnedo A, Larson G, Svensson L. The G428A Nonsense Mutation in

FUT2 Provides Strong but Not Absolute Protection against Symptomatic GII.4 Norovirus Infection. Lopman BA, editor. *PLoS ONE* 2009; 4: e5593.

73. Barton SJ, Murray R, Lillycrop KA, Inskip HM, Harvey NC, Cooper C, Karnani N, Zolezzi IS, Sprenger N, Godfrey KM, Binia A. *FUT2* Genetic Variants and Reported Respiratory and Gastrointestinal Illnesses During Infancy. *The Journal of Infectious Diseases* 2019; 219: 836–843.

74. Innes AL, McGrath KW, Dougherty RH, McCulloch CE, Woodruff PG, Seibold MA, Okamoto KS, Ingmundson KJ, Solon MC, Carrington SD, Fahy JV. The H antigen at epithelial surfaces is associated with susceptibility to asthma exacerbation. *Am. J. Respir. Crit. Care Med.* 2011; 183: 189–194.

75. Santos-Cortez RLP, Chiong CM, Frank DN, Ryan AF, Giese APJ, Bootpetch Roberts T, Daly KA, Steritz MJ, Szeremeta W, Pedro M, Pine H, Yarza TKL, Scholes MA, Llanes EG d.V., Yousaf S, Friedman N, Tantoco MaLC, Wine TM, Labra PJ, Benoit J, Ruiz AG, de la Cruz RAR, Greenlee C, Yousaf A, Cardwell J, Nonato RMA, Ray D, Ong KMC, So E, Robertson CE, et al. FUT2 Variants Confer Susceptibility to Familial Otitis Media. *The American Journal of Human Genetics* 2018; 103: 679–690.

76. Taylor SL, Woodman RJ, Chen AC, Burr LD, Gordon DL, McGuckin MA, Wesselingh S, Rogers GB. *FUT2* genotype influences lung function, exacerbation frequency and airway microbiota in non-CF bronchiectasis. *Thorax* 2017; 72: 304–310.

77. Blackwell CC, Jónsdóttir K, Hanson M, Todd WTA, Chaudhuri AKR, Mathew B, Brettle RP, Weir DM. Non-secretion of abo antigens predisposing to infection by Neisseria Meningitidis and Streptococcus Pneumoniae. *The Lancet* 1986; 328: 284–285.

78. Kachuri L, Francis SS, Morrison M, Boss&eacute Y, Cavazos TB, Rashkin SR, Ziv E, Witte JS. The landscape of host genetic factors involved in infection to common viruses and SARS-CoV-2 [Internet]. Genetic and Genomic Medicine; 2020 MayAvailable from: http://medrxiv.org/lookup/doi/10.1101/2020.05.01.20088054.

79. Kousathanas A, Pairo-Castineira E, Rawlik K, Stuckey A, Odhams CA, Walker S, Russell CD, Malinauskas T, Wu Y, Millar J, Shen X, Elliott KS, Griffiths F, Oosthuyzen W, Morrice K, Keating S, Wang B, Rhodes D, Klaric L, Zechner M, Parkinson N, Siddiq A, Goddard P, Donovan S, Maslove D, Nichol A, Semple MG, Zainy T, Maleady-Crowe F, Todd L, et al. Whole-genome sequencing reveals host factors underlying critical COVID-19. *Nature* 2022; 607: 97–103.

80. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature* [Internet] 2021 [cited 2021 Sep 10]; Available from: http://www.nature.com/articles/s41586-021-03767-x.

81. Lindesmith L, Moe C, Marionneau S, Ruvoen N, Jiang X, Lindblad L, Stewart P, LePendu J, Baric R. Human susceptibility and resistance to Norwalk virus infection. *Nat. Med.* 2003; 9: 548–553.

82. Borén T, Falk P, Roth KA, Larson G, Normark S. Attachment of Helicobacter pylori to human gastric epithelium mediated by blood group antigens. *Science* 1993; 262: 1892–1895.

83. Wacklin P, Mäkivuokko H, Alakulppi N, Nikkilä J, Tenkanen H, Räbinä J, Partanen J, Aranko K, Mättö J. Secretor genotype (FUT2 gene) is strongly associated with the composition of Bifidobacteria in the human intestine. *PLoS ONE* 2011; 6: e20113.

84. Wacklin P, Tuimala J, Nikkilä J, Sebastian Tims null, Mäkivuokko H, Alakulppi N, Laine P, Rajilic-Stojanovic M, Paulin L, de Vos WM, Mättö J. Faecal microbiota composition in adults is associated with the FUT2 gene determining the secretor status. *PLoS ONE* 2014; 9: e94863.

85. Rausch P, Rehman A, Künzel S, Häsler R, Ott SJ, Schreiber S, Rosenstiel P, Franke A, Baines JF. Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and FUT2 (Secretor) genotype. *Proc. Natl. Acad. Sci. U.S.A.* 2011; 108: 19030–19035.

86. Galeev A, Suwandi A, Cepic A, Basu M, Baines JF, Grassl GA. The role of the blood group-related glycosyltransferases FUT2 and B4GALNT2 in susceptibility to infectious disease. *International Journal of Medical Microbiology* 2021; 311: 151487.

87. Cohen M, Hurtado-Ziola N, Varki A. ABO blood group glycans modulate sialic acid recognition on erythrocytes. *Blood* 2009; 114: 3668–3676.

88. Walters RW, Pilewski JM, Chiorini JA, Zabner J. Secreted and Transmembrane Mucins Inhibit Gene Transfer with AAV4 More Efficiently than AAV5. *J. Biol. Chem.* 2002; 277: 23709–23713.

89. Hurd EA, Holmén JM, Hansson GC, Domino SE. Gastrointestinal mucins of Fut2-null mice lack terminal fucosylation without affecting colonization by Candida albicans. *Glycobiology* 2005; 15: 1002–1007.

90. Magalhães A, Rossez Y, Robbe-Masselot C, Maes E, Gomes J, Shevtsova A, Bugaytsova J, Borén T, Reis CA. Muc5ac gastric mucin glycosylation is shaped by FUT2 activity and functionally impacts Helicobacter pylori binding. *Sci Rep* 2016; 6: 25575.

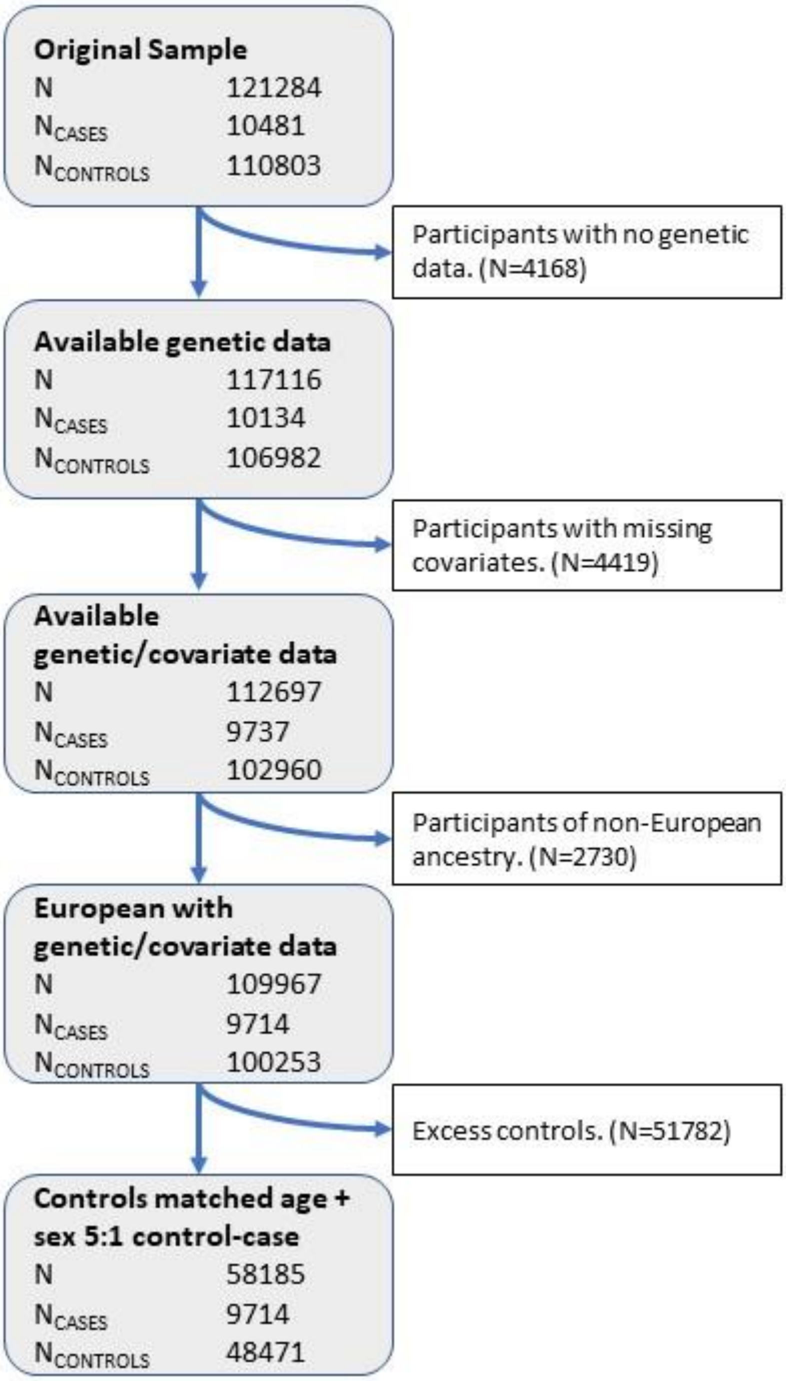
91. Sajuthi SP, Everman JL, Jackson ND, Saef B, Rios CL, Moore CM, Mak ACY, Eng C, Fairbanks-Mahnke A, Salazar S, Elhawary J, Huntsman S, Medina V, Nickerson DA, Germer S, Zody MC, Abecasis G, Kang HM, Rice KM, Kumar R, Zaitlen NA, Oh S, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Rodríguez-Santana J, Burchard EG, Seibold MA. Nasal airway transcriptome-wide association study of asthma reveals genetically driven mucus pathobiology. *Nat Commun* 2022; 13: 1632.

92. Radicioni G, Ceppe A, Ford AA, Alexis NE, Barr RG, Bleecker ER, Christenson SA, Cooper CB, Han MK, Hansel NN, Hastie AT, Hoffman EA, Kanner RE, Martinez FJ, Ozkan E, Paine R, Woodruff PG, O'Neal WK, Boucher RC, Kesimer M. Airway mucin MUC5AC and MUC5B concentrations and the initiation and progression of chronic obstructive pulmonary disease: an analysis of the SPIROMICS cohort. *Lancet Respir Med* 2021; : S2213-2600(21)00079-5.

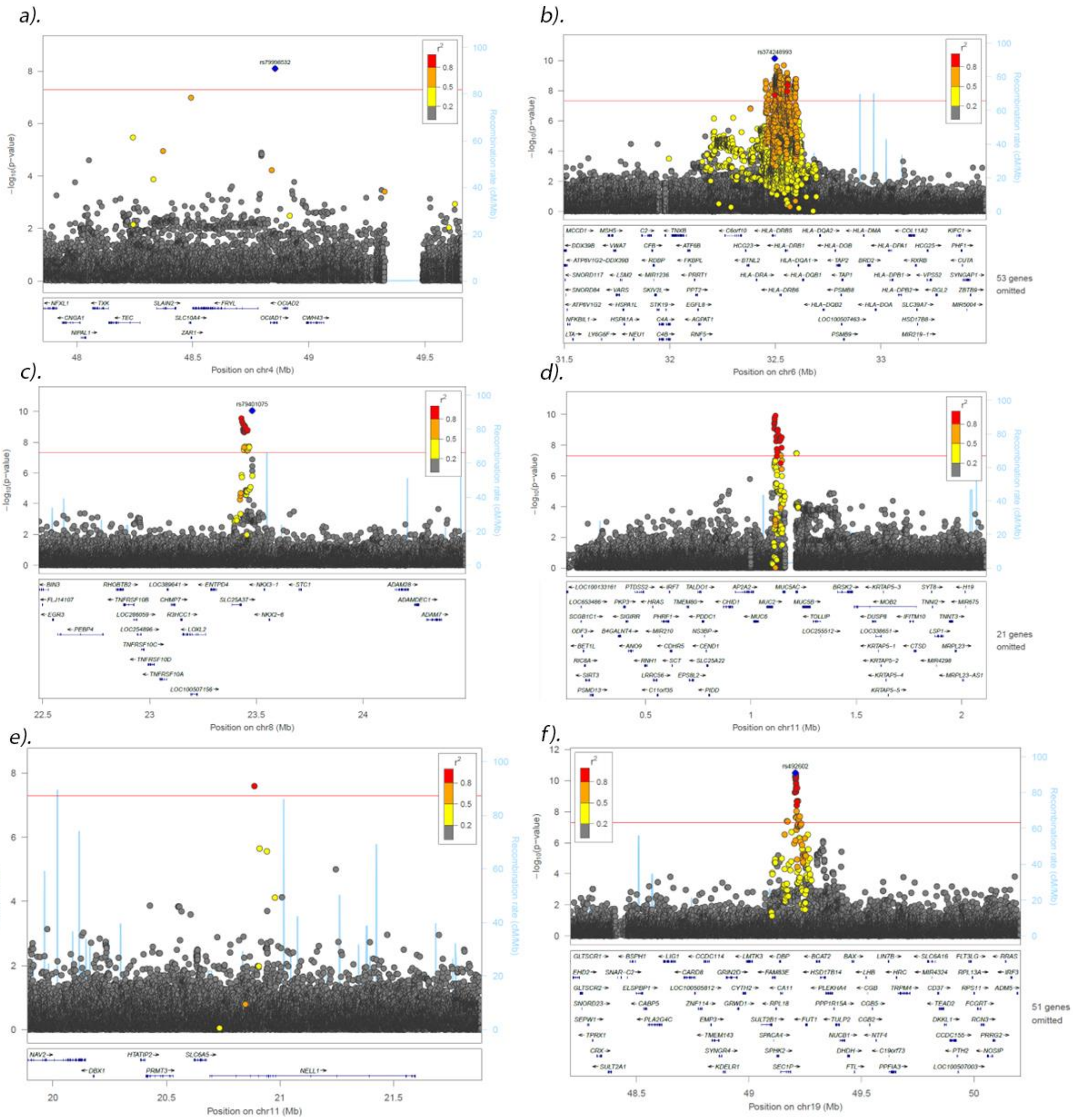
93. Ralazamahaleo M, Elsermans V, Top I, Guidicelli G, Visentin J. Characterization of the novel *HLA-DRB1*03:147* allele by sequencing-based typing. *HLA* 2019; 93: 53–54.

94. Tobin MD, Izquierdo AG. Improving ethnic diversity in respiratory genomics research. *Eur Respir J* 2021; 58: 2101615.

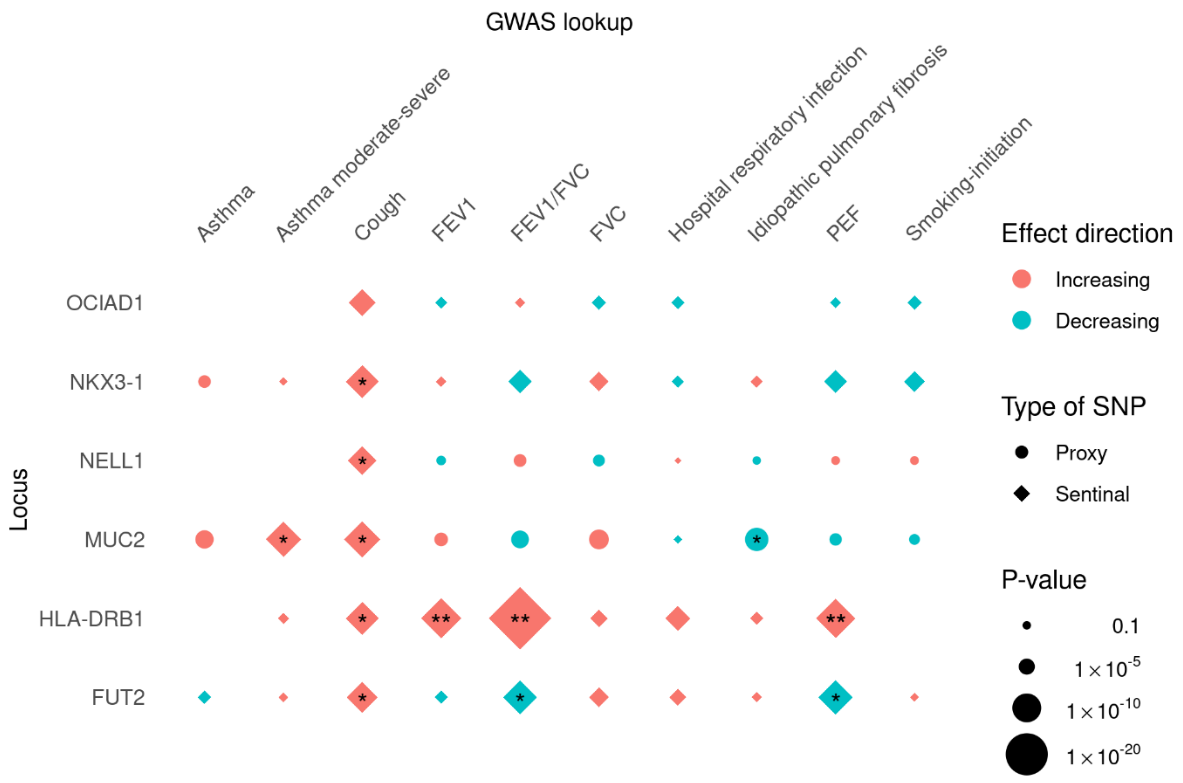
95. Okeley NM, Alley SC, Anderson ME, Boursalian TE, Burke PJ, Emmerton KM, Jeffrey SC, Klussman K, Law C-L, Sussman D, Toki BE, Westendorf L, Zeng W, Zhang X, Benjamin DR, Senter PD. Development of orally active inhibitors of protein and cellular fucosylation. *Proceedings of the National Academy of Sciences* 2013; 110: 5404–5409.


Original Sample
N 121284
N CASES 10481
N CONTROLS 110803
Participants with no genetic data. (N=4168)
Available genetic data
N 117116
N CASES 10134
N CONTROLS 106982
Participants with missing covariates. (N=4419)
Available genetic/covariate data
N 112697
N CASES 9737
N CONTROLS 102960
Participants of non-European ancestry. (N=2730)
European with genetic/covariate data
N 109967
N CASES 9714
N CONTROLS 100253
Excess controls. (N=51782)
Controls matched age + sex 5:1 control-case
N 58185
N CASES 9714
N CONTROLS 48471

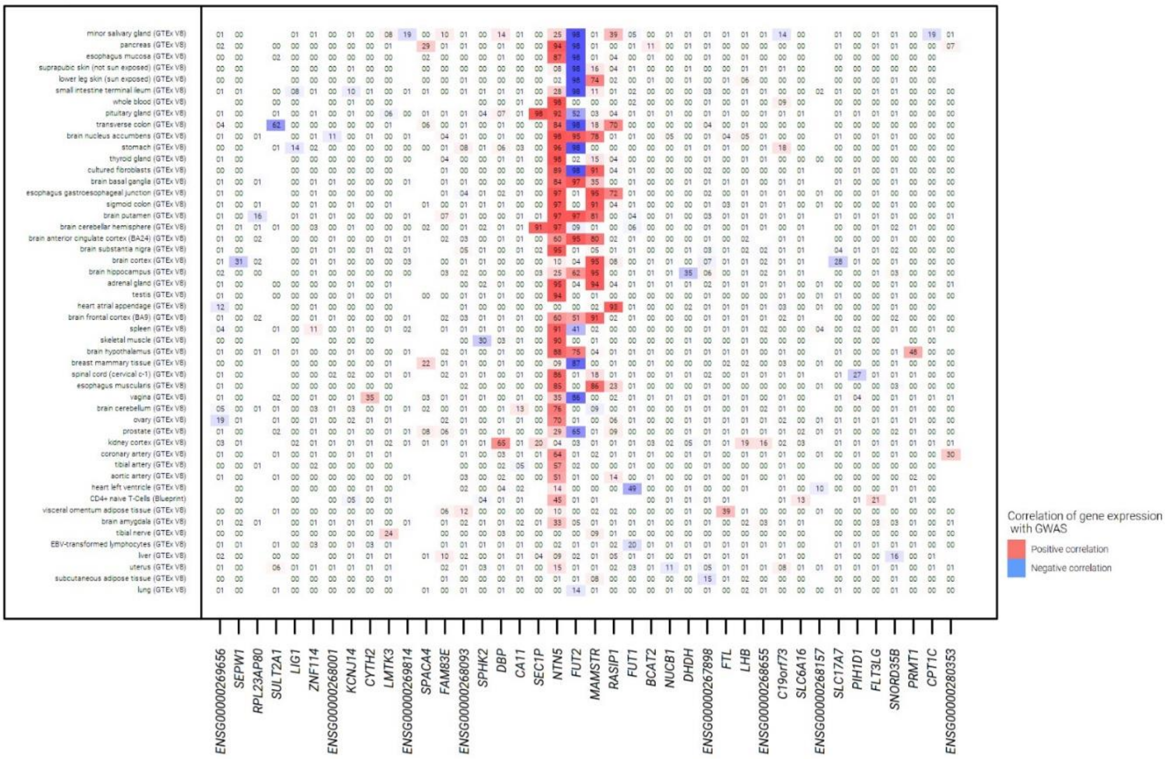
Study flow chart detailing case control selection from the UK Biobank cohort.



LocusZoom plots of the six sentinel signals, a). OCIAD1 signal (rs79998532), b). HLA-DRB5 signal (rs374248993), c). NKX3-1 signal (rs79401075), d). MUC2 signal (rs779167905), e). NELL1 signal (rs529240826), and f). FUT2 signal (rs492602)



Results for association of sentinel variant risk alleles with respiratory traits. Results are aligned to the risk allele for chronic sputum production, effect direction 'Increasing' can be read as increasing risk for binary traits and increasing values in quantitative traits. Chronic bronchitis and smoking age of onset, cigarettes per day and cessation phenotype lookups were omitted as no associations with P<0.05 found. *P <5.95x10-4 (Bonferroni adjustment for 84 association tests) ** P<5x10-8.



Results for eQTL colocalization for the FUT2 locus using variant rs492602. The numbers within the grid are the posterior probability of colocalization (H4), with results aligned to the risk allele G for the rs492602 variant. Missing numbers indicate no data was available for the respective gene and tissue.

Supplementary Materials

Contents

Supplementary Text

Replication: Description of independent replication cohorts (Generation Scotland, LifeLines 1, LifeLines 2, Vlagtwedde—Vlaardingen and EXCEED Study)

The Generation Scotland study

The Generation Scotland study (GS) is a population- and family-based cohort with broad consent for genetic, health, well-being and lifestyle studies. (Smith *et al.*, 2013) The main recruitment (24,096 individuals in 5501 family groups) took place during 2006–11.

In 2020, a series of CovidLife surveys were conducted during the COVID-19 pandemic. Survey invitations were sent to 22,796 members of GS who provided an e-mail address for recontact, as well as to other adults in the UK through collaborators and social media channels (Fawns-Ritchie *et al.*, 2021). The sputum question was asked within the COVID-19 surveys and phrased as "Do you usually bring up phlegm/sputum/mucus from the lungs or do you usually feel like you have mucus in your lungs that is difficult to bring up, with having a cold?" with yes or no as possible answers, yes defined cases and no controls.

The analysis for this paper was performed using PLINK 2, restricted to those of European ancestry and used sex, age, smoking status, and the first 10 principal components as covariates in the regression.

LifeLines 1 and 2 and Vlagtwedde-Vlaardingen

Genotyped individuals from the first (n = 7,976) and second (n = 5,260) data release of the LifeLines cohort study (2006–2011) (Scholtens *et al.*, 2015) and 1,529 subjects from the last survey

(1989/1990) from the Vlagtwedde-Vlaardingen cohort (de Jong *et al.*, 2014; van Diemen *et al.*, 2005), a prospective general population based cohort including Caucasians of Dutch descent were used as replication cohorts (Zeng *et al.*, 2017).

In these cohorts, genotyping was performed using IlluminaCytoSNP-12 arrays. The applied genotyping quality control criteria have been described before (de Jong *et al.*, 2015; Scholtens *et al.*, 2015): Samples with call-rates of less than 95% were excluded as were samples of non-Caucasians and first degree relatives. SNPs were excluded if they had a genotype call-rate < 95%, minor allele frequency (MAF) < 1%, or a Hardy-Weinberg equilibrium (HWE) p-value < 10−4.

Phlegm was measured by standardized questionnaires from the European Community Respiratory Health Survey (ECRHS) (Burney *et al.*, 1994). Phlegm was defined as at least one positive answer to the questions: "do you usually bring up any phlegm from your chest first thing in the morning in winter?" or "do you usually bring up any phlegm from your chest during the day, or at night, in winter?".

The analysis on the presence of phlegm was performed using PLINK version 1.07 (Purcell *et al.*, 2007). We used an additive genetic model adjusted the logistic regression analysis for age, sex, and current smoking.

EXCEED

EXCEED is a longitudinal population-based cohort which facilitates investigation of genetic, environmental and lifestyle-related determinants of a broad range of diseases and of multiple morbidity through data collected at baseline and via electronic healthcare record linkage. Recruitment has taken place in Leicester, Leicestershire and Rutland since 2013 and is ongoing, with 11,000 participants. Recruitment was widened to anyone over 18 years of age living in the Midlands in 2020. Participants provided a DNA sample, have consented to follow-up for up to 25 years through electronic health records and additional bespoke data collection is planned. Data available includes baseline demographics, anthropometry, spirometry, lifestyle factors (smoking and alcohol use) and longitudinal health information from primary care records, with additional linkage to other EHR datasets planned. Patients have consented to be contacted for recall-by-genotype and recall-by-phenotype sub-studies. Further details about the study can be accessed in the Cohort Profile Paper(John *et al.*, 2019), with additional information about our COVID-19 Focus available as a Data Note Paper (Lee *et al.*, 2021).

The sputum question was included in COVID questionnaire 1, the question was "Do you usually bring up phlegm/sputum/mucus from the lungs, or do you feel like you have mucus in your lungs that is difficult to bring up, when you don't have a cold?", the possible potions were "No", "Yes, sometimes, "Yes, always", and "Unsure". Those who responded "Yes, always" were counted as case and "No" controls. The analysis was performed in PLINK 2 using the following as covariates, age, sex, ever/never smoking status and the first 10 principal component.

Replication: Defining sputum phenotype using primary care data in UK Biobank

We evaluated whether primary care codes for sputum production could be used to define an independent case-control dataset within UK Biobank for replication of our discovery GWAS. To do this, we first evaluated the overlap of Read v3 codes for sputum (Supplementary Table 12) in the available V1.0 September 2019 primary care data for the 121,283 participants who had answered “yes” or “no” for UK Biobank field 22504 (“do you bring up phlegm/sputum/mucus daily?”). To be deemed useful to identify new cases from the remaining participants with primary care data, we required a positive predictive value to predict a case based on primary care records to be above 80%.

Presence of one or more sputum codes had a Positive Predictive Value (PPV) of 29% (7.9% of cases and 1.9% of controls had one or more Read codes indicative of sputum production). This is unlikely to change with the release of additional UK Biobank primary care data and this we concluded that we were unable to use primary care data in UK Biobank to define an independent replication dataset.

Associations with other phenotypes: chronic cough and chronic bronchitis

We defined four phenotypes using UK Biobank data for association analysis with the sentinel variants. We defined cough and chronic bronchitis phenotypes selecting cases as those answering yes and controls answering no to the following UK Biobank data-fields, 22502 (cough) 22124 (chronic bronchitis). All phenotypes were limited to European ancestry using the same methods as the primary analysis, association with cough and chronic bronchitis used the same covariates as the primary analysis, all association analyses were run in PLINK 2.0 (Chang et al., 2015).

Associations with other phenotypes: PheWAS analysis

Traits included UK Biobank baseline measures (from questionnaires and physical measures), self-reported medication usage, and operative procedures, as well as those captured in Office of Population Censuses and Surveys codes from the electronic health record. We also included self-reported disease variables and those from hospital episode statistics (ICD-10 codes truncated to three-character codes and combined in block and chapter groups), combining these where possible to maximize power. A total of 2,150 traits were defined with >200 cases and were included for analysis. Analyses were conducted in unrelated European-ancestry individuals (KING kinship coefficient < 0.0442), and were adjusted for age, sex, genotyping array and first ten principal components. Logistic and linear models were fitted for binary and quantitative outcomes, respectively. Biomarker measurements were adjusted for statins according to the ‘Statin identification and LDL adjustment’ methods described by Sinnott-Armstrong 2019 (Sinnott-Armstrong *et al.*, 2021). Statin-adjusted phenotype values were further adjusted (in residualization) for age at assessment center visit, genetic sex, genotyping array, fasting time, sample dilution factor, socio-economic status indicator, blood sample hour, and urine sample hour with assessment center as a random effect. We then conducted rank-based inverse normal transformation of these residuals. The rank-based inverse-normal transformed residuals were used as inputs for our GWAS linear regression in Hail. False discovery rate (FDR) was calculated using the Benjamin H method (Benjamini and Hochberg, 1995) adjusting for the 2,172 traits tested. Associations with a FDR <0.05 are reported.

To test association with the HLA allele we used, DeepPheWAS (Packer *et al.*, 2022). The platform includes 2,504 phenotypes in UK Biobank, a subset of 2,246 are recommended for association testing. Deep PheWAS then filters these requiring a minimum case number, we chose to keep the default settings of 50 case minimum for binary phenotypes and a 100-case minimum for quantitative phenotypes. After limiting to European ancestry, filtering for case numbers and removing related pairs (KING kinship coefficient >=0.0884) this left 1,907 phenotypes for association analysis.

References

Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 289–300.

Burney,P.G. *et al.* (1994) The European Community Respiratory Health Survey. *Eur Respir J*, **7**, 954–960.

Chang,C.C. *et al.* (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaSci*, **4**, 7.

van Diemen,C.C. *et al.* (2005) A disintegrin and metalloprotease 33 polymorphisms and lung function decline in the general population. *Am J Respir Crit Care Med*, **172**, 329–333.

Fawns-Ritchie,C. *et al.* (2021) CovidLife: a resource to understand mental health, well-being and behaviour during the COVID-19 pandemic in the UK. *Wellcome Open Res*, **6**, 176.

John,C. *et al.* (2019) Cohort profile: Extended Cohort for E-health, Environment and DNA (EXCEED). *International Journal of Epidemiology*, dyz175.

de Jong,K. *et al.* (2014) Association of occupational pesticide exposure with accelerated longitudinal decline in lung function. *Am J Epidemiol*, **179**, 1323–1330.

de Jong,K. *et al.* (2015) Genome-wide interaction study of gene-by-occupational exposure and effects on FEV1 levels. *J Allergy Clin Immunol*, **136**, 1664-1672.e14.

Lee,P.H. *et al.* (2021) Extended Cohort for E-health, Environment and DNA (EXCEED) COVID-19 focus. *Wellcome Open Res*, **6**, 349.

Packer,R. *et al.* (2022) Deep-PheWAS: a pipeline for phenotype generation and association analysis for phenome-wide association studies Genetic and Genomic Medicine.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, **81**, 559–575.

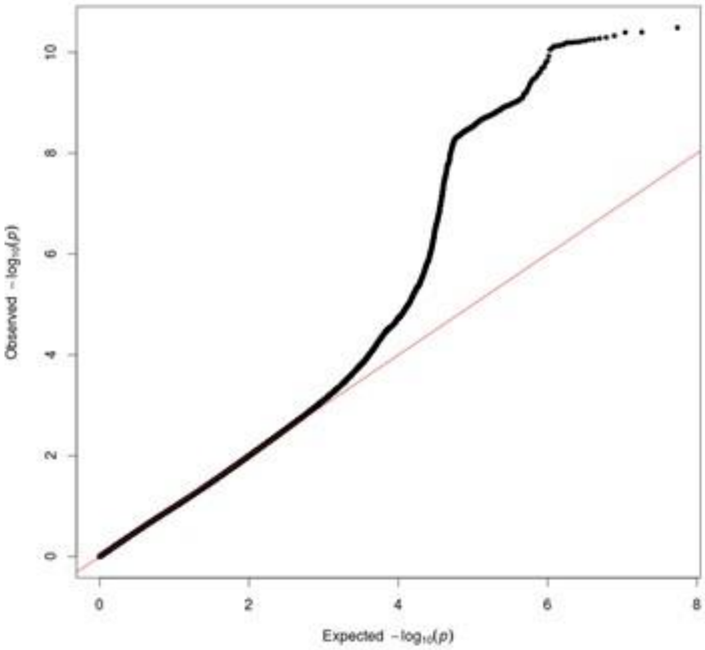
Scholtens,S. *et al.* (2015) Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int J Epidemiol*, **44**, 1172–1180.

Sinnott-Armstrong,N. *et al.* (2021) Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat Genet*, **53**, 185–194.

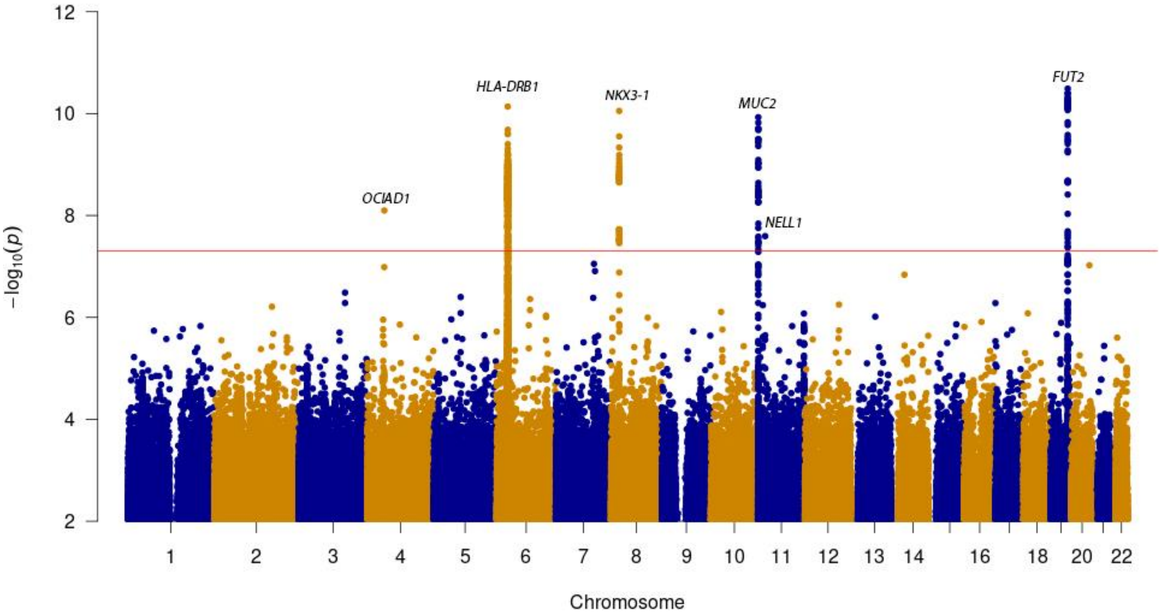
Smith,B.H. *et al.* (2013) Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *International Journal of Epidemiology*, **42**, 689–700.

Zeng,X. *et al.* (2017) No convincing association between genetic markers and respiratory symptoms: results of a GWA study. *Respir Res*, **18**, 11.

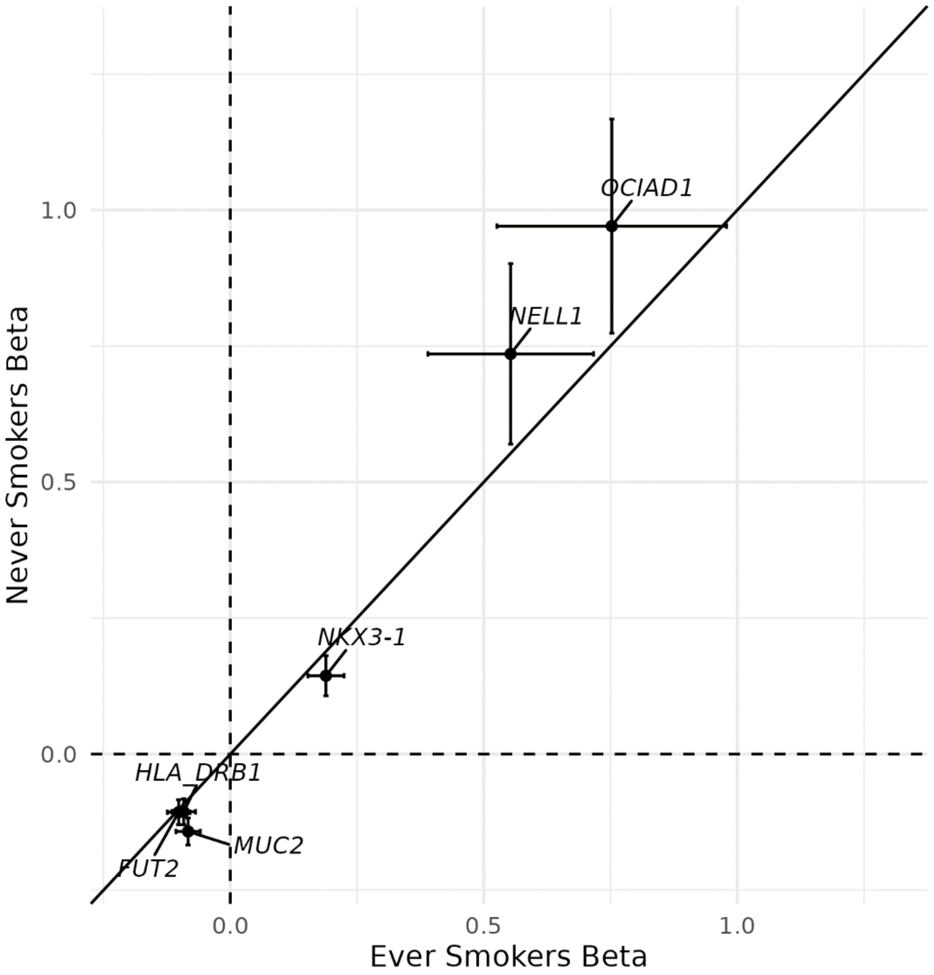
Supplementary Figures



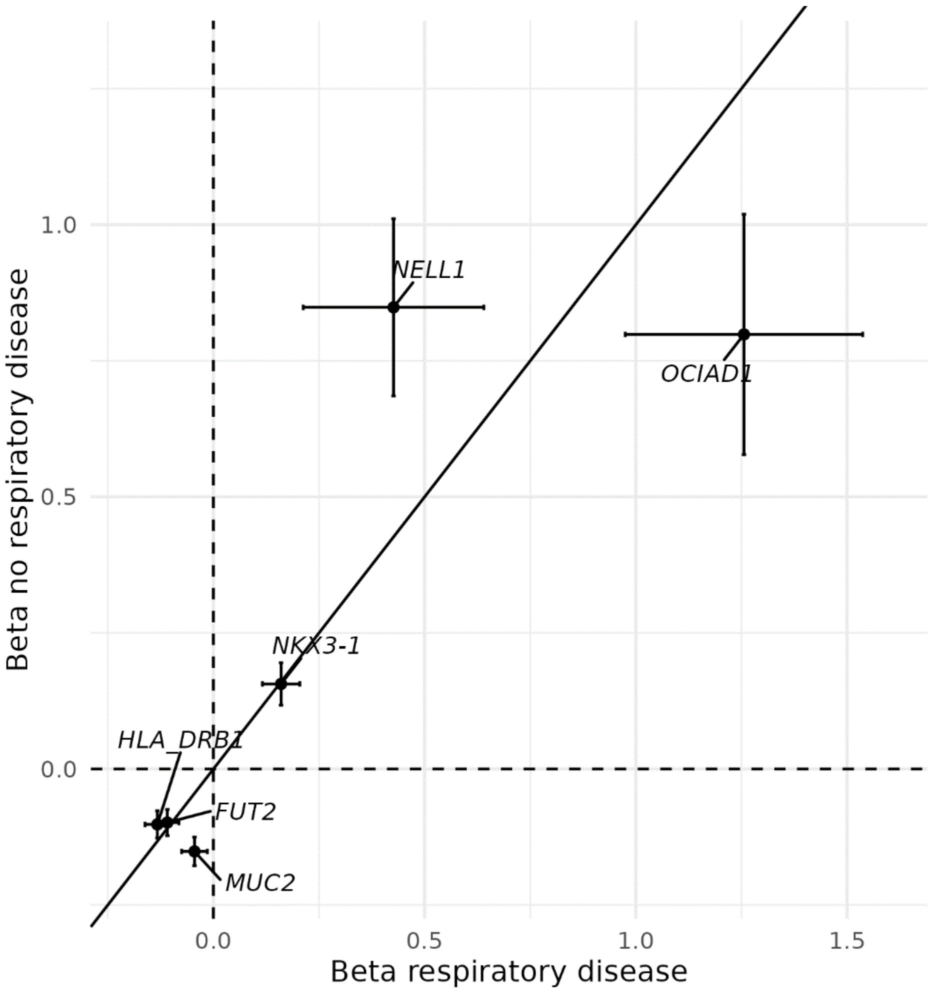
Supplementary Figure 1: Quantile-quantile (QQ) plot for results of genome-wide association study of chronic sputum production in UK Biobank. Variants with imputation quality INFO <0.5 and minor allele count (MAC) <20 were excluded. The genomic control inflation factor, lambda, was 1.026.



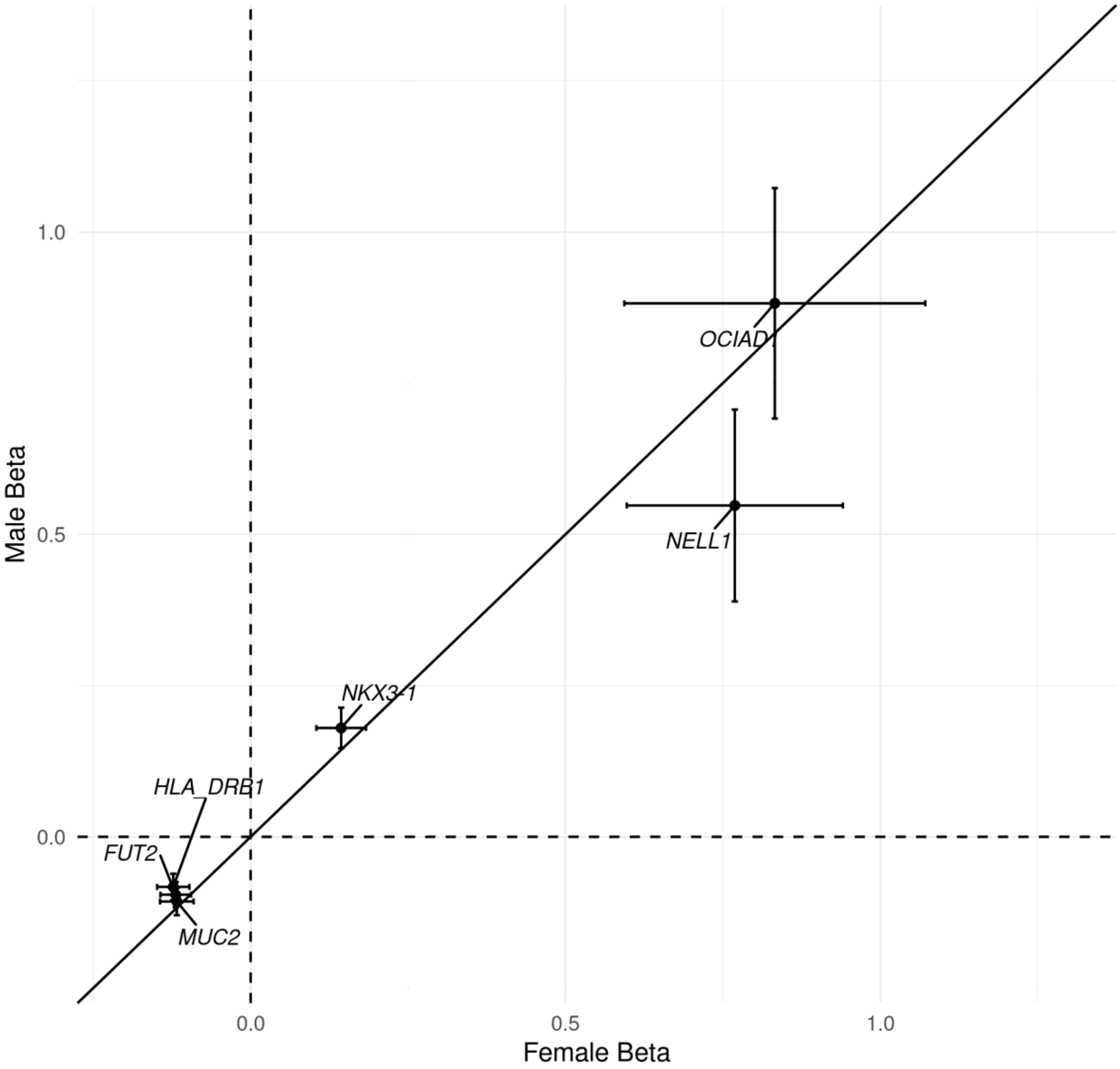
Supplementary Figure 2*: Manhattan plot for the genome-wide association study of chronic sputum production. The red line indicates genome-wide significance.*



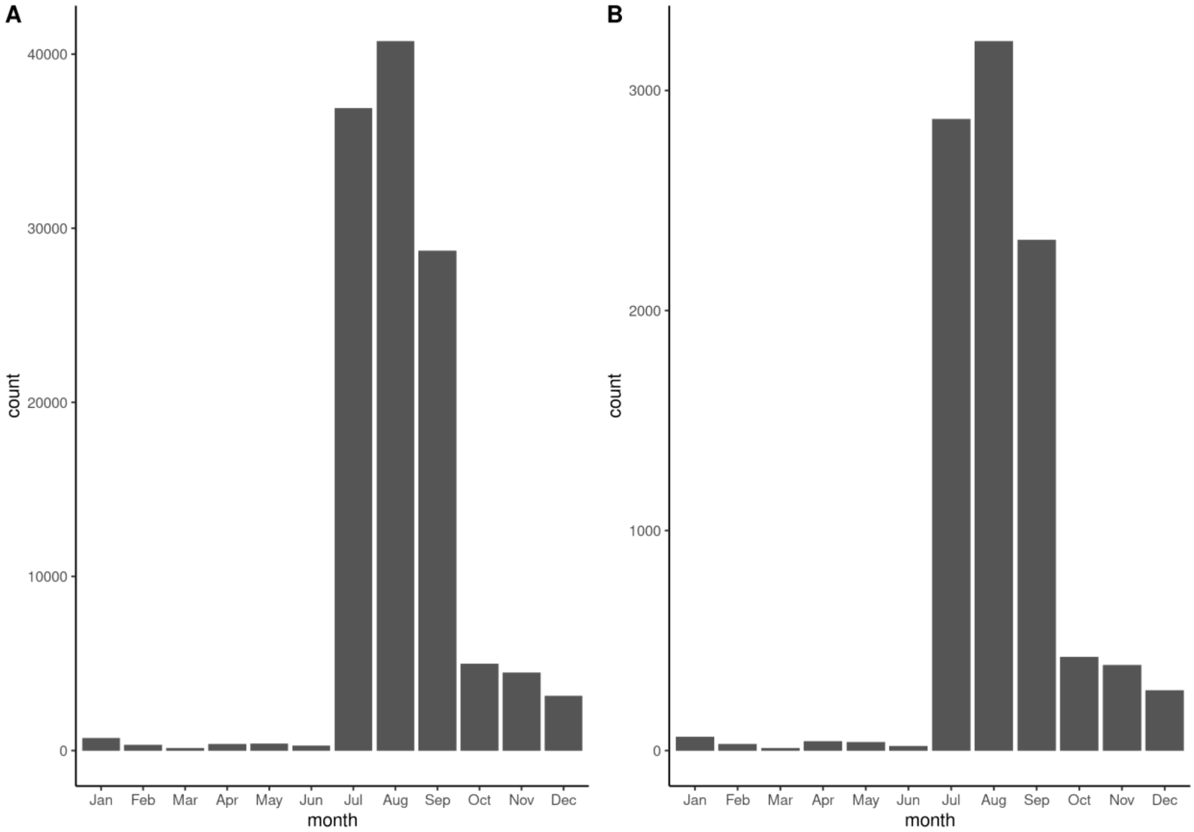
Supplementary Figure 3: Plot of beta values for association for sentinel variants in ever smokers (N cases = 5,161, N controls = 20,229) against beta values for sentinel variants in never smokers (N cases = 4,522, N controls = 28,030), sentinel variants labelled with corresponding loci.



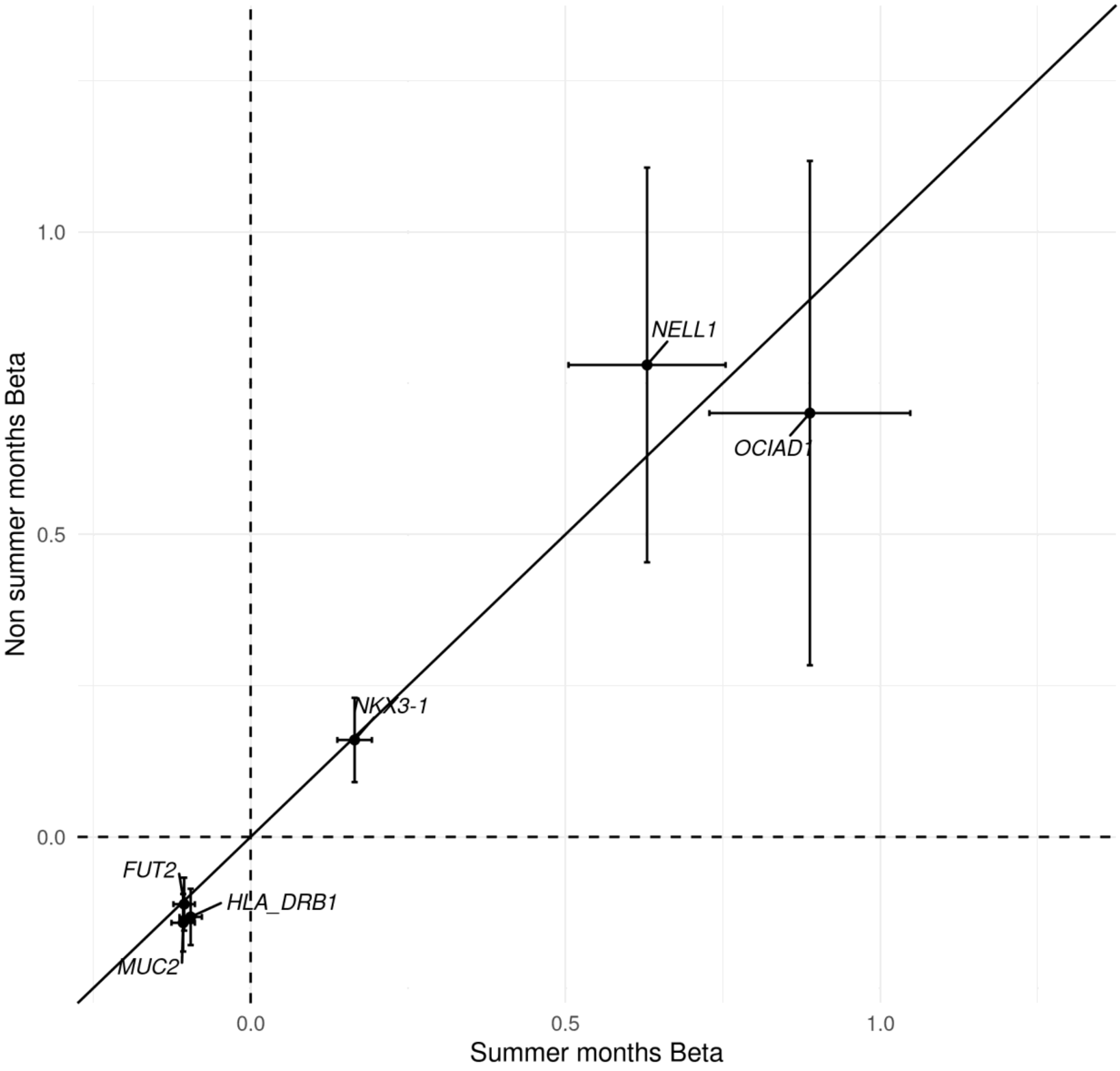
Supplementary Figure 4: Plot of beta values for association for sentinel variants in those with no history of chronic respiratory disease (N cases = 4037, N controls = 28,477), against beta values for sentinel variants in those with a history of chronic respiratory disease (N cases = 3,704, N controls = 9,049), sentinel variants labelled with corresponding loci. History of respiratory disease defined as one or more of, spirometry defined COPD GOLD1+, doctor diagnosed asthma or doctor diagnosed chronic bronchitis.



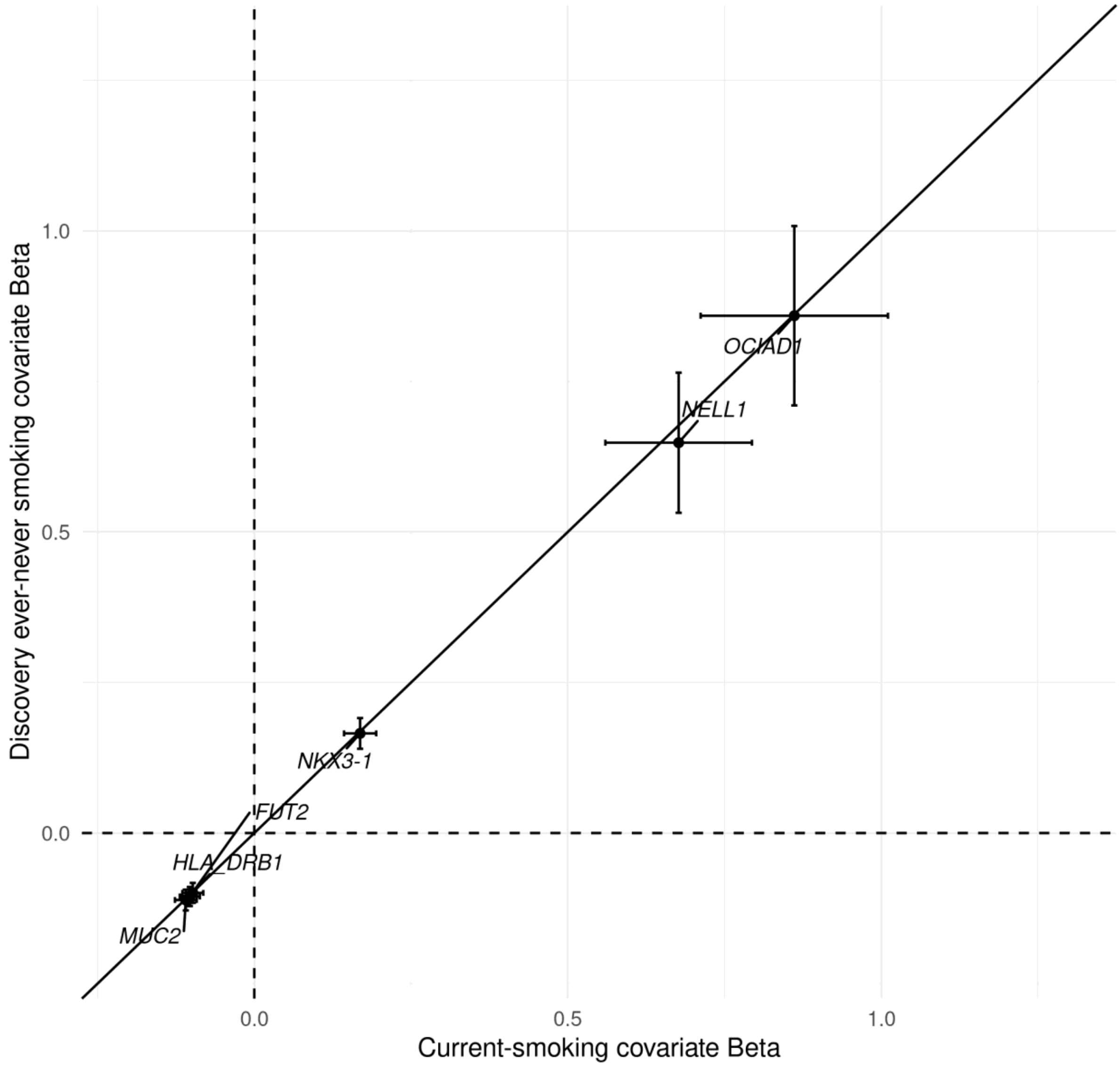
Supplementary Figure 5: Plot of beta values for association for sentinel variants in males (N cases = 5,589, N controls = 27,880), against beta values for sentinel variants in females (N cases = 4,124, N controls = 20,579), sentinel variants labelled with corresponding loci.



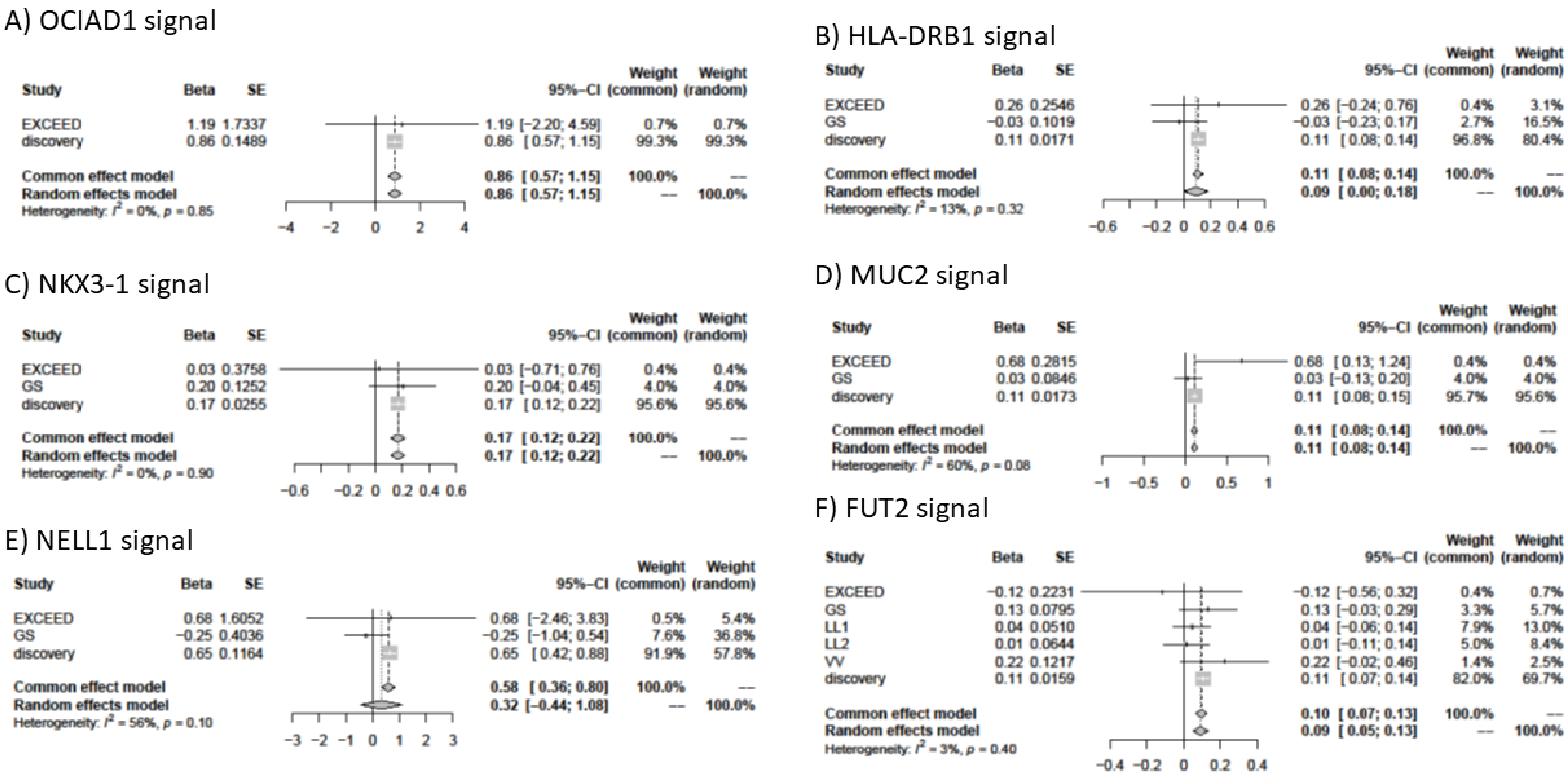
Supplementary Figure 6: Histogram of month that online questionnaire was completed (field-ID 22500) for A) all cases and controls and B) cases only in UK Biobank.



Supplementary Figure 7: Plot of beta values for association for sentinel variants in those who completed the online questionnaire in July, August, and September (N cases = 8,416, N controls = 42,627), against beta values for sentinel variants in those who completed the online questionnaire in the other months (N cases = 1,297, N controls = 5,832) sentinel variants labelled with corresponding loci. July, August, and September selected as these contained the most correspondents in months with higher allergen exposure, see Supplementary Figure 11.



Supplementary Figure 8: Plot of beta values for association for sentinel variants with covariate ‘current smoker’ replacing ‘ever smoker’ in discovery sample (N cases = 9,714, N controls = 48,471), against beta values for sentinel variants in the discovery with original covariates (N cases = 9,714, N controls = 48,471), sentinel variants labelled with corresponding loci.



A) OCIAD1 signal

Study	Beta	SE	95%-CI	Weight (common)	Weight (random)
EXCEED	1.19	1.7337	1.19 [−2.20; 4.59]	0.7%	0.7%
discovery	0.86	0.1489	0.86 [0.57; 1.15]	99.3%	99.3%
Common effect model			**0.86 [0.57; 1.15]**	**100.0%**	**--**
Random effects model			**0.86 [0.57; 1.15]**	**--**	**100.0%**

Heterogeneity: $I^2 = 0\%$, $p = 0.85$

B) HLA-DRB1 signal

Study	Beta	SE	95%-CI	Weight (common)	Weight (random)
EXCEED	0.26	0.2546	0.26 [−0.24; 0.76]	0.4%	3.1%
GS	−0.03	0.1019	−0.03 [−0.23; 0.17]	2.7%	16.5%
discovery	0.11	0.0171	0.11 [0.08; 0.14]	96.8%	80.4%
Common effect model			**0.11 [0.08; 0.14]**	**100.0%**	**--**
Random effects model			**0.09 [0.00; 0.18]**	**--**	**100.0%**

Heterogeneity: $I^2 = 13\%$, $p = 0.32$

C) NKX3-1 signal

Study	Beta	SE	95%-CI	Weight (common)	Weight (random)
EXCEED	0.03	0.3758	0.03 [−0.71; 0.76]	0.4%	0.4%
GS	0.20	0.1252	0.20 [−0.04; 0.45]	4.0%	4.0%
discovery	0.17	0.0255	0.17 [0.12; 0.22]	95.6%	95.6%
Common effect model			**0.17 [0.12; 0.22]**	**100.0%**	**--**
Random effects model			**0.17 [0.12; 0.22]**	**--**	**100.0%**

Heterogeneity: $I^2 = 0\%$, $p = 0.90$

D) MUC2 signal

Study	Beta	SE	95%-CI	Weight (common)	Weight (random)
EXCEED	0.68	0.2815	0.68 [0.13; 1.24]	0.4%	0.4%
GS	0.03	0.0846	0.03 [−0.13; 0.20]	4.0%	4.0%
discovery	0.11	0.0173	0.11 [0.08; 0.15]	95.7%	95.6%
Common effect model			**0.11 [0.08; 0.14]**	**100.0%**	**--**
Random effects model			**0.11 [0.08; 0.14]**	**--**	**100.0%**

Heterogeneity: $I^2 = 60\%$, $p = 0.08$

E) NELL1 signal

Study	Beta	SE	95%-CI	Weight (common)	Weight (random)
EXCEED	0.68	1.6052	0.68 [−2.46; 3.83]	0.5%	5.4%
GS	−0.25	0.4036	−0.25 [−1.04; 0.54]	7.6%	36.8%
discovery	0.65	0.1164	0.65 [0.42; 0.88]	91.9%	57.8%
Common effect model			**0.58 [0.36; 0.80]**	**100.0%**	**--**
Random effects model			**0.32 [−0.44; 1.08]**	**--**	**100.0%**

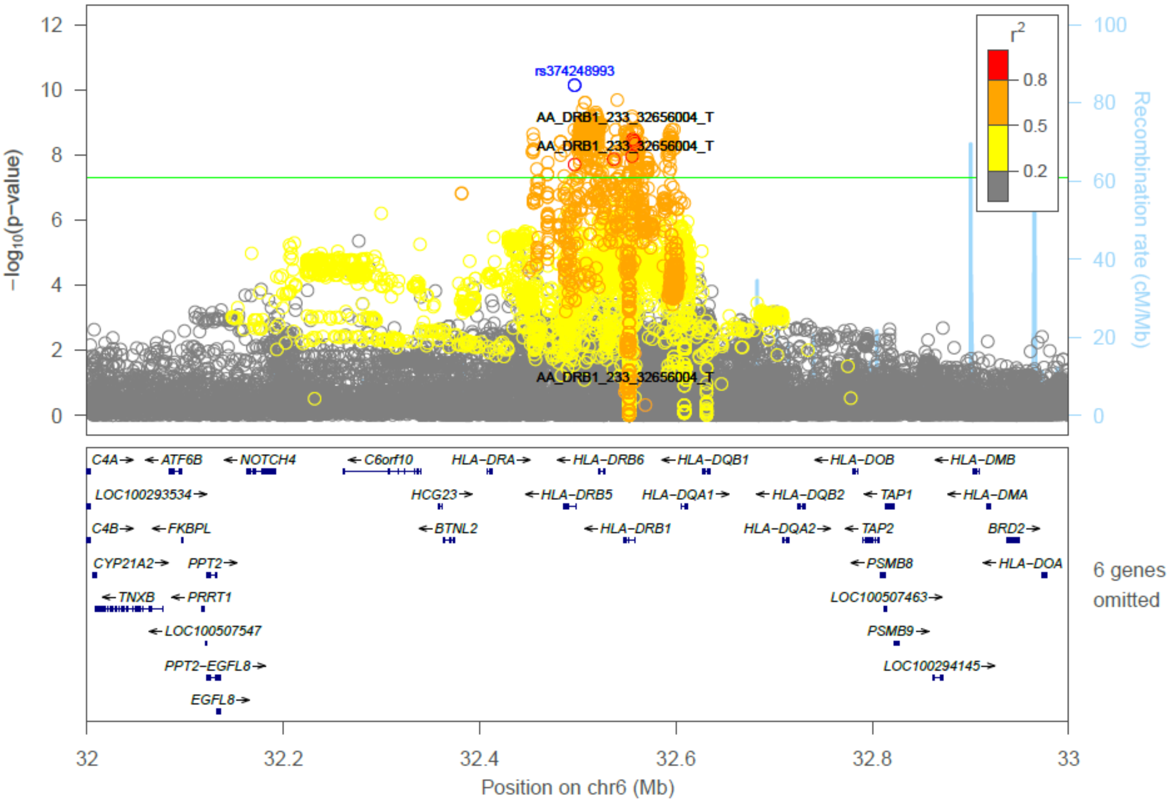
Heterogeneity: $I^2 = 56\%$, $p = 0.10$

F) FUT2 signal

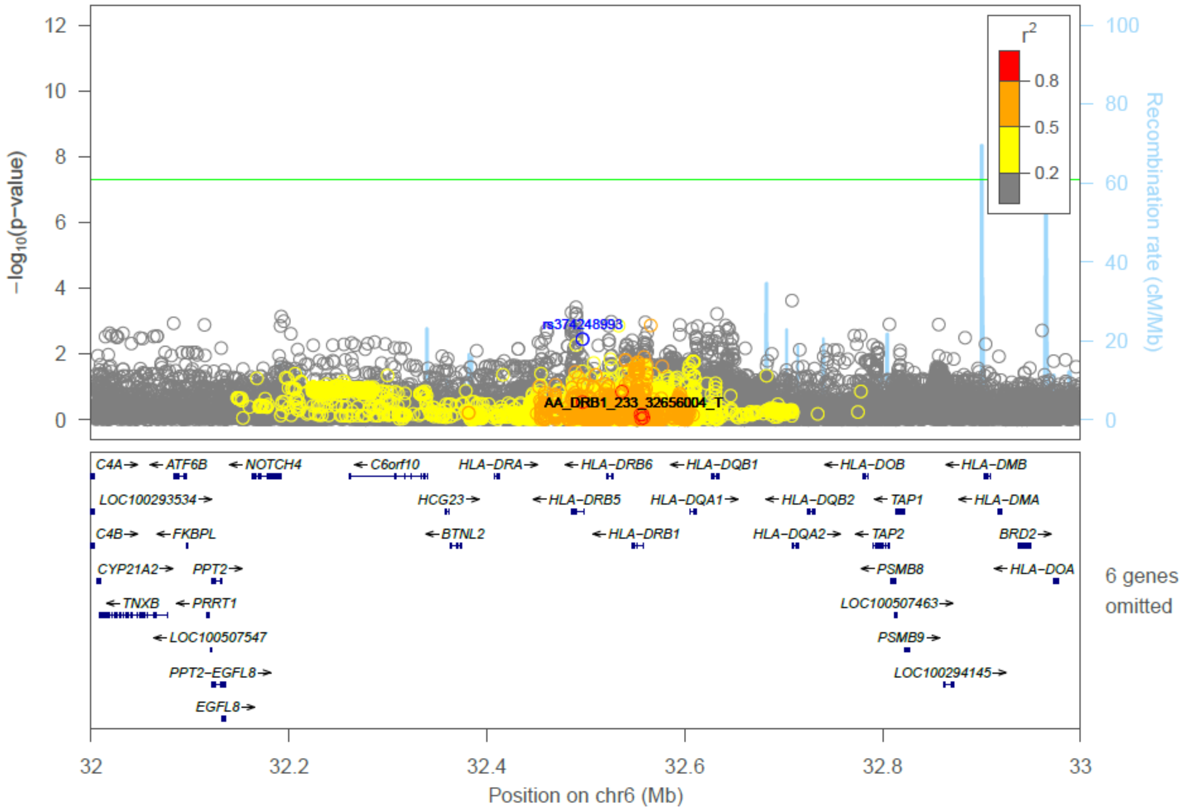
Study	Beta	SE	95%-CI	Weight (common)	Weight (random)
EXCEED	−0.12	0.2231	−0.12 [−0.56; 0.32]	0.4%	0.7%
GS	0.13	0.0795	0.13 [−0.03; 0.29]	3.3%	5.7%
LL1	0.04	0.0510	0.04 [−0.06; 0.14]	7.9%	13.0%
LL2	0.01	0.0644	0.01 [−0.11; 0.14]	5.0%	8.4%
VV	0.22	0.1217	0.22 [−0.02; 0.46]	1.4%	2.5%
discovery	0.11	0.0159	0.11 [0.07; 0.14]	82.0%	69.7%
Common effect model			**0.10 [0.07; 0.13]**	**100.0%**	**--**
Random effects model			**0.09 [0.05; 0.13]**	**--**	**100.0%**

Heterogeneity: $I^2 = 3\%$, $p = 0.40$

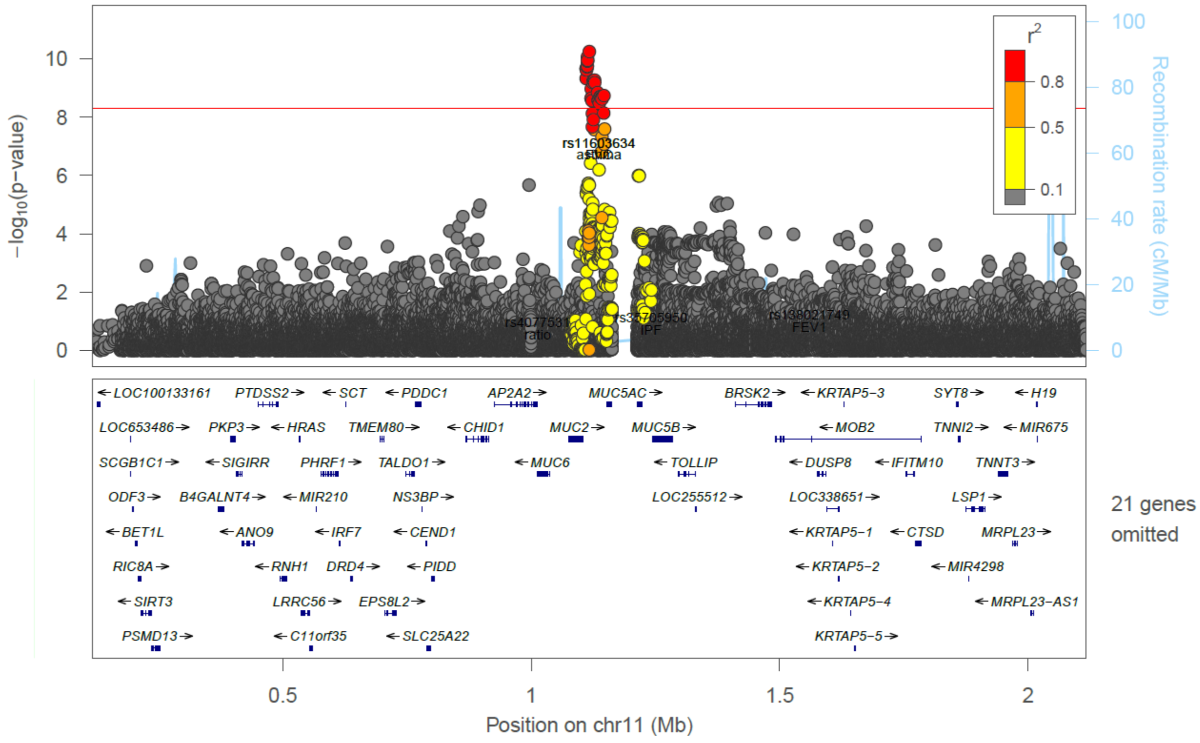
Supplementary Figure 9 Forest plots for results from discovery and replication studies. GS: Generation Scotland, LL1: LifeLines 1, LL2: LifeLines 2, VV: Vlagtwedde-Vlaardingen.



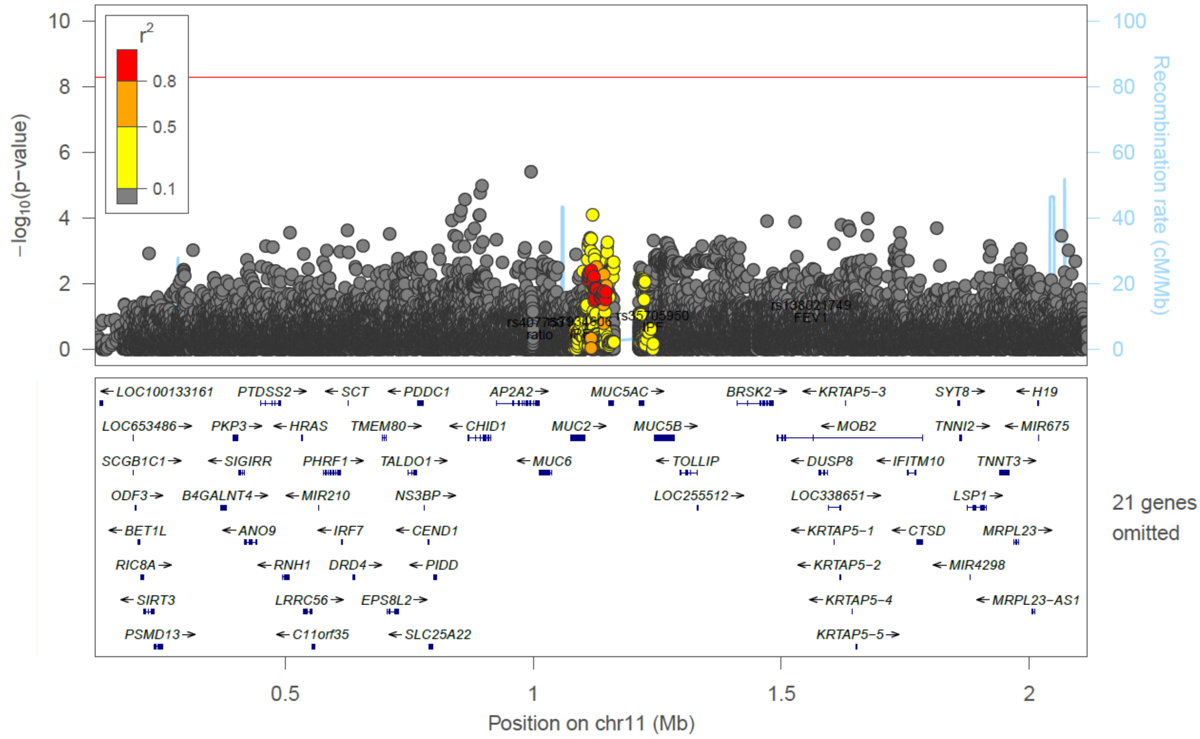
Supplementary Figure 10: Locuszoom plot for combined results SNPs, HLA allele and amino acid changes for *HLA-DRB1* locus, HLA alleles have been labelled in black, sentinel variant is labelled and colored blue.



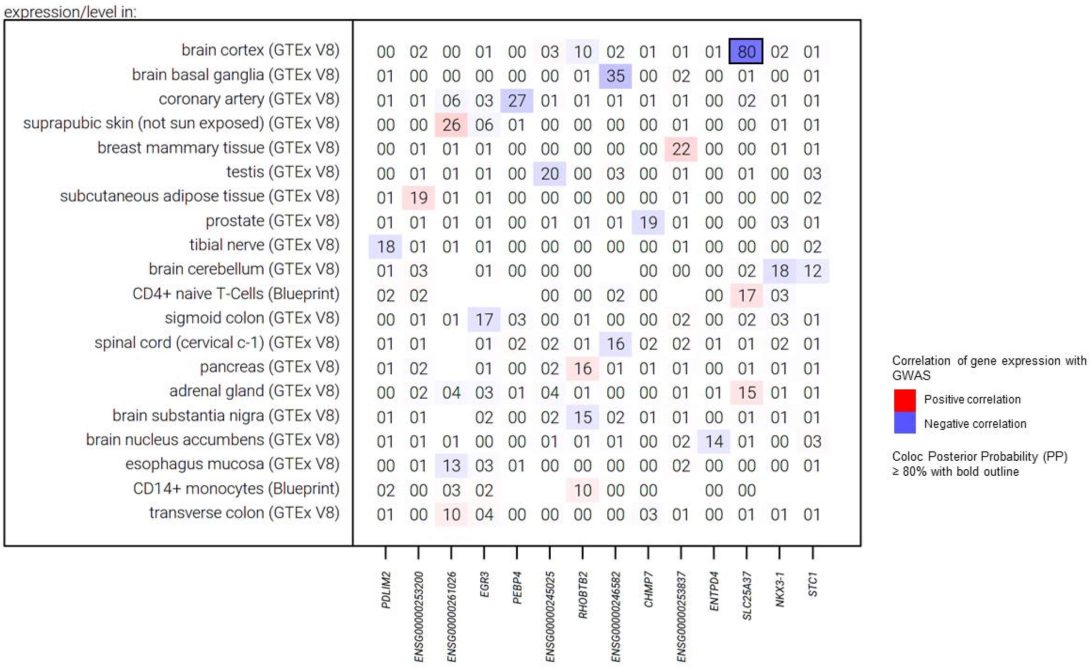
Supplementary Figure 11: Locuszoom plot for combined results SNPs, HLA allele and amino acid changes for *HLA-DRB1* locus conditioned on HLA allele AA_DRB1_233_32656004_T (*DRB1**03:147), HLA alleles have been labelled in black, sentinel variant is labelled and colored blue



Supplementary Figure 12: Region plot of **rs779167905 signal** conditioned on IPF signal **rs35705950** (r^2 = 0.005 with chronic sputum sentinel rs779167905, after conditioning chronic sputum P = 5.67×10^{-11})



Supplementary Figure 13: Region plot of **rs779167905 signal** conditioned on moderate to severe asthma signal **rs11603634** (r^2 = 0.451 with chronic sputum sentinel rs779167905, after conditioning chronic sputum P = 0.0039)



	PDLIM2	*ENSG00000253200*	*ENSG00000261026*	*EGR3*	*PEBP4*	*ENSG00000245025*	*RHOBTB2*	*ENSG00000246582*	*CHMP7*	*ENSG00000253837*	*ENTPD4*	*SLC25A37*	*NKX3-1*	*STC1*
brain cortex (GTEx V8)	00	02	00	01	00	03	10	02	01	01	01	80	02	01
brain basal ganglia (GTEx V8)	01	00	00	00	00	00	01	35	00	02	00	01	00	01
coronary artery (GTEx V8)	01	01	06	03	27	01	01	01	01	01	00	02	01	01
suprapubic skin (not sun exposed) (GTEx V8)	00	00	26	06	01	00	00	00	00	01	00	00	01	01
breast mammary tissue (GTEx V8)	00	01	01	01	00	00	00	00	00	22	00	00	01	01
testis (GTEx V8)	00	01	01	01	00	20	00	03	00	01	00	01	00	03
subcutaneous adipose tissue (GTEx V8)	01	19	01	01	00	00	00	00	00	01	00	00	00	02
prostate (GTEx V8)	01	01	01	01	00	01	01	01	19	01	00	00	03	01
tibial nerve (GTEx V8)	18	01	01	01	00	00	00	00	00	01	00	00	00	02
brain cerebellum (GTEx V8)	01	03		01	00	00	00		00	00	00	02	18	12
CD4+ naive T-Cells (Blueprint)	02	02				00	00	02	00		00	17	03	
sigmoid colon (GTEx V8)	00	01	01	17	03	00	01	00	00	02	00	02	03	01
spinal cord (cervical c-1) (GTEx V8)	01	01		01	02	02	01	16	02	02	01	01	02	01
pancreas (GTEx V8)	01	02		01	00	02	16	01	01	01	00	01	01	01
adrenal gland (GTEx V8)	00	02	04	03	01	04	01	00	00	01	01	15	01	01
brain substantia nigra (GTEx V8)	01	01		02	00	02	15	02	01	01	00	01	01	01
brain nucleus accumbens (GTEx V8)	01	01	01	00	00	01	01	01	00	02	14	01	00	03
esophagus mucosa (GTEx V8)	00	01	13	03	01	00	00	00	00	02	00	00	00	01
CD14+ monocytes (Blueprint)	02	00	03	02			10	00	00		00	00		
transverse colon (GTEx V8)	01	00	10	04	00	00	00	00	03	01	00	01	01	01

Supplementary Figure 14: Results for eQTL colocalization for the *NKX3*-1 locus using variant **rs79401075.** The numbers within the table are the posterior probability of colocalization (H4), with results aligned to the risk allele A for the **rs79401075** variant. Missing numbers indicate no data was available for the respective gene and tissue.

Supplementary Table Legends

Supplementary Table 1: Diagnostic codes used for bronchiectasis. Source of codes V3=Read V3 codes, V2=Read V2 codes, ICD10=International classification of disease 10th edition, ICD9 = International classification of disease 10th edition, MD = Mortality data (ICD10 code), SR = Self report UK Biobank code (data-field 20002). Read V3 and Read V2 are primary care codes extracted from the primary care records. ICD10 codes are extracted from the hospital episodic statistics and mortality data.

Supplementary Table 2: Diagnostic codes used for cystic fibrosis. Source of codes V3=Read V3 codes, V2=Read V2 codes, ICD10=International classification of disease 10th edition, ICD9 = International classification of disease 10th edition, MD = Mortality data (ICD10 code). Read V3 and Read V2 are primary care codes extracted from the primary care records. ICD10 codes are extracted from the hospital episodic statistics and mortality data.

Supplementary Table 3: eQTL tissues and data source searched.

Supplementary Table 4a: Demographics, ever smoking status (UK Biobank data-field 22016), doctor diagnosed asthma (UK Biobank data-field 22127), doctor diagnosed chronic bronchitis (UK Biobank data-field 22129), cough and moderate to severe asthma status of those who answered yes or no to the question “do you bring up phlegm/sputum/mucus daily?” and were viable cases or controls.

Supplementary Table 4b: COPD status based on spirometry of cases and controls (spirometry-derived phenotypes are only available in those with spirometry data that passed quality control).

Supplementary Table 4c: Bronchiectasis and cystic fibrosis status of cases and controls (primary care, secondary care, mortality and self-reported phenotypes). Only individuals with linked primary care data included in the comparison.

Supplementary Table 5: Sensitivity analysis, for each sensitivity analysis the discovery sample was stratified into subgroups and the sentinel SNPs tested for association, in all analyses except for those on smoking phenotypes the same covariates used in the discovery were used, for smoking analyes the 'ever smoking' covariate was removed. Respiratory disease made from self-reported doctor diagnosed asthma (field-ID 22127), self-reported doctor diagnoses chronic bronchitis (field-ID 22129) and those with spirometry indicative of COPD GOLD 1+, no respiratory disease includes only those with availble spirometry who do not meet the criteria for respiratory disease. All smoking derived using field 22506, summer refers to date of completed questionnaire and refers to months July, August, and September, non_summer all other months. #CHROM = Chromosome, ALT = alternative allele, CT = count, A1 is the coded allele.

Supplementary Table 6: Results for GWAS look-ups, results aligned the chronic sputum production risk allele for all loci.

Supplementary Table 7: PheWAS results, all results aligned to risk allele for chronic sputum production. HLA-DRB1 results obtained from Deep-PheWAS. SE = standard error, FRD=false discovery rate, only results with FDR <0.01 are reported.

Supplementary Table 8: Associations with previously reported GWAS in Open Targets Genetics Portal. PMID=Pubmed ID. Only results with $P<5x10^{-8}$ reported. OR=Odds ratio, P=p-value.

Supplementary Table 9: Ensemble variant effect predictor (VEP) results for *FUT2* locus.

Supplementary Table 10: Results for credible sets for each of the 5 non HLA signals, results aligned to chronic sputum production risk allele for all variants.

Supplementary Table 11: eQTL co-localisation results, results aligned to risk allele for chronic sputum production. H4 = posterior probability of the signals being the same.

Supplementary Table 12: Read V3 codes for sputum in original sample and stratified by current smoking status (field-ID 22506). Read V3 are primary care codes extracted from the primary care records. * =Overall N_cases = 4498, N_controls = 45403, **= Overall N_cases = 440, N_controls = 1509, ***= Overall N_cases = 4041, N_controls = 43676.

Supplementary Table 13: Results and meta-analysis from the five additional studies for the six sentinel variants. *= rs10902094 proxy, **= rs143032234 proxy, ª= rs504963 proxy. P=p-value, OR= Odds ratio, CI= Confidence Interval.

TABLE 1. STREGA reporting recommendations, extended from STROBE Statement

Item	Item number	STROBE Guideline	Extension for Genetic Association Studies (STREGA)	Relevant text from manuscript
Title and Abstract	1	(a) Indicate the study's design with a commonly used term in the title or the abstract.		Title and Abstract: "genome-wide association study"
		(b) Provide in the abstract an informative and balanced summary of what was done and what was found.		Abstract: Methods (including sample size, significance threshold used) and Findings
Introduction				
Background rationale	2	Explain the scientific background and rationale for the investigation being reported.		Abstract: Background Introduction
Objectives	3	State specific objectives, including any pre-specified hypotheses.	***State if the study is the first report of a genetic association, a replication effort, or both.***	Introduction: Paragraph 4 – first report of genetic associations for chronic sputum production.
Methods				
Study design	4	Present key elements of study design early in the paper.		Abstract: Methods and Findings. Hypothesis and overall approach outlined at end of Introduction.

Item	Item number	STROBE Guideline	Extension for Genetic Association Studies (STREGA)	Relevant text from manuscript
				Methods
Setting	5	Describe the setting, locations and relevant dates, including periods of recruitment, exposure, follow-up, and data collection.		Methods: Subjects Data Availability
Participants	6	(a) **Cohort study –** Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up. **Case-control study –** Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls. **Cross-sectional study –** Give the eligibility criteria, and the sources and methods of selection of participants.	***Give information on the criteria and methods for selection of subsets of participants from a larger study, when relevant*.**	Case-control study Methods: Study population
		(b) **Cohort study –** For matched studies, give matching criteria and number of exposed and unexposed. **Case-control study –** For matched studies, give matching criteria and the number of controls per case.		Case-control study Methods: Study population

Item	Item number	STROBE Guideline	Extension for Genetic Association Studies (STREGA)	Relevant text from manuscript
Variables	7	*(a)* Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.	***(b) Clearly define genetic exposures (genetic variants) using a widely-used nomenclature system. Identify variables likely to be associated with population stratification (confounding by ethnic origin).***	Methods: Genome-wide association study of chronic sputum production Rsids used to describe associated variants.
Data sources measurement	8*****	*(a)* For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group.	***(b) Describe laboratory methods, including source and storage of DNA, genotyping methods and platforms (including the allele calling algorithm used, and its version), error rates and call rates. State the laboratory/centre where genotyping was done. Describe comparability of laboratory methods if there is more than one group. Specify whether genotypes were assigned using all of the data from the study simultaneously or in smaller batches.***	Methods: Genome-wide association study of chronic sputum production
Bias	9	*(a)* Describe any efforts to address potential sources of bias.	***(b) For quantitative outcome variables, specify if any investigation of potential bias resulting from pharmacotherapy was undertaken. If relevant, describe the nature and magnitude of the potential bias,***	Sensitivity analyses undertaken to investigate contribution of pre-existing respiratory disease or smoking status, time of year for completion of

Item	Item number	STROBE Guideline	Extension for Genetic Association Studies (STREGA)	Relevant text from manuscript
			and explain what approach was used to deal with this.	survey, sex, current smoking vs ever smoking.
Study size	10	Explain how the study size was arrived at.		Methods: Study population
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why.	***If applicable, describe how effects of treatment were dealt with.***	Methods: Genome-wide association study of chronic sputum production
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding.	***State software version used and options (or settings) chosen.***	Methods: Genome-wide association study of chronic sputum production
		(b) Describe any methods used to examine subgroups and interactions.		Methods: Associations with other phenotypes
		(c) Explain how missing data were addressed.		Methods: Study population
		(d) **Cohort study –** If applicable, explain how loss to follow-up was addressed.		Methods: Study population

Item	Item number	STROBE Guideline	Extension for Genetic Association Studies (STREGA)	Relevant text from manuscript
		Case-control study – If applicable, explain how matching of cases and controls was addressed. **Cross-sectional study –** If applicable, describe analytical methods taking account of sampling strategy.		
		(e) Describe any sensitivity analyses.		Methods: Associations with other phenotypes
			***(f) State whether Hardy-Weinberg equilibrium was considered and, if so, how*.**	Methods: Genome-wide association study of chronic sputum production (Paragraph 1)
			(g) Describe any methods used for inferring genotypes or haplotypes.	Methods: Genome-wide association study of chronic sputum production (Paragraph 1)
			(h) Describe any methods used to assess or address population stratification.	Methods: Genome-wide association study of chronic sputum production
			(i) Describe any methods used to address multiple comparisons or to control risk of false positive findings.	Methods: Genome-wide association study of chronic sputum production

Item	Item number	STROBE Guideline	Extension for Genetic Association Studies (STREGA)	Relevant text from manuscript
			(j) Describe any methods used to address and correct for relatedness among subjects	Methods: Study population
Results				
Participants	13*	(a) Report the numbers of individuals at each stage of the study – e.g., numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed.	***Report numbers of individuals in whom genotyping was attempted and numbers of individuals in whom genotyping was successful.***	Results: Paragraph 1, Table 1, figure 1
		(b) Give reasons for non-participation at each stage.		Results: Paragraph 1, figure 1
		(c) Consider use of a flow diagram.		Described in: Methods: Study population, figure 1
Descriptive data	14*	(a) Give characteristics of study participants (e.g., demographic, clinical, social) and information on exposures and potential confounders.	***Consider giving information by genotype***.	Results: Table 1

Item	Item number	STROBE Guideline	Extension for Genetic Association Studies (STREGA)	Relevant text from manuscript
		(b) Indicate the number of participants with missing data for each variable of interest.		Results: Table 1
		(c) **Cohort study –** Summarize follow-up time, e.g. average and total amount.		Not Applicable
Outcome data	15 *****	**Cohort study-**Report numbers of outcome events or summary measures over time.	***Report outcomes (phenotypes) for each genotype category over time***	
		Case-control study – Report numbers in each exposure category, or summary measures of exposure.	***Report numbers in each genotype category***	Not applicable for genome-wide study. Coded allele frequencies for variants of interest reported throughout.
		Cross-sectional study – Report numbers of outcome events or summary measures.	***Report outcomes (phenotypes) for each genotype category***	
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95% confidence intervals). Make clear which confounders		Results: Table 2

Item	Item number	STROBE Guideline	Extension for Genetic Association Studies (STREGA)	Relevant text from manuscript
		were adjusted for and why they were included.		
		(b) Report category boundaries when continuous variables were categorized.		Not applicable
		(c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period.		Not applicable
			(d) Report results of any adjustments for multiple comparisons.	Results: Table 2, Figure 4.
Other analyses	17	(a) Report other analyses done – e.g., analyses of subgroups and interactions, and sensitivity analyses.		Results: page 7, paragraph 1, Supplementary Figures 7 and 8, Supplementary Table 5.
			(b) If numerous genetic exposures (genetic variants) were examined, summarize results from all analyses undertaken.	Results: Figure 2

Item	Item number	STROBE Guideline	Extension for Genetic Association Studies (STREGA)	Relevant text from manuscript
			(c) If detailed results are available elsewhere, state how they can be accessed.	Data Availability
Discussion				
Key results	18	Summarize key results with reference to study objectives.		Discussion: Paragraph 1
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias.		Discussion: Paragraph 7
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence.		Discussion: Paragraph 8
Generalizability	21	Discuss the generalizability (external validity) of the study results.		Discussion: Paragraph 6
Other Information				
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based.		Acknowledgements

STREGA = STrengthening the REporting of Genetic Association studies; STROBE = STrengthening the Reporting of Observational Studies in Epidemiology.

* Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.