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.

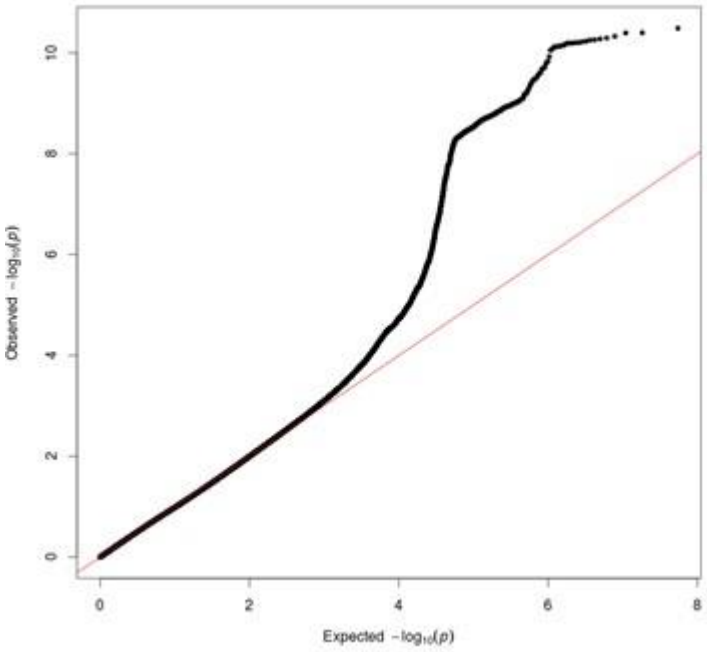
Scholtens,S. *et al.* (2015) Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int J Epidemiol*, **44**, 1172–1180.

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

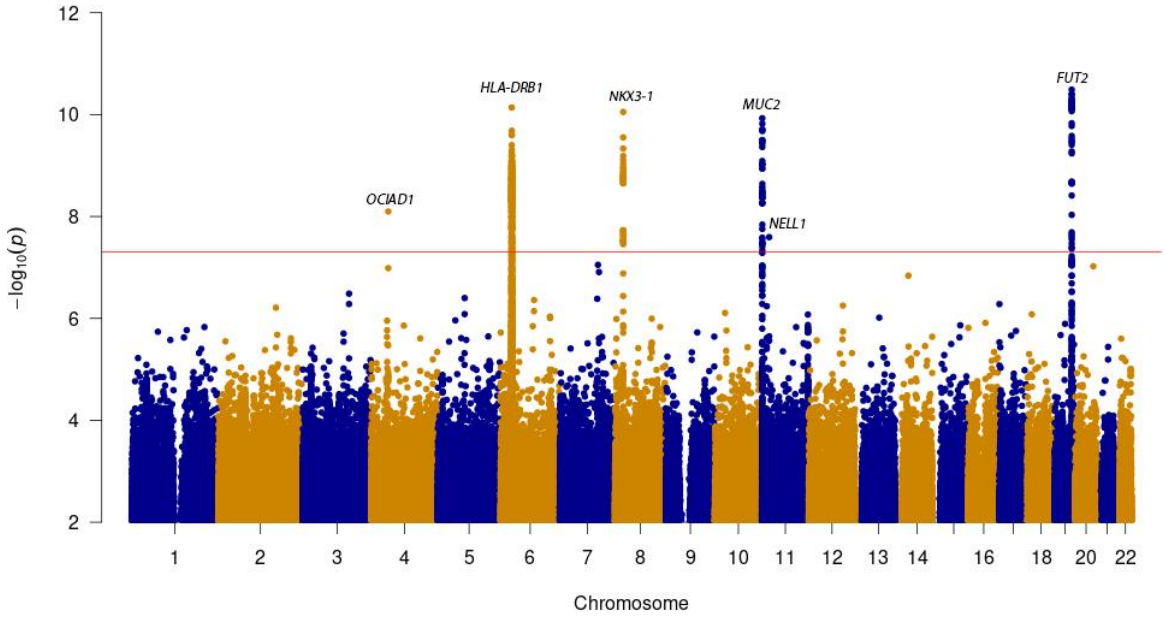
Smith,B.H. *et al.* (2013) Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *International Journal of Epidemiology*, **42**, 689–700.

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

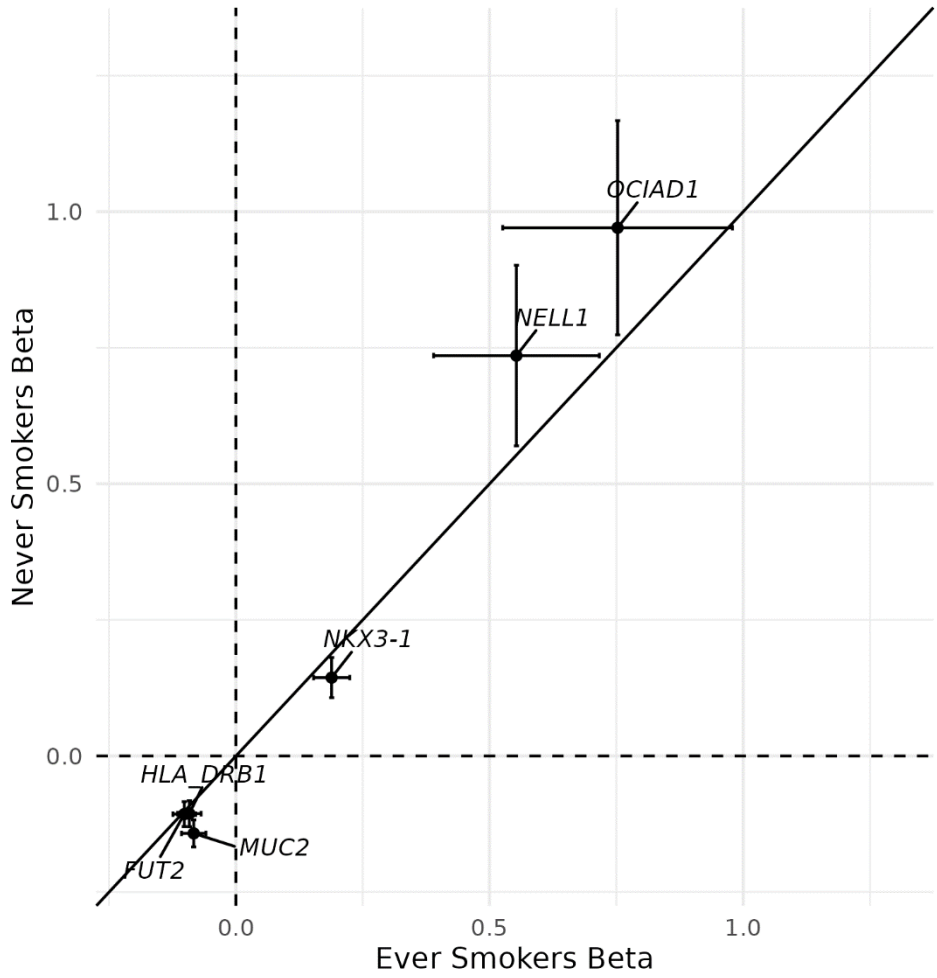
Supplementary Figures



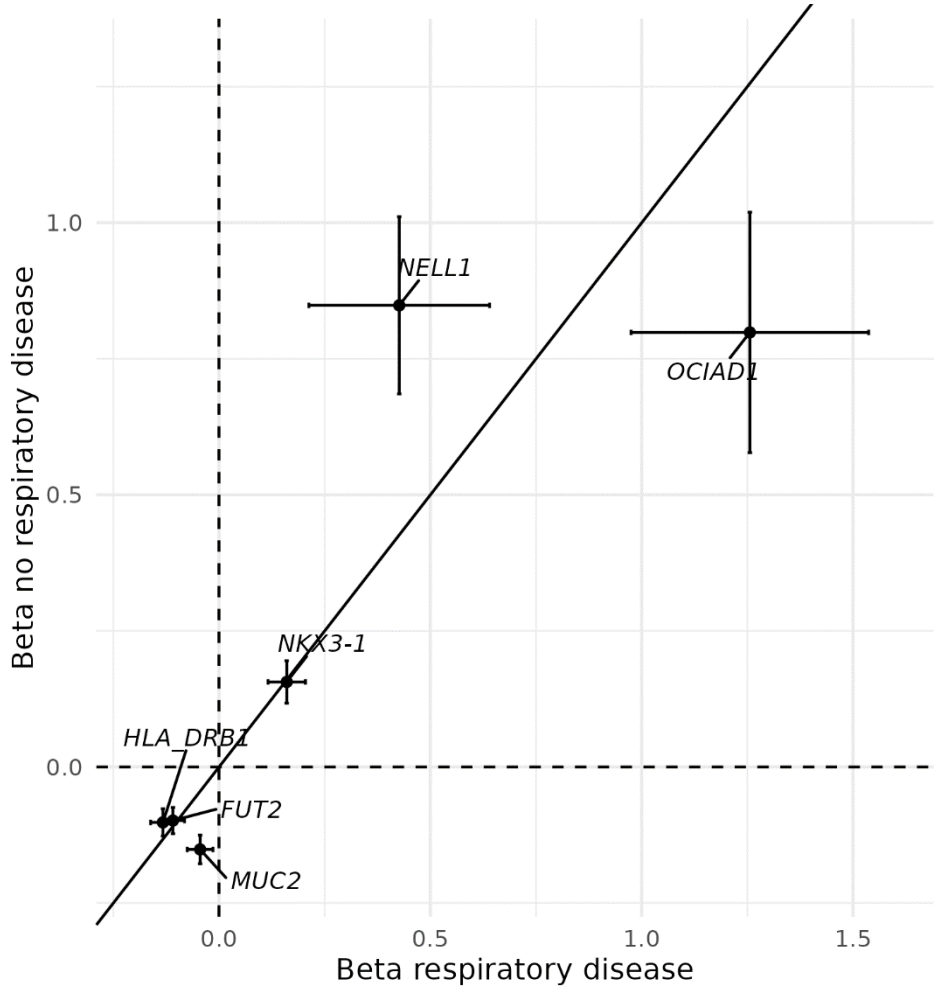
Supplementary Figure S1: Quantile-quantile (QQ) plot for results of genome-wide association study of chronic sputum production in UK Biobank. Variants with imputation quality INFO <0.5 and minor allele count (MAC) <20 were excluded. The genomic control inflation factor, lambda, was 1.026.



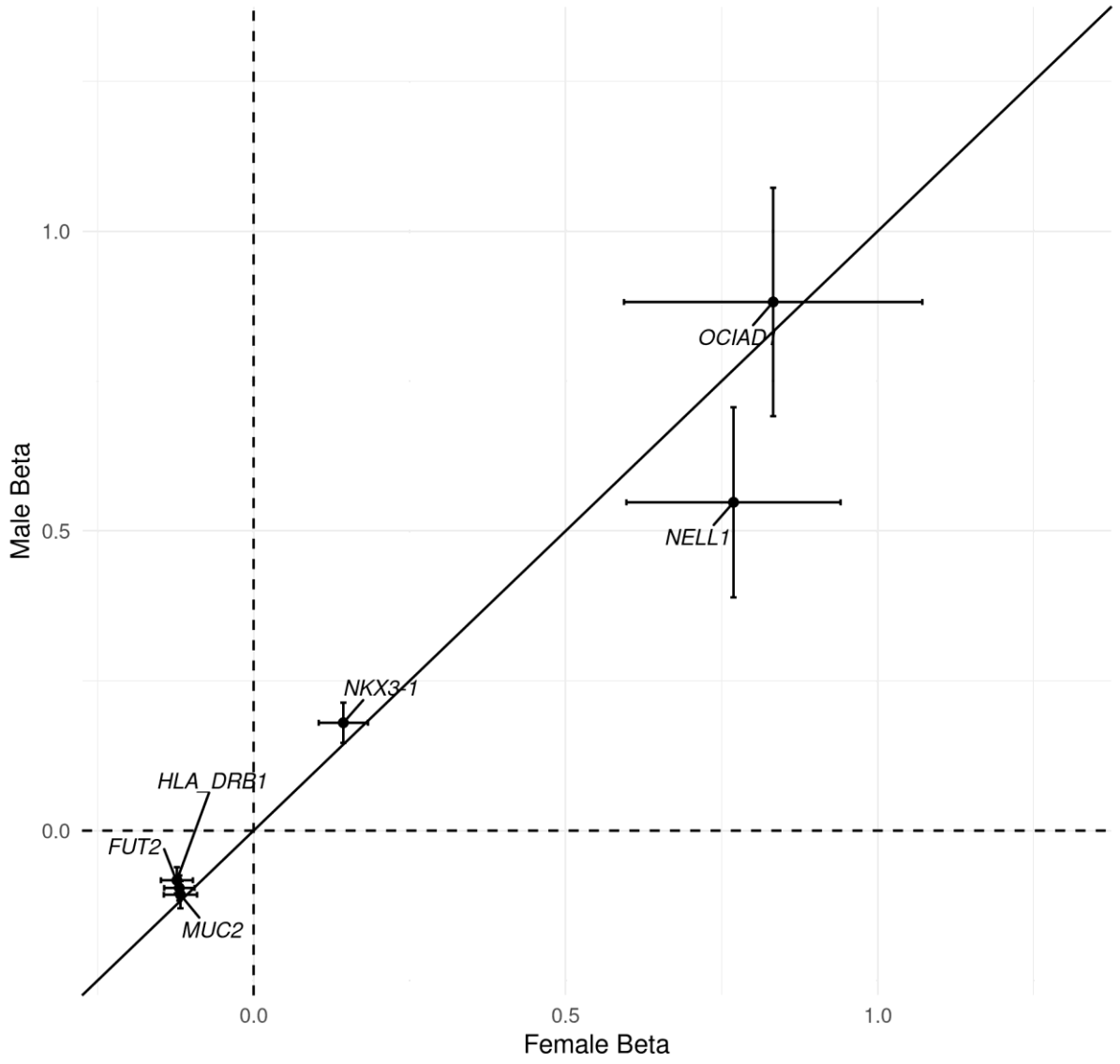
***Supplementary Figure S2**: Manhattan plot for the genome-wide association study of chronic sputum production. The red line indicates genome-wide significance.*



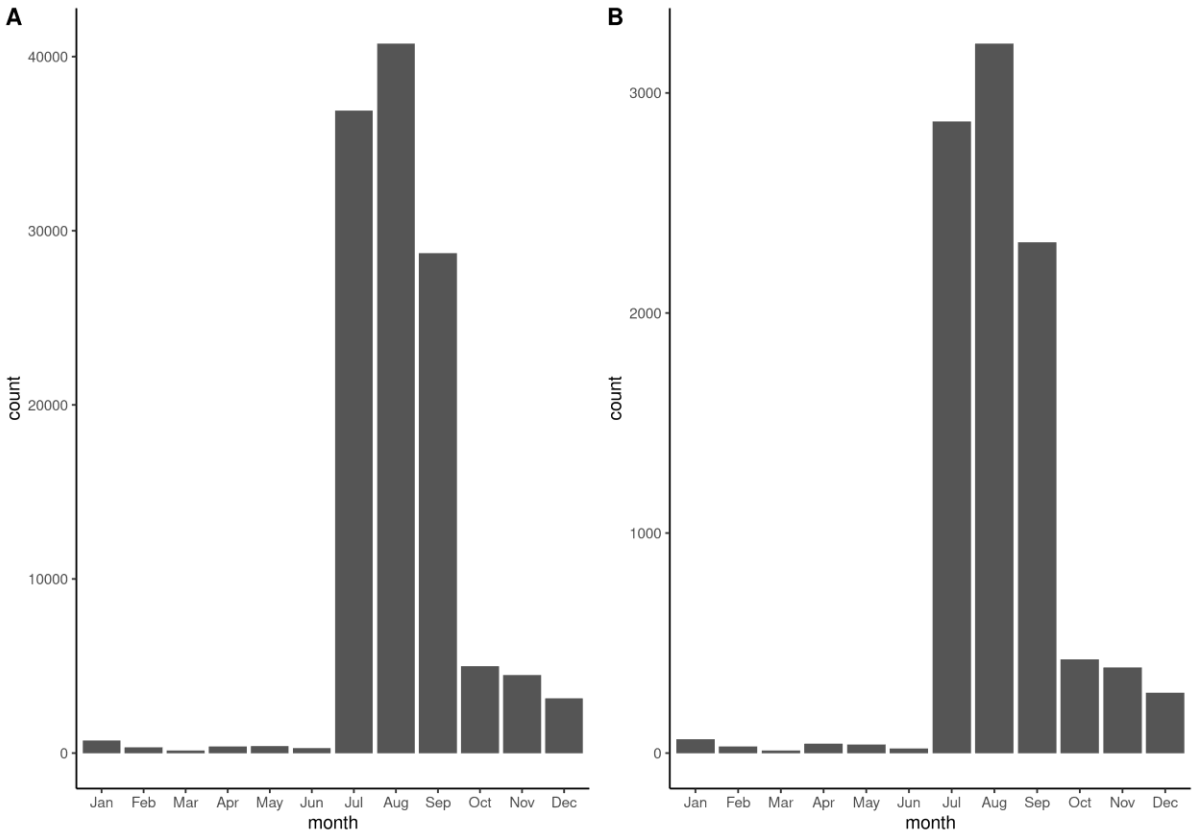
Supplementary Figure S3: Plot of beta values for association for sentinel variants in ever smokers (N cases = 5,161, N controls = 20,229) against beta values for sentinel variants in never smokers (N cases = 4,522, N controls = 28,030), sentinel variants labelled with corresponding loci.



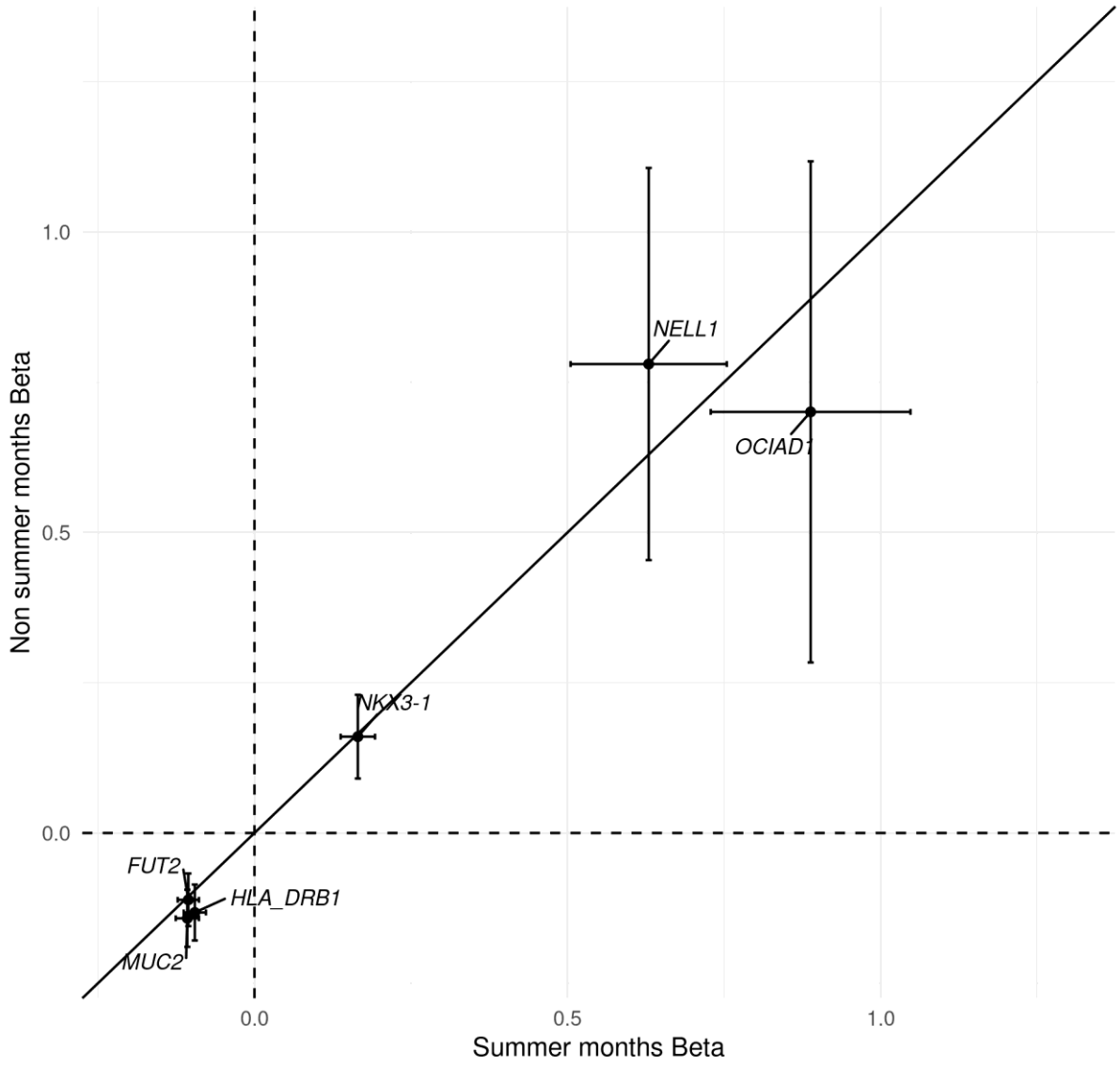
Supplementary Figure S4: Plot of beta values for association for sentinel variants in those with no history of chronic respiratory disease (N cases = 4037, N controls = 28,477), against beta values for sentinel variants in those with a history of chronic respiratory disease (N cases = 3,704, N controls = 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.



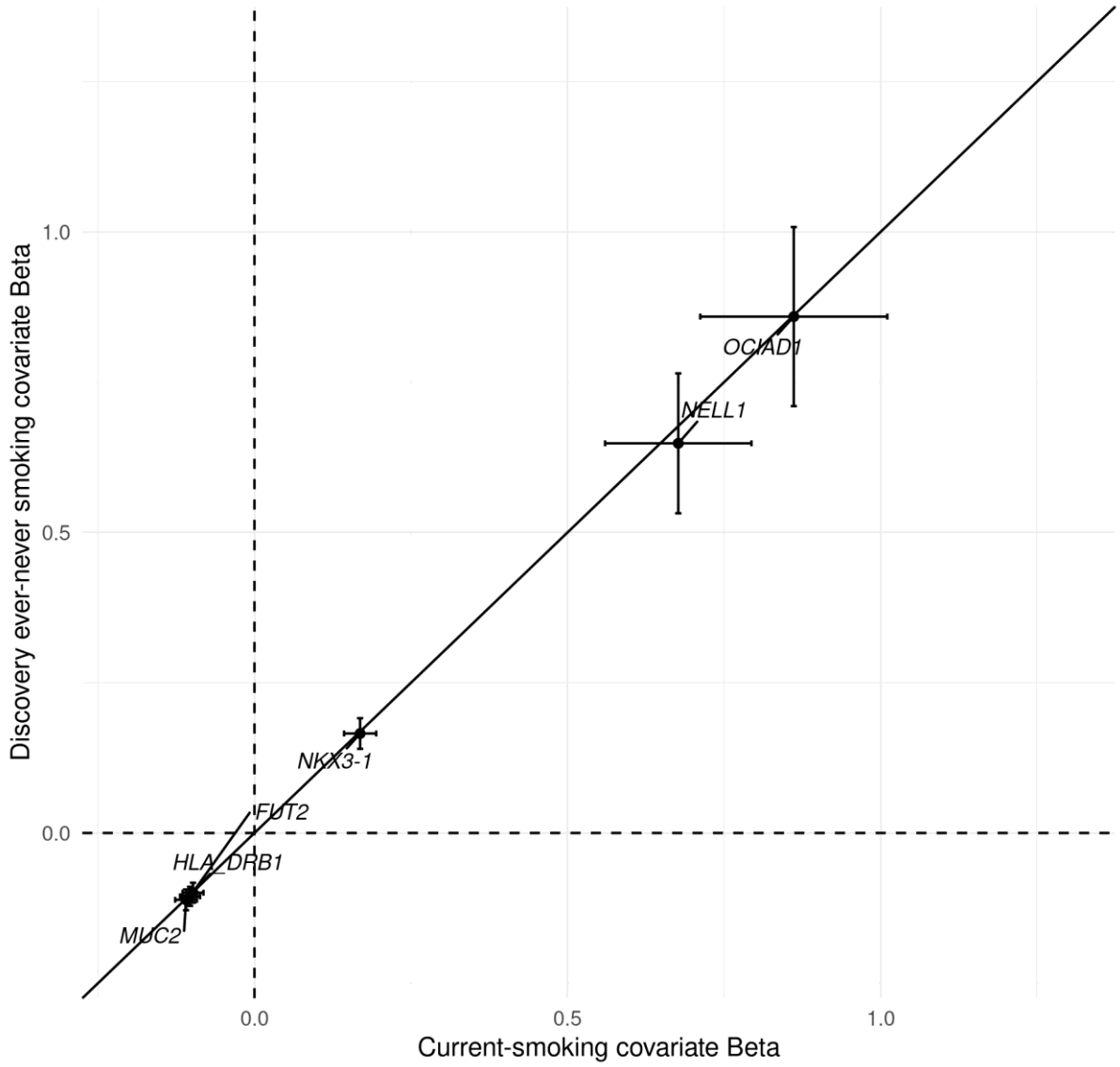
Supplementary Figure S5: Plot of beta values for association for sentinel variants in males (N cases = 5,589, N controls = 27,880), against beta values for sentinel variants in females (N cases = 4,124, N controls = 20,579), sentinel variants labelled with corresponding loci.



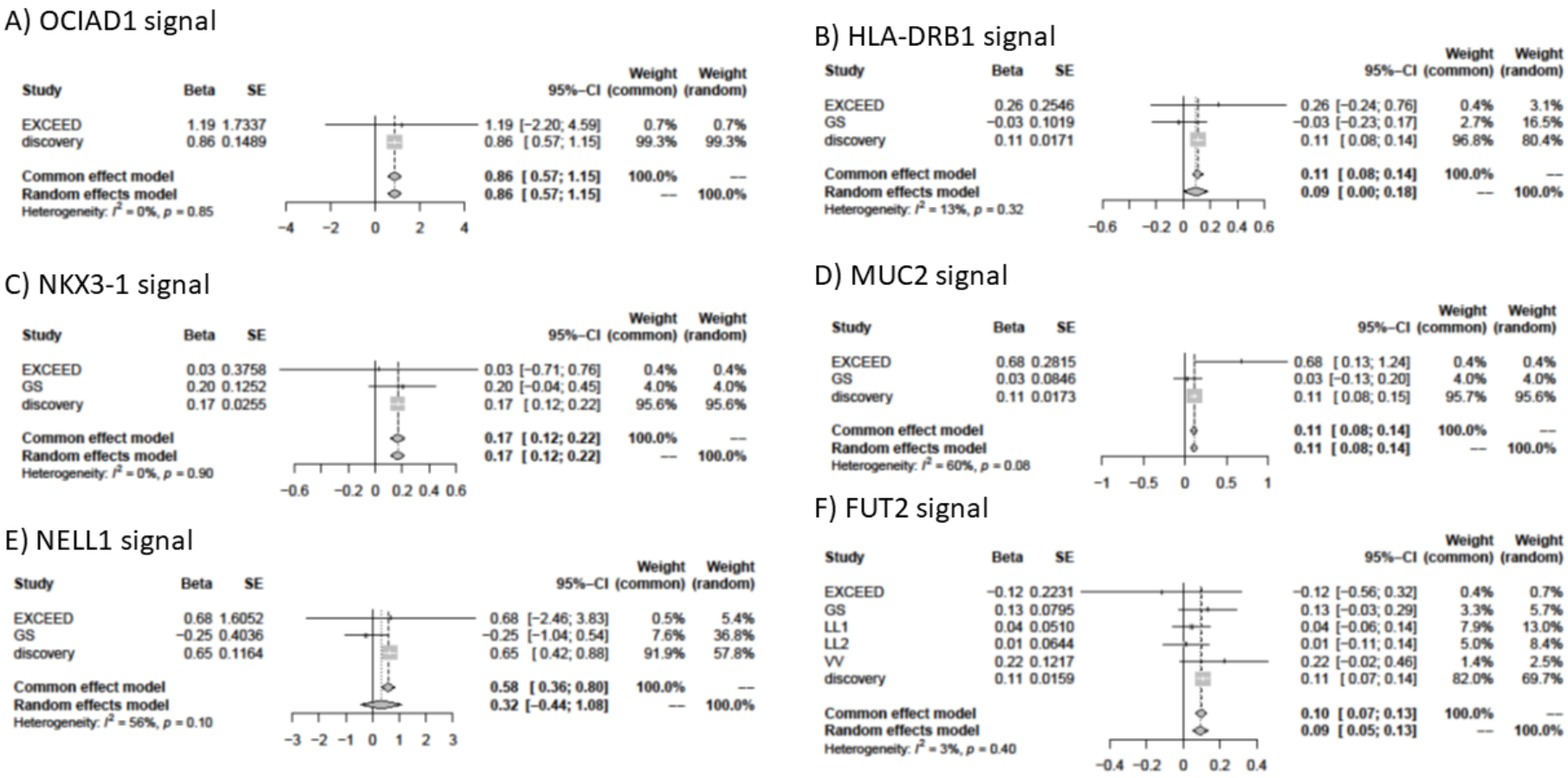
Supplementary Figure S6: Histogram of month that online questionnaire was completed (field-ID 22500) for A) all cases and controls and B) cases only in UK Biobank.



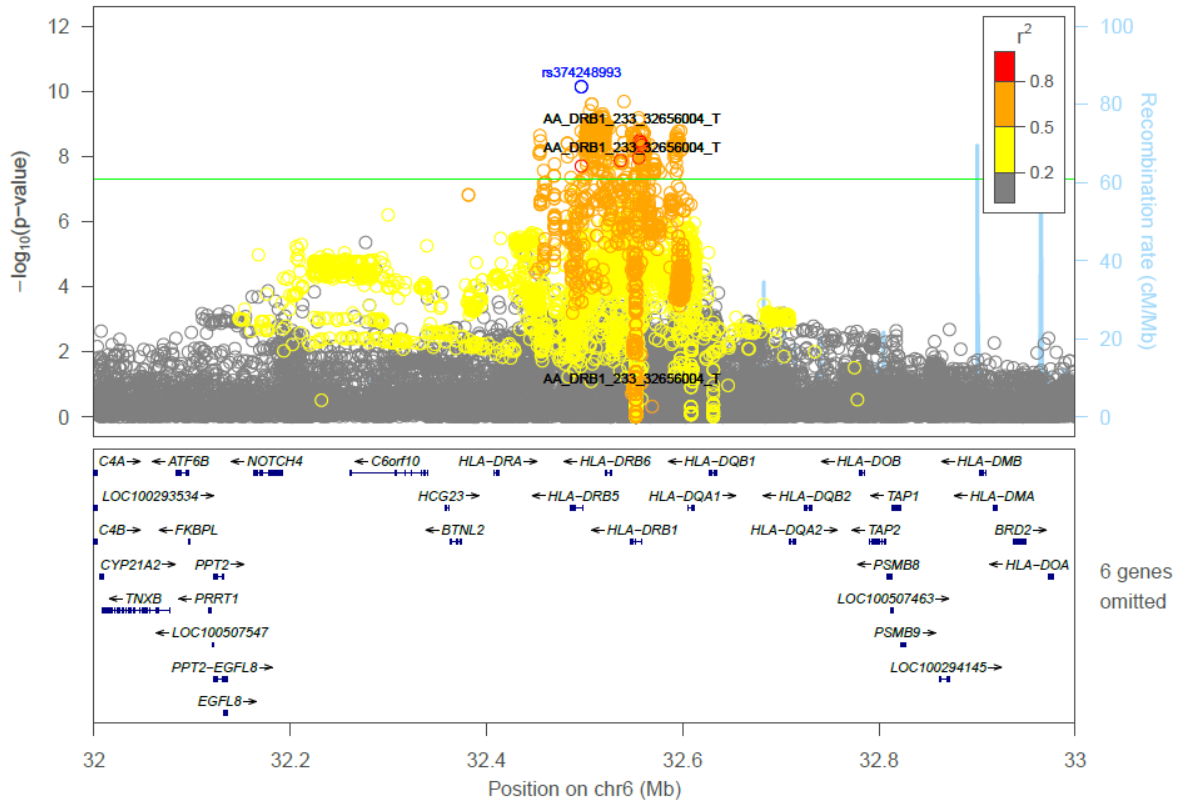
Supplementary Figure S**7:** Plot of beta values for association for sentinel variants in those who completed the online questionnaire in July, August, and September (N cases = 8,416, N controls = 42,627), against beta values for sentinel variants in those who completed the online questionnaire in the 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.



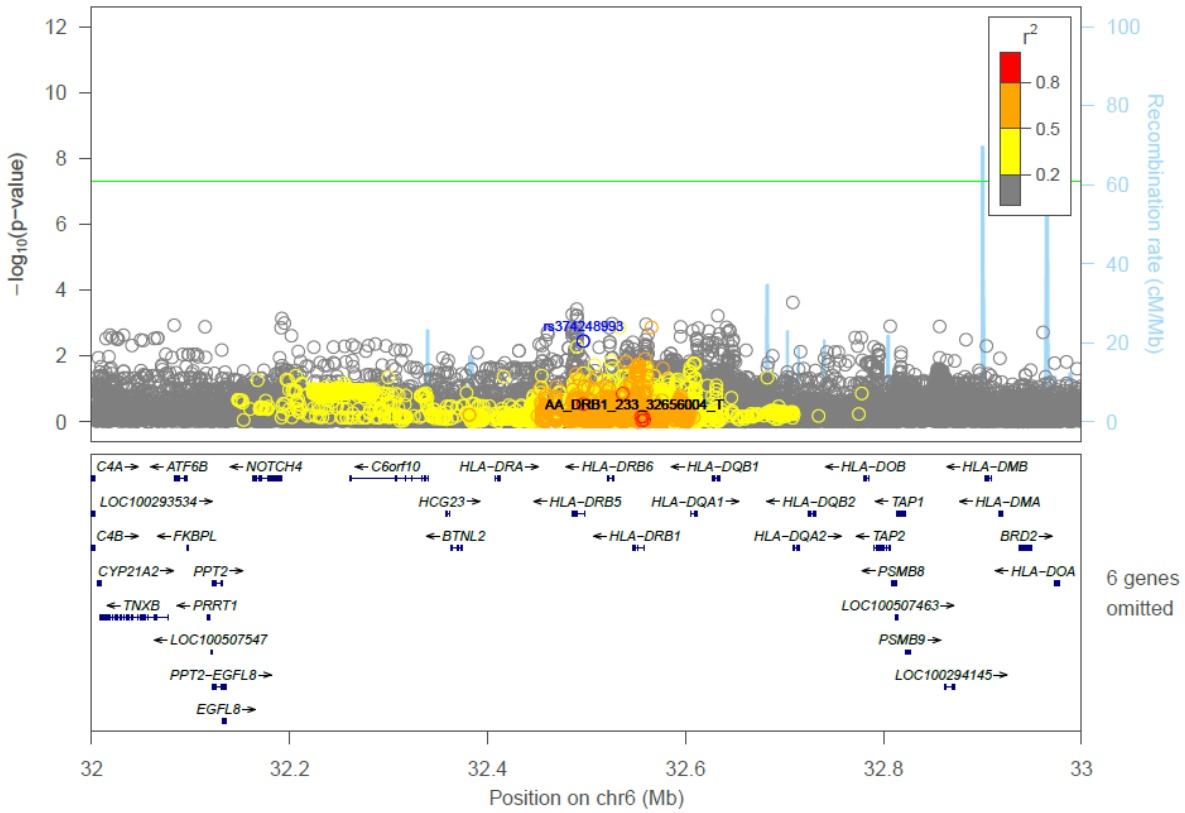
Supplementary Figure S**8:** Plot of beta values for association for sentinel variants with covariate 'current smoker' replacing 'ever smoker' in discovery sample (N cases = 9,714, N controls = 48,471), against beta values for sentinel variants in the discovery with original covariates (N cases = 9,714,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



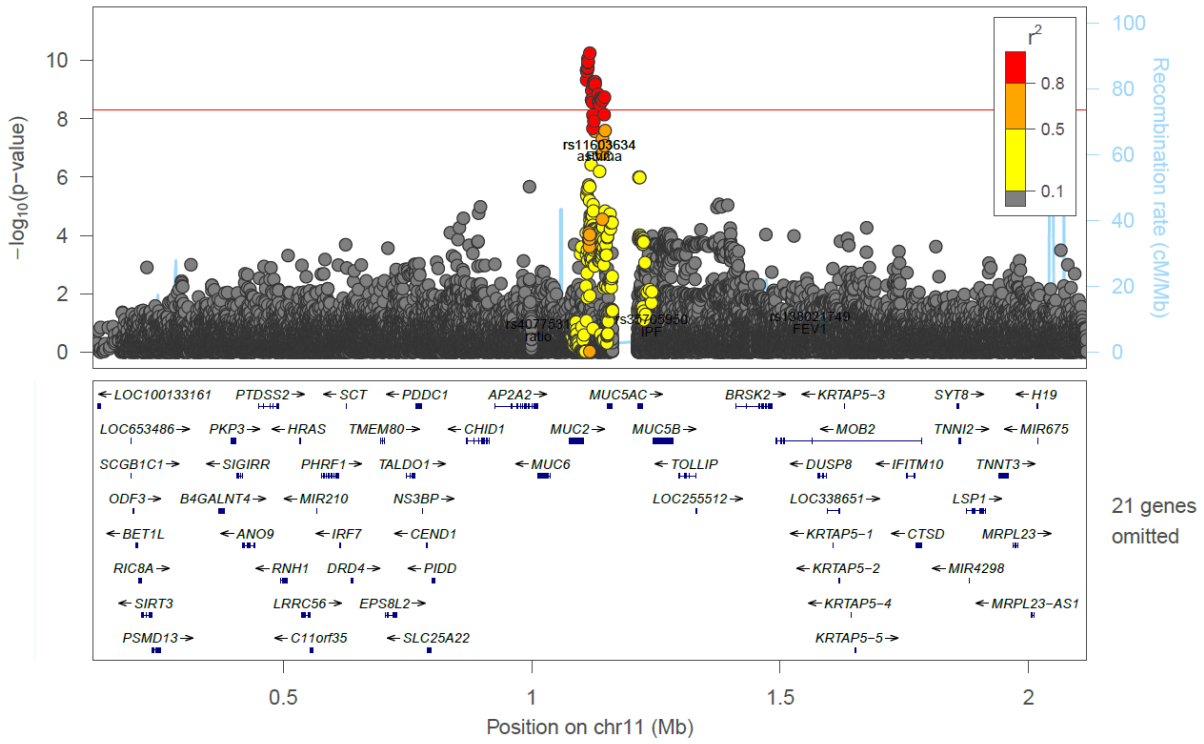
Supplementary Figure S9 Forest plots for results from discovery and replication studies. GS: Generation Scotland, LL1: LifeLines 1, LL2: LifeLines 2, VV: Vlagtwedde-Vlaardingen.



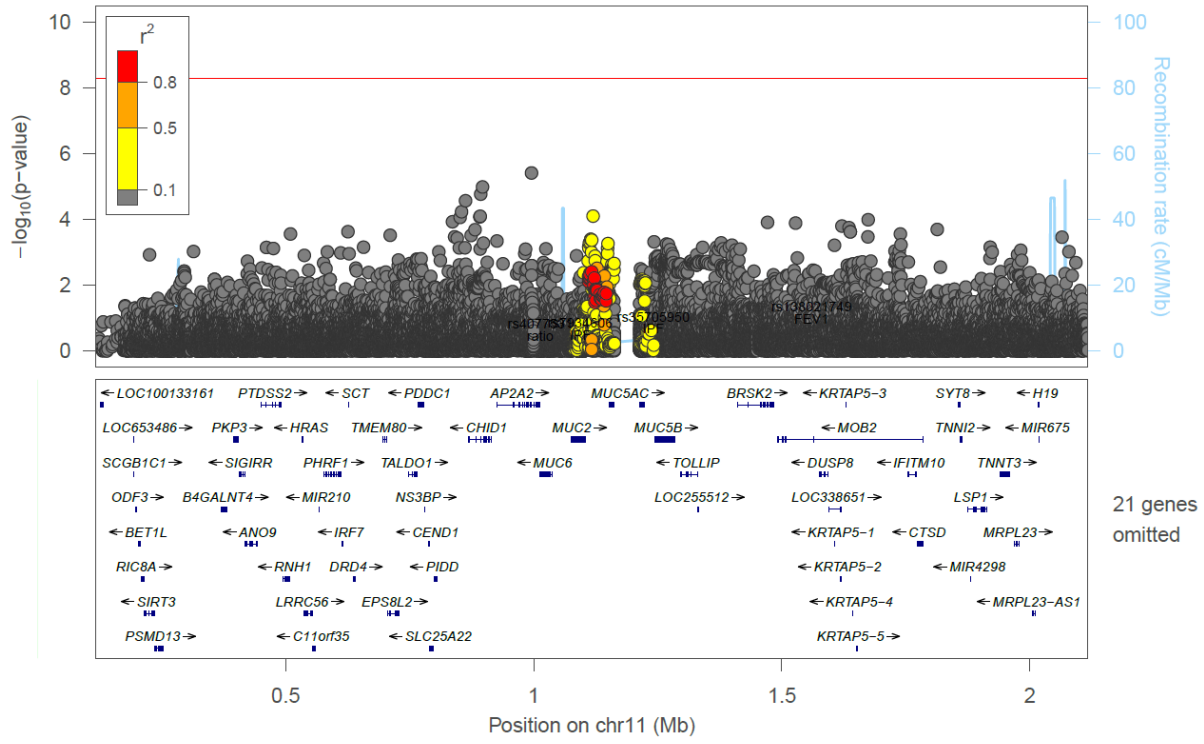
Supplementary Figure S10: Locuszoom plot for combined results SNPs, HLA allele and amino acid changes for *HLA-DRB1* locus, HLA alleles have been labelled in black, sentinel variant is labelled and colored blue.



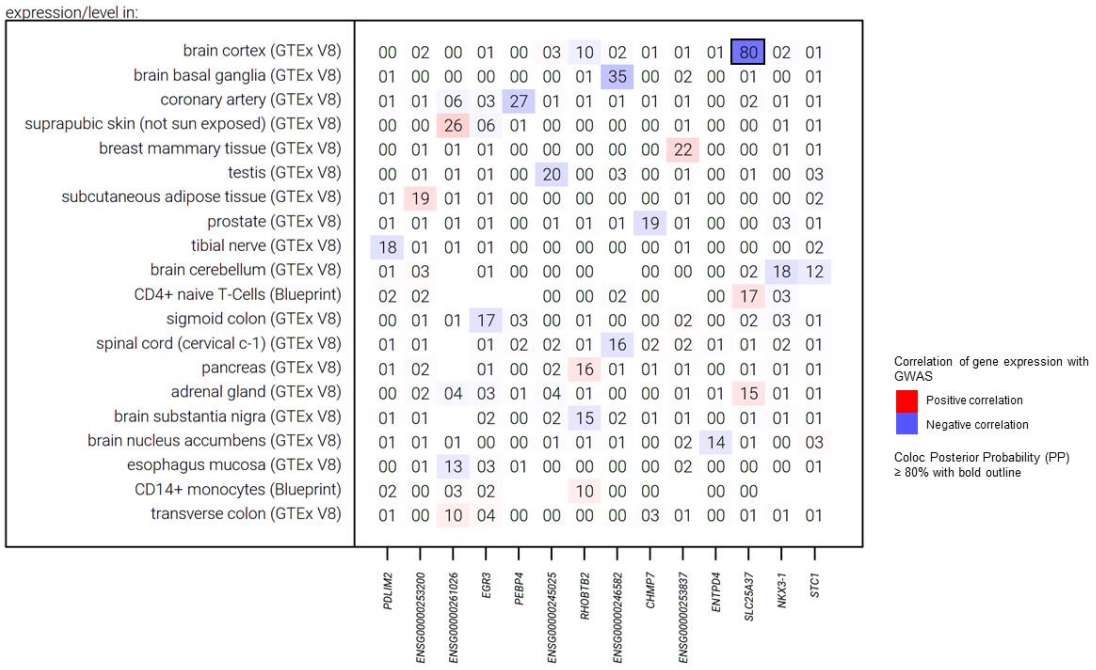
Supplementary Figure S11: Locuszoom plot for combined results SNPs, HLA allele and amino acid changes for *HLA-DRB1* locus conditioned on HLA allele AA_DRB1_233_32656004_T (*DRB1**03:147), HLA alleles have been labelled in black, sentinel variant is labelled and colored blue



Supplementary Figure S12: 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 S13: 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)



Supplementary Figure S14: 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 S1: 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 S2: 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 S3: eQTL tissues and data source searched.
Supplementary Table S4a: 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 S4b: 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 S4c: 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 S5: 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 S6: Results for GWAS look-ups, results aligned the chronic sputum production risk allele for all loci.
Supplementary Table S7: 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 S8: 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 S9: Ensemble variant effect predictor (VEP) results for *FUT2* locus.
Supplementary Table S10: 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 S11: eQTL co-localisation results, results aligned to risk allele for chronic sputum production. H4 = posterior probability of the signals being the same.
Supplementary Table S12: 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 S13: 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.