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

Packer R1, Shrine N2, Melbourne C1, Thompson R1, Williams AT1, Paynton ML1, Davitte J3, Hessel E4, Michalovich D4, Betts JC4, Sayers I5, Yeo A1, Hall IP2, Tobin MD1,5, Wain LV1,5

1 Department of Health Sciences, University of Leicester, Leicester, UK.
2 School of Medicine and NIHR Nottingham Biomedical Research Centre, University of Nottingham, Nottingham, UK
3. GSK R&D, Collegeville, PA, USA
4. GSK R&D, Stevenage, UK
5. Leicester NIHR Biomedical Research Centre, Glenfield Hospital, Leicester, UK

Corresponding Author: Louise V Wain louisewain@leicester.ac.uk

Acknowledgements

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. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. 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.

Conflicts of interest

L.V.W. holds a GSK/British Lung Foundation 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. LVW, MDT, IS and IPH report collaborative research funding from GSK to undertake the submitted work. LVW, MDT and RP 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.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
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 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 the Human Leukocyte Antigen (HLA) locus, chromosome 11 mucin locus (containing MUC2, MUC5AC and MUC5B) and the FUT2 locus. 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.

Interpretation

Novel signals at the FUT2 and mucin loci highlight mucin flucosylation as 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 [1–4] 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 [5].

The determinants of chronic sputum production in disease are not completely understood [6]. Most studies of excess sputum production have been in subjects with chronic bronchitis and COPD where it has been associated with lower lung function [7, 8] and higher risk of both exacerbation and respiratory symptoms [9]. Risk factors for excess sputum production include smoking and occupational and environmental pollutants [8, 10–12]. 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 [13–15].

Genome-wide association studies have highlighted pathways underlying a range of respiratory traits and diseases, and highlighted potentially relevant drug targets [16, 17]. Previous genome wide association studies of sputum production have been limited to cohorts of smokers [18, 19] or those with diagnosed COPD [20, 21] 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. 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 [22]). For this study, we defined cases as those who answered “yes” to the question “do you bring up phlegm/sputum/mucus daily?” (UK Biobank 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 [23], with available smoking data (field 20160). Related individuals were removed, with cases preserved over controls when excluding one of a pair (or more) of related individuals (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 (field 22005) was retained.
From all available controls, we defined a subset of controls with a similar age (field id 34) and sex (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 field 22127), moderate-to-severe asthma (as previously described [24]), doctor-diagnosed chronic bronchitis (field 22129), cough on most days (field 22502), smoking status (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 [23], [25]) bronchiectasis and cystic fibrosis (Supplementary Tables 1 and 2).

The overlap of primary care sputum codes with the chronic sputum production question (UK Biobank 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).

UK Biobank has ethical approval from North West - Haydock Research Ethics Committee (21/NW/0157).

**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.

Association testing was performed using logistic regression under an additive genetic model in PLINK 2.0 [26] 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 [27]. All variant coordinates are for genome build GRCh37. Region plots were created using LocusZoom [28].

**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).

For signals within the HLA region (chr6:29,607,078-33,267,103 (b37)), we re-imputed our discovery samples using IMPUTE2 v2.3.1 with a reference panel for 424 classical HLA alleles and 1,276 amino acid changes [29] and 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 [30], PolyPhen-2 and CADD, all implemented using the Ensemble GRCh37 Variant Effect Predictor (VEP) [31], alongside FATHMM
[32]. 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 [33] and BLUEPRINT [34] (see Supplementary Table 3 for list of tissues). We tested for colocalisation of GWAS and eQTL signals using coloc [35]. 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 twelve respiratory or respiratory-related traits from GWAS results (moderate-to-severe asthma (N cases=5,135, controls=25,675) [24], lung function ( Forced Expired Volume in 1 second [FEV1], Forced Vital Capacity [FVC], FEV1/FVC, peak expiratory flow (PEF)) (N=400,102) [23], respiratory infection (N cases=19,459, controls=101,438) [36], idiopathic pulmonary fibrosis (IPF) (N cases=2,668, controls=8,591) [37], 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) [38]). Smoking trait results were from the UK Biobank component of [39]. 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 72 association tests was applied requiring a P < 6.94x10^{-4} for association to be classified as statistically significant. HLA gene allele or amino acid changes were used in lookups where HLA imputation had been performed (available in the lung function [23], IPF [37], respiratory infection [36] and moderate-to-severe asthma [24] GWAS).

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, or between individuals with and without evidence of chronic respiratory disease separately.

To investigate associations of the chronic sputum-associated variants with a wider range of phenotypes, we performed a phenome-wide association study (PheWAS) of 2172 traits in UK Biobank (FDR<0.01, Supplementary Text) and searched Open Targets Genetics Portal (P<5x10^{-8}, version 0.4.0 (bd664ca) - accessed 16th April 2021 [40]).

**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). A total of 9,714 cases and 48,471 controls (Table 1), and 27,317,434 variants, were included in the GWAS. Ever-smoking and respiratory disease were more common in the cases than in the controls. The genomic control inflation factor (lambda) was 1.026 so no adjustments to the test statistics were
applied (Figure 1). Six independent novel signals met the genome-wide significance threshold of \( P<5\times10^{-8} \) (Figure 2 and Table 2, Supplementary Figures 1 to 6).

**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.

<table>
<thead>
<tr>
<th></th>
<th>Cases N=9714</th>
<th>Controls N=48471</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean age (years)</td>
<td>57.7</td>
<td>57.7</td>
</tr>
<tr>
<td>% female</td>
<td>42.5</td>
<td>42.5</td>
</tr>
<tr>
<td>Ever smoked (%)</td>
<td>5306 (54.6)</td>
<td>20912 (43.1)</td>
</tr>
<tr>
<td>Doctor-diagnosed chronic bronchitis (%)</td>
<td>407 (4.2)</td>
<td>416 (0.86)</td>
</tr>
<tr>
<td>Doctor-diagnosed asthma (%)</td>
<td>2630 (27.1)</td>
<td>5251 (10.8)</td>
</tr>
<tr>
<td>Cough on most days (%)</td>
<td>7022 (72.3)</td>
<td>3999 (8.3)</td>
</tr>
<tr>
<td>Moderate-to-severe asthma (%)</td>
<td>520 (5.4)</td>
<td>521 (1.1)</td>
</tr>
<tr>
<td>Meets spirometry criteria GOLD 1-4 (%)</td>
<td>1511 (21.8)*</td>
<td>4766 (13.1)*</td>
</tr>
</tbody>
</table>

No systematic difference was seen in effect sizes between never and ever smokers for the six sentinel variants (Supplementary Figure 7). We observed that the MUC2 locus signal showed a larger effect amongst those without a history of chronic respiratory disease (and was more similar to the overall results) than those with respiratory disease (Supplementary Figure 8 and Supplementary Table 5). 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).
Figure 1 Quantile-quantile (QQ) plot for results of genome-wide association study of sputum production. Variants with imputation quality INFO <0.5 and minor allele count (MAC) <20 were excluded. The genomic control inflation factor, lambda, was 1.026.

Figure 2 Manhattan plot for genome-wide association study of sputum production. The red line indicates genome-wide significance.
Table 2: Novel genome-wide significant signals of association with sputum production.

<table>
<thead>
<tr>
<th>Chr:Position (GRCh37)</th>
<th>RSID</th>
<th>Locus (BP distance from gene)*</th>
<th>coded/non-coded</th>
<th>Coded allele frequency % (count)</th>
<th>OR (95 CI)</th>
<th>P</th>
<th>Imputation quality (INFO)</th>
<th># variants in 99% credible set (highest posterior probability)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4:48854355</td>
<td>rs79998532</td>
<td>OCIAD1 (intrinsic)</td>
<td>A/G</td>
<td>0.2% (233)</td>
<td>2.36 (1.76-3.16)</td>
<td>8.00x10⁻⁰⁹</td>
<td>0.92</td>
<td>3 (0.86)</td>
</tr>
<tr>
<td>6:32496534</td>
<td>rs374248993</td>
<td>HLA-DRB1 †</td>
<td>G/C</td>
<td>57% (66355)</td>
<td>1.12 (1.08-1.16)</td>
<td>7.30x10⁻¹¹</td>
<td>0.87</td>
<td>HLA-DRB1*03:147†</td>
</tr>
<tr>
<td>8:23480686</td>
<td>rs79401075</td>
<td>NKX3-1 (59,765)</td>
<td>A/G</td>
<td>10% (11620)</td>
<td>1.18 (1.12-1.24)</td>
<td>8.90x10⁻¹¹</td>
<td>0.98</td>
<td>30 (0.32)</td>
</tr>
<tr>
<td>11:1116931</td>
<td>rs779167905</td>
<td>MUC2 (12,513)</td>
<td>T/TTCTA</td>
<td>67% (78158)</td>
<td>1.12 (1.08-1.16)</td>
<td>1.20x10⁻¹⁰</td>
<td>0.98</td>
<td>30 (0.15)</td>
</tr>
<tr>
<td>11:20887601</td>
<td>rs529240826</td>
<td>NELL1 (intrinsic)</td>
<td>GC/G</td>
<td>0.51% (588)</td>
<td>1.91 (1.52-2.4)</td>
<td>2.50x10⁻⁰⁸</td>
<td>0.67</td>
<td>2 (0.83)</td>
</tr>
<tr>
<td>19:49206417</td>
<td>rs492602</td>
<td>FUT2 (exonic)</td>
<td>G/A</td>
<td>51% (58803)</td>
<td>1.11 (1.08-1.15)</td>
<td>3.20x10⁻¹¹</td>
<td>1</td>
<td>32 (0.07)</td>
</tr>
</tbody>
</table>

† = 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
**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=3.43x10⁻⁹). The amino acid change was in linkage disequilibrium with the GWAS sentinel variant rs374248993 (R²=0.74) and the signal for rs374248993 was attenuated when conditioned on the amino acid change (Supplementary Figures 9 and 10).

*HLA-DRB1*03:147 was significantly associated with FEV₁, FEV₁/FVC and PEF at genome-wide significance (P<5x10⁻⁸). The amino acid associated with increased risk of sputum production (threonine) was associated with increased lung function (Supplementary Table 6); this has not been previously reported.

**MUC2 locus**

The mucin locus signal (rs779167905) for risk of chronic sputum production on chromosome 11 was also significantly associated with asthma (OR 1.06, P=0.0027), moderate-to-severe asthma (OR 1.13, P=6.3x10⁻⁷, FVC (beta 0.0087, P=6x10⁻⁴) and IPF (OR 0.84, P=7.5x10⁻⁶)). (Figure 3, Supplementary Table 6). There were no associations with any gene expression for rs779167905.

Genome-wide significant associations with IPF [41] and moderate-severe asthma [24] have previously been reported at this locus and so we undertook a conditional analysis to identify whether the sputum production signal was independent of these previous signals. Repeating the association testing for this variant conditioning on the previously reported variants (rs35705950 [41] and rs11603634 [24]) identified that the sputum GWAS signal was independent of the IPF signal (rs779167905, conditional P=1.18x10⁻⁶) but was not independent of the moderate-to-severe asthma signal (rs779167905, conditional P=0.0039), see Supplementary Figures 11 and 12.

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 A) was associated with higher risk of asthma and asthma-related traits [42–44] and 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, 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 (see Supplementary Table 9 and Supplementary Table 10).

Sentinel variant rs492602 at the *FUT2* locus was associated with expression of *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 FEV₁/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 [43, 45, 46], type 1 diabetes [47] and Crohn’s disease [48–51], elevated vitamin B12 [52–55] and cholesterol and fat metabolites [42, 43, 56–60], hypertension/cardiovascular disease [43, 45, 61], excess alcohol with associated sequelae [45, 62–64], increased risk of mumps and lower risk of childhood ear infections [65]. 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 13).
**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. Smoking age of onset, cigarettes per day and cessation phenotype lookups were omitted as no associations with $P<0.05$ found. *$P < 6.94 \times 10^{-4}$ (Bonferroni adjustment for 72 association tests) ** $P < 5 \times 10^{-8}$. 

GWAS

<table>
<thead>
<tr>
<th>Locus</th>
<th>Asthma</th>
<th>Asthma moderate-severe</th>
<th>FEV1</th>
<th>FEV/ FVC</th>
<th>Hospital respiratory infection</th>
<th>Idiopathic pulmonary fibrosis</th>
<th>PEF</th>
<th>Smoking initiation</th>
</tr>
</thead>
<tbody>
<tr>
<td>OCIAD1</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NKK3-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NELL1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MUC2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA-DRB1</td>
<td>•</td>
<td>•</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FUT2</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Effect direction
- • Increasing
- • Decreasing

Type of SNP
- • Proxy
- • Sentinel

P-value
- • $0.1$
- • $1 \times 10^{-5}$
- • $1 \times 10^{-10}$
- • $1 \times 10^{-20}$
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.
Discussion

We describe the first GWAS of chronic sputum production to identify genome-wide significant signals and our findings implicate genes involved in mucin production and fucosylation, as well as the HLA class II histocompatibility antigen, HLA-DRB1.

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 [66] 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 [67]. Candidate gene studies of this locus have identified that non-secretors (who are associated with high risk of chronic sputum production according to our study) have a lower risk of H. Pylori infection [68], rotavirus A infection [69, 70], norovirus infection [71–73], infant (12-24 months) respiratory illness [74], asthma exacerbations [75], otitis media [76], exacerbation in non-cystic fibrosis bronchiectasis and Pseudomonas aeruginosa airway infection in the same group [77], some evidence of slower HIV progression [73] and a higher risk of pneumococcal and meningococcal infection [78]. 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 [79]. A recent GWAS of critically ill cases of COVID-19 (cases N=7491), showed that the risk allele for chronic phlegm production (G) of rs492602 was protective against life threatening COVID-19 (P=4.55x10^{-9}, OR 0.88, CI 0.87-0.90) [80]. However, this finding was not replicated in the latest COVID-19 Host Genetics Initiative results for a similar phenotype [81].

Epitopes that are fucosylated by FUT2 play a role in cell-cell interaction including host-microbe interaction [82, 83] and mediate interaction with intestinal microbiota, thereby influencing its composition [84–87]. 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 [88]. Sialic acid binding has been shown to be important for adenovirus binding in cell models [89] and modulating this binding has been implicated as a possible mechanism for increasing risk of mumps infection [65].

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). 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 [90]. FUT2 has also been shown to determine the O-glycosylation pattern of Muc5ac in mice [91]. The significant signal at MUC2 in our analysis was not independent of the previously reported moderate-to-severe asthma signal [24] for which MUC5AC was implicated as the most likely causal gene using gene expression data from bronchial epithelial
cells. Although our analysis did not identify an association at the MUC2 locus with COPD-related traits (FEV1 and FEV1/FVC), a recent study has also highlighted MUC5AC as a potential biomarker for COPD prognosis [92].

Our study demonstrates the value in studying a disease-relevant phenotype in a population that is unascertained with respect to inclusion (or exclusion) of respiratory disease or smoking status. 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). 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.

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 [93].

**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].
References


26. Purcell, Shaun C Christopher. Plink 2.0 [Internet]. Available from: www.coggenomics.org/plink/2.0/.


64. 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,


75. 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


