**Genetic Regulators of Sputum Mucin Concentration and Their Associations with COPD Phenotypes**

Eric Van Buren[1,2], Giorgia Radicioni[3], Sarah Lester[4], Wanda K. O'Neal[3], Hong Dang[3], Silva Kasela[5,6], Suresh Garudadri[7], Jeffrey L. Curtis[8,9], Meilan Han[8], Jerry Krishnan[10], Emily S. Wan[11,12], Edwin K. Silverman[11], Annette Hastie[13], Victor E. Ortega[14], Tuuli Lappalainen[5,6], Stephanie A. Christenson[7], Yun Li[1,4], Michael H. Cho[11], Mehmet Kesimer[3], and Samir N. P. Kelada[3,4*]

[1] Department of Biostatistics, University of North Carolina, Chapel Hill, NC
[2] Department of Biostatistics, Harvard T.H. Chan School of Public Health, MA
[3] Marsico Lung Institute, University of North Carolina, Chapel Hill, NC
[4] Department of Genetics, University of North Carolina, Chapel Hill, NC
[5] New York Genome Center, New York, NY, USA
[6] Department of Systems Biology, Columbia University, New York, NY, USA
[7] Division of Pulmonary, Critical Care, Allergy, & Sleep Medicine, Department of Medicine, University of California San Francisco, San Francisco, CA, USA
[8] Pulmonary & Critical Care Medicine Division, University of Michigan, Ann Arbor, MI
[9] Medical Service, VA Ann Arbor Healthcare System, Ann Arbor, MI
[10] Breathe Chicago Center, University of Illinois Chicago
[11] Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA
[12] VA Boston Healthcare System, Jamaica Plain, MA
[13]Department of Internal Medicine, Wake Forest School of Medicine, Winston Salem, NC
[14] Department of Internal Medicine, Division of Respiratory Medicine, Mayo Clinic, Scottsdale, AZ

**Author Contributions:** SNPK had full access to all of the data in the study and takes responsibility for the integrity of the data and accuracy of the analysis. EVB, GR, WKO, ESW, EKS, YL, MHC and SNPK contributed to the conception and design of the study. GR, SR, JLC, MH, JK, ESW, EKS, AH, VEO, TL, SAC, MHC, and MK contributed to the acquisition of the data. EVB, GR, WKO, HD, SK, YL, MHC, and SNPK contributed to data analysis. EVB, JLC, SK, ESW, EKS, AH, VEO, YL, MHC and SNPK contributed to the drafting of the manuscript. All of the authors approved this version of the manuscript to be published.

**Disclosures:**
JLC received consulting fees from AstraZeneca and CLS Bering , received travel support from AstraZeneca Global, and is on the data safety monitoring board or advisory board of Novartis.
MH received grants or contracts from Sanofi, Novartis, Nuvaira, Sunovion, Gala Therapeutics, AstraZeneca, Boehringer Ingelheim, and Biodesix; royalties from Uptodate and Norton Publishing Penguin Random House; consulting fees from AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Novartis, Pulmonx, Teva, Verona, Merck, Mylan, Sanofi, DevPro, Aerogen, Polarian, United Therapeutics, Regeneron, and Altesa BioPharma; honoraria from Cipla, Chiesi, AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Medscape, Integrity, and NACE; and holds stock or stock options Meissa Vaccines and Altesa BioPharma.
ESW received honoraria from Pri-Med.

EKS received grant support from GlaxoSmithKline and Bayer.
VEO is on the Data Safety Monitoring Board or Advisory Board for Sanofi and Regeneron.
TL has stock options from Variant Bio.
SAC received consulting fees from Sanofi/Regeneron, GlaxoSmithKline, AstraZeneca, Glenmark Pharmaceuticals, and Amgen; honoraria from Sanofi/Regeneron, MJH Holdings LLC: Physicians' Education Resource, Sunovion, UpToDate, and Wolters Kluwer Health; travel support from AstraZeneca and GlaxoSmithKline; and sits on the data safety board or advisory boards of Sanofi/Regeneron, and AstraZeneca, and GlaxoSmithKline.
MHC has received grant support from GlaxoSmithKline and Bayer, consulting fees from AstraZeneca, and speaking fees from Illumina.
MK received consulting fees from Arrowhead pharmaceuticals and has a patent pending for mucin measurements in chronic bronchitis.
EVB, GR, SL, WKO, HD, SG, ATH, YL, and SNPK have no conflicts of interest to declare.

***Corresponding Author**
Samir Kelada, PhD MPH
Department of Genetics
120 Mason Farm Road, Room 5072
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599
919-962-2148
samir_kelada@med.unc.edu

**Abstract**

Hyper-secretion and/or hyper-concentration of mucus is a defining feature of multiple obstructive lung diseases, including chronic obstructive pulmonary disease (COPD). Mucus itself is composed of a mixture of water, ions, salt and proteins, of which the gel-forming mucins, MUC5AC and MUC5B, are the most abundant. Recent studies have linked the concentrations of these proteins in sputum to COPD phenotypes, including chronic bronchitis (CB) and acute exacerbations (AE). We sought to determine whether common genetic variants influence sputum mucin concentrations and whether these variants are also associated with COPD phenotypes, specifically CB and AE. We performed a GWAS to identify quantitative trait loci for sputum mucin protein concentration (pQTL) in the Sub-Populations and InteRmediate Outcome Measures in COPD Study (SPIROMICS, n=708 for total mucin, n=215 for MUC5AC, MUC5B). Subsequently, we tested for associations of mucin pQTL with CB and AE using regression modeling (n=822-1300). Replication analysis was conducted using data from COPDGene (n =5740) and by examining results from the UK Biobank. We identified one genome-wide significant pQTL for MUC5AC (rs75401036) and two for MUC5B (rs140324259, rs10001928). The strongest association for MUC5B, with rs140324259 on chromosome 11, explained 14% of variation in sputum MUC5B. Despite being associated with lower MUC5B, the C allele of rs140324259 conferred increased risk of CB (odds ratio (OR) = 1.42; 95% confidence interval (CI): 1.10-1.80) as well as AE ascertained over three years of follow up (OR=1.41; 95% CI: 1.02-1.94). Associations between rs140324259 and CB or AE did not replicate in COPDGene. However, in the UK Biobank, rs140324259 was associated with phenotypes that define CB, namely chronic mucus production and cough, again with the C allele conferring increased risk. We conclude that sputum MUC5AC and MUC5B concentrations are associated with common

genetic variants, and the top locus for MUC5B may influence COPD phenotypes, in particular CB.

Word count: 300

**Author Summary**

Chronic obstructive pulmonary disease (COPD) is characterized by presence of emphysema and/or chronic bronchitis. Excessive mucus production is a defining phenotype of chronic bronchitis, and is associated with several important features of COPD, including exacerbations and loss of lung function. Recent studies have demonstrated that the amount of mucus produced in COPD patients is an important marker of disease state. We investigated whether common genetic variants are associated with the concentration of two key proteins in mucus, MUC5AC and MUC5B, and whether the variants we identified are also associated with COPD outcomes. We identified multiple genetic variants that were associated with MUC5AC or MUC5B concentration. The strongest association we detected, for MUC5B on chromosome 11, was also associated with features of COPD, including chronic bronchitis and acute exacerbations, in one COPD study population but not another. Results from a much larger study, the UK Biobank, indicate that this variant is associated with chronic mucus production and chronic cough, which are key features of chronic bronchitis. Thus, we conclude that the concentration of key proteins in mucus are influenced by genetic variation, and that a variant on chromosome 11 that affects MUC5B may in turn alter COPD outcomes.

Word count: 197

## INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is a smoking-related disease that affects more than 200 million people and is the fourth leading cause of death worldwide [1,2]. The disease is characterized by the presence of emphysema and/or chronic bronchitis (CB). Chronic mucus hyper-secretion is a defining phenotype of CB and is associated with airway obstruction due to mucus plugs [3], acute exacerbations (AE) [4], and accelerated loss of lung function over time [5,6].

Mucus itself is composed of a mixture of water, ions, salt and proteins, and mucin glycoproteins, most prominently the gel-forming mucins, MUC5AC and MUC5B. Their concentrations and biochemical properties (e.g., size and oxidation state) largely determine the viscoelastic properties of mucus in health and disease [7]. Recent studies emanating from the Sub-Populations and InteRmediate Outcome Measures in COPD Study (SPIROMICS) have provided compelling evidence that the concentration of mucin proteins in induced sputum is an important biomarker in COPD [8,9]. Kesimer et al. showed that mucin concentrations (total, MUC5AC and/or MUC5B) was associated with smoking history, phlegm production, CB, risk of AE, and disease severity [8]. A subsequent study showed that the concentrations of MUC5AC was associated with disease initiation and progression [9].

Here, we report our findings of a genome-wide search for common genetic variants associated with variation in sputum mucin protein concentration, that is, mucin protein quantitative trait loci (pQTL). Previous studies have identified genetic variants associated with mucin gene expression (eQTL) that are located within or near *MUC5AC* [10–12] (in asthma) and *MUC5B* (in idiopathic

24 pulmonary fibrosis, IPF [13]). That said, we conducted a genome-wide search for mucin pQTL
25 because our prior work in a mouse model system indicated distal (or trans) pQTL for mucins
26 were possible and perhaps even likely [14]. We leveraged quantitative mass spectrometry-based
27 measurements of samples from SPIROMICS that were generated previously [8,9], to identify
28 main effect pQTL and pQTL that result from genotype × smoking interactions. Subsequently, we
29 tested whether the pQTL we identified were associated with COPD outcomes, namely CB and
30 AE, in SPIROMICS, followed by replication analysis in COPDGene and the UK Biobank.
31
32 **RESULTS**
33 **GWAS for Mucin pQTL**
34 We conducted a GWAS of total and specific (MUC5AC and MUC5B) mucin concentrations in
35 sputum to identify novel regulators of these biomarkers in SPIROMICS. Descriptive statistics of
36 study participants are provided in Table 1. The mucin phenotype data represent a subset of
37 subjects described in two previous studies [8,9], and comprise a subset of participants in
38 SPIROMICS (Figure S1). In this sample, there was a clear effect of smoking history on total
39 mucin concentration (Figure S2), but among COPD cases, there was not a linear or
40 monotonically increasing relationship between total mucin concentration and disease severity (as
41 reflected by Global Initiative for Chronic Obstructive Lung Disease (GOLD) stage). Similar
42 patterns were observed for MUC5AC and MUC5B (Figure S2).
43
44 We did not detect any genome-wide significant loci ($p<5.0x10^{-8}$) associated with total mucin
45 concentration based on data from SPIROMICS participants of European ancestry (EA, N=576)
46 or African ancestry (AA, N=132) participants (Figures S3 and S4), nor in combined analysis of

47 EA and AA subjects (not shown). Testing for joint effects of SNP and SNP × smoking (pack-
48 years) interactions did not reveal any loci associated with total mucin concentration either. In
49 contrast, despite relatively small sample size (N=215 EA subjects), we identified three genome-
50 wide significant pQTL for MUC5AC or MUC5B (Figure 1, Table S1, and Figures S5 and S6). In
51 addition to the pQTL for MUC5AC on chromosome (Chr) 7 (rs75401036), we identified one
52 highly suggestive locus on Chr 2 (rs16866419, $p=7.2x10^{-8}$), thus both MUC5AC pQTL are
53 located on chromosomes other than Chr 11 where *MUC5AC* and *MUC5B* are located (i.e., act in
54 trans). We note that restricting our analysis to variants located in/near MUC5AC, i.e., with a
55 reduced multiple testing correction, did not reveal any local pQTL for MUC5AC, and this
56 includes testing variants previously associated with *MUC5AC* gene expression [10–12] including
57 rs12788104, rs11602802, rs11603634, rs1292198170, and rs1132436. Overall, SNP-based
58 heritability estimates for MUC5AC ($h^2_{SNP}$=0.712; S.E. = 1.322) and MUC5B ($h^2_{SNP}$ = 0.608,
59 S.E.=1.475) were high but imprecise, which is not surprising given the relatively small sample
60 size.
61
62 For MUC5B, one local pQTL was detected on Chr 11 (rs140324259), and one distal pQTL was
63 located on Chr 4 (rs10001928). One additional MUC5B pQTL (rs6043852), located in the intron
64 of *KIF16B* on chromosome 20 was detected by testing for the joint effects of SNP + SNP ×
65 smoking interactions (joint test p-value = $1.3x10^{-9}$, Figure 2A). Further analysis revealed that the
66 interaction itself contributed substantially to the joint effect ($p_{interaction} = 1.1x10^{-7}$), such that
67 rs6043852 was associated with MUC5B concentration only in subjects that are not current
68 smokers (Figure 2B). Given the relatively low minor allele frequency of rs6043852 (3%), the
69 number of subjects harboring genotypes with the minor allele (A) and of contrasting smoking

70 status was not large (n=6 A allele carriers in current smokers and n=6 in the combined never plus
71 former smoker group). Hence this SNP × smoking interaction pQTL must be interpreted with
72 caution.
73
74 The strongest pQTL we identified was for MUC5B on Chr 11. The lead variant, rs140324259, is
75 located approximately 100 kb upstream of *MUC5B,* in between *MUC2* and *MUC5AC* (Figure
76 3A). A second variant located in intron 6 of *MUC5AC*, rs28668859, was also associated with
77 MUC5B concentration; conditional analysis revealed this this signal was partially dependent on
78 linkage disequilibrium (LD, $R^2$=0.20) with rs140324259 (conditional p-value = $1.6x10^{-3}$).
79 Neither rs140324259 nor rs28668859 are in LD ($R^2$=0.02 and 0.01, respectively) with the
80 *MUC5B* promoter variant rs35705950 that is a well-known *MUC5B* eQTL and is associated with
81 IPF [13]. After adjusting for covariates, rs140324259 genotype explained ~14% of variation in
82 sputum MUC5B, and each minor allele (C) contributed a ~2.3 pmol/ml unit decrease in MUC5B
83 (Figure 3B), an effect size that is greater than the effect of current smoking status (yes vs. no,
84 ~1.4 pmol/ml). Adjusting for disease severity (using GOLD stage) did not materially change
85 these results.
86
87 We asked whether lead MUC5B pQTL, rs140324259, was associated with *MUC5B* gene
88 expression in the SPIROMICS participants or other studies. In a subset of SPIROMICS
89 participants (n=144) for whom airway brush RNA-seq data exist [15], we used a tagSNP for
90 rs140324259, namely rs55680540 (LD $R^2$=0.72 in entire SPIROMICS population), but found no
91 correlation between genotype and *MUC5B* expression (Figure S7A). No other variants in the
92 region were significantly associated with *MUC5B* expression (FDR = 0.497, Figure S7B).

93 Additionally, rs140324259 was also not associated with *MUC5B* expression in the nasal
94 epithelium of subjects with cystic fibrosis [16] or asthma [12], nor was it reported as an eQTL in
95 any tissue in the GTEx dataset [17], including homogenized lung tissue.
96
97 Given that power for eQTL detection could be an issue underlying the negative eQTL
98 association results, we asked whether rs140324259 or four variants in LD (rs55680540,
99 rs28668859, rs11604917, and rs76498418) could potentially affect gene expression by altering
100 transcription factor binding or chromatin state using Haploreg [18]. As shown in Table S2,
101 rs140324259, rs11604917, and rs55680540 are predicted to alter binding of transcription factors,
102 and there is some evidence that rs11604917 and rs55680540 alter chromatin state in relevant cell
103 types or tissues (Table S3). Perhaps most notably, rs11604917 lies in an enhancer region in
104 multiple cell types and tissues and is predicted to alter binding of the transcription factor RBP-J.
105 The alternate allele (C) of rs11604917 disrupts the consensus sequence at the first position of an
106 almost invariant motif (Figure S8). Given that RBP-J is part of the Notch signaling pathway that
107 determines ciliated vs. secretory cell fate in murine airways [19], this finding potentially merits
108 further investigation.
109
110 **Association of MUC5B pQTL with COPD Phenotypes**
111 Subsequently, we tested whether rs140324259 was related to clinically-relevant COPD
112 phenotypes, namely CB and AE. In the subset of SPIROMICS participants with complete
113 phenotype, genotype, sputum MUC5B, and clinical outcomes data (n=141), we found that
114 rs140324259 was not associated with CB at baseline/enrollment (p=0.25, Table S4).
115 rs140324259 was not associated with AE in the year prior to enrollment (p=0.14, Figure 4A)

116 unless we accounted for sputum MUC5B concentration (p=0.02, Figure 4B, and Table S5).
117 Surprisingly, in this analysis, we found that while MUC5B concentration was positively
118 associated with AE ($\beta_1$=0.45, p=0.01, Figure 4B), the effect of rs140324259 genotype ($\beta_2$=0.74,
119 p=0.02) was opposite our expectation based on pQTL analysis. That is, the C allele of
120 rs140324259, which was associated with lower MUC5B ($\gamma$ =–0.77, p=2.6 x$10^{-6}$) and therefore
121 would be expected to confer decreased risk of AE, was associated with increased risk of AE
122 compared to the T allele.
123
124 These results suggested the possibility that rs140324259 may exert effects on AE through both
125 direct and indirect paths, the latter via MUC5B (Figure 4B). To examine this further, we
126 employed a mediation analysis approach, based on the framework developed by Baron and
127 Kenny [20], in which the effect of rs140324259 on AE is modeled as the sum of direct
128 (rs140324259 → AE) and indirect paths (rs140324259 → AE via MUC5B). We leveraged an
129 intersection union test [21] to jointly test that both components of the indirect path (rs140324259
130 → MUC5B and MUC5B → AE) are statistically significant. Indeed, we found evidence that this
131 is the case (p=0.02), which is consistent with a model of partial mediation by MUC5B. Thus,
132 overall, we conclude from these results that rs140324259 likely affects AE in two ways, both
133 directly and indirectly, but with contrasting allele effects in each case, and overall the net effect
134 is that rs140324259 C allele confers increased risk of AE. We note also that contrasting direction
135 of effects of rs140324259 → MUC5B (–) and MUC5B → AE (+) likely explain why the
136 magnitude of the association between rs140324259 and AE is weak and therefore not statistically
137 significant (Figure 4A) [20].
138

139 We then examined associations between rs140324259 and clinical phenotypes in the larger
140 SPIROMICS population for which genotype and clinical data exist but there is not sputum mucin
141 concentration data (n≈1,250). In this sample, rs140324259 was associated with CB at baseline
142 (p=0.02, Table 2). Similar to results in the smaller subset of subjects described above, the C
143 allele was associated with increased risk of CB (odds ratio (OR) = 1.42; 95% confidence interval
144 (CI): 1.10-1.80). The effect of rs140324259 on AE in the larger SPIROMICS sample with
145 clinical data was examined using both retrospectively and prospectively ascertained data.
146 rs140324259 was not significantly associated with AE in the year prior to enrollment (Table S6),
147 nor in the year following enrollment (Table S7). However, in both of these analyses, the results
148 were suggestive and the direction of effect was again positive for the C allele.
149
150 We then asked whether rs140324259 genotype was associated with AE over a period of three
151 years of follow up. SPIROMICS participants exacerbation frequency was categorized based on a
152 previous study as never (n=433), inconsistent (n=331) or consistent (n=58) over the three years
153 [22] (see Methods). Using a proportional odds model, we analyzed whether rs140324259
154 genotype distinguished never exacerbators versus inconsistent and consistent AE, and whether
155 rs140324259 genotype distinguished consistent exacerbators versus never and inconsistent
156 exacerbators. We found that rs140324259 genotype was associated with the former contrast
157 (p=0.03), with the C allele conferring increased risk of being either an inconsistent or consistent
158 exacerbator (Tables S8 and S9). rs140324259 genotype was not associated with the contrast
159 between consistent exacerbators versus never and inconsistent exacerbators. To simplify the
160 interpretation of the effect of rs140324259 on prospectively ascertained AE, we then
161 dichotomized subjects into two groups, those who experienced AE (inconsistent and consistent)

162 versus those that did not. As shown in Table 3, the rs140324259 C allele conferred increased risk
163 of AE over three years of follow up (OR=1.41; 95% CI: 1.02-1.94).
164
165 **Replication analyses**
166 Finally, we analyzed data from the COPDGene study population and UK Biobank in an attempt
167 to replicate results from SPIROMICS. For COPDGene, we utilized phenotype data from COPD
168 cases of European ancestry, and genotype data was based on whole genome sequencing from
169 TOPMed. Sample sizes in these analyses ranged from 5300-5700 depending on the outcome. In
170 this population of COPD cases, rs140324259 was not associated with CB (OR = 1.08 (95% CI:
171 0.94-1.24), nor AE (Supplementary Tables S10-S12), though we were unable to directly evaluate
172 whether rs140324259 was associated with the exacerbation frequency categories (never,
173 inconsistent, consistent) described in SPIROMICS. In the UK Biobank, however, we found that
174 rs140324259 was associated with two CB-related phenotypes, namely bringing up
175 phlegm/sputum/mucus on most days and cough on most days (Table 4). Importantly, the C allele
176 was enriched in cases vs. controls for these two phenotypes, thus these results are directionally
177 consistent with results from SPIROMICS. Results for other variants in LD are shown in Table
178 S13. As UK Biobank results were not adjusted for smoking, we additionally assessed whether
179 rs140324259 was associated with smoking. rs140324259 was either not associated with
180 smoking history variables or was weakly associated, but in these cases the C allele frequency
181 was higher in controls than cases (Table 4), suggesting that the associations between
182 rs140324259 and CB-related phenotypes in the UK Biobank are unlikely to be mediated by, or
183 confounded with, smoking.
184

185 **DISCUSSION**

186 Using quantitative measurements of sputum mucin concentrations, we identified three genome-
187 wide significant loci and one highly suggestive locus associated with MUC5AC or MUC5B. The
188 strongest signal we detected, with rs140324259, accounted for a large percent of variation in
189 MUC5B, and is independent of the common *MUC5B* promoter variant associated with IPF.
190 Surprisingly, rs140324259 does not appear to be an eQTL for *MUC5B*, though we note that our
191 sample size for eQTL analysis was not large and that the tagSNP we used is not in very high LD
192 with rs140324259. The mechanism by which rs140324259 (or a variant in LD) acts as a pQTL
193 remains to be determined. One nearby variant, rs11604917, is intriguing given that it potentially
194 disrupts binding of the transcription factor RBP-J, a key player in the Notch signaling pathway
195 that determines ciliated vs. secretory cell fate in murine airways [19]. This could suggest that the
196 MUC5B pQTL is a function of cell type composition of the airway epithelium, an idea supported
197 by the lack of an association with gene expression. However, this variant is in low LD with
198 rs140324259, and the association of rs11604917 with CB-related phenotypes in the UK Biobank
199 was not nearly as strong as for rs140324259, arguing against a causal role for rs11604917. Still,
200 future studies will need to address the possibility that this locus affects sputum MUC5B
201 concentration without a corresponding effect on *MUC5B* mRNA, which has been observed for
202 other genes [23,24].

203

204 Given that previous studies have identified eQTL for *MUC5AC* [10–12] in asthma and *MUC5B*
205 [13] in IPF located near the genes themselves ("local eQTL"), one potential *a priori* prediction
206 could have been that these same variants would be associated with MUC5AC and MUC5B
207 protein concentrations. This was not the case, even in the context of a regional association

208 analysis (i.e., not a genome-wide significance threshold). This is perhaps not surprising for at
209 least two reasons. First, there are clear differences between our study and the previous studies as
210 a function of disease state (COPD vs. asthma vs. IPF) and anatomical location (upper vs. lower
211 airways). Second, mucin protein concentration is the product of several pathways beyond just
212 mucin gene transcription, including protein synthesis, post-translation modifications, packaging
213 into vesicles, secretion, airway hydration via ion transport, and mucociliary clearance. Thus, one
214 could reasonably expect that genetic variants that regulate any of these processes could be
215 associated with mucin concentration. It remains to be determined whether any of the distal/off-
216 chromosome pQTL identified here play a role in one or more of these pathways. That we did not
217 identify any associations in/near genes with known roles in these processes suggests that either
218 we were underpowered to detect these associations and/or that there is limited functional genetic
219 variation in/near these genes.
220
221 Our analysis of associations between rs140324259 and clinical outcomes, namely CB and AE,
222 produced intriguing results in the SPIROMICS cohort, including that the variant was associated
223 with AE ascertained prospectively over three years. These results did not replicate in
224 COPDGene, but we did find an association with sputum production and cough in a much larger
225 dataset, the UK Biobank. These data argue in support of a role for the MUC5B pQTL in CB-
226 related phenotypes. However, we acknowledge that the results of CB and AE association
227 analyses with rs140324259 in SPIROMICS would not a survive multiple testing correction based
228 on the number of outcomes/models we evaluated; in addition, we were unable to replicate these
229 results COPDGene, thus raising the potential that the results in SPIROMICS represent false
230 positives. We note here that failure to replicate genetic associations with AE is unfortunately

231 common [25], and future studies in which standardized definitions of AE can be employed will
232 certainly facilitate the best comparisons across studies [25]. It is also worth noting that a previous
233 study identified significant blood biomarkers of susceptibility for AE in SPIROMICS and
234 separately in COPDGene, but there was essentially no overlap in associations between the two
235 populations [26], which points to the difficulty in identifying reproducible predictors of
236 exacerbations. The UK Biobank analysis, while supportive of our results, did not have the same
237 degree of detailed respiratory phenotypes and was performed in a general population sample.
238 Additional analyses adjusting for disease state and other covariates could be beneficial [27]).
239 Further attempts to replicate these finding in other populations would also be useful, in particular
240 to address the question of generalizability across populations of different genetic ancestries.
241
242 In aggregate, the results of association tests between rs140324259 and COPD phenotypes
243 suggest an apparent paradox. While MUC5B concentration was positively associated with AE
244 and CB in SPIROMICS, and the C allele was associated with significantly reduced MUC5B, the
245 C allele overall was associated increased risk of AE and CB. This result suggests that higher
246 expression of MUC5B may in fact be protective against AE and CB, perhaps by virtue of
247 normalizing the ratio of elevated MUC5AC to MUC5B, making it more clearable, as has been
248 suggested before in relationship to the IPF-associated variant rs35705950 [28].
249
250 While we examined associations between loci associated with mucins, CB, and AE in COPD
251 patients specifically, others have examined the genetics of CB/chronic mucin hypersecretion in
252 combined analysis of the general population and patients with COPD [27,29] or in smokers
253 without COPD [30]. In the study with COPD cases and the general population [29], the most

254 consistent association signal was for rs6577641, which was also shown to act as an eQTL for the
255 gene *SATB1*. In look up analysis, this variant was not associated with either sputum MUC5B
256 concentration or CB in the SPIROMICS population, nor was the lead variant (rs10461985) from
257 another study [30]. The most recent study reported an association of variants on proximal Chr 11
258 (near *MUC2*) with chronic sputum production using the same UK Biobank phenotype codes we
259 used [27], but the LD between lead variant in that study (rs779167905) and rs140324259 is
260 minimal ($R^2$=0.08), making it unlikely that these are the same signals.
261
262 In summary, we identified pQTL for MUC5AC and MUC5B in sputum, demonstrating that
263 common genetic variants influence these biomarkers. The lead MUC5B pQTL, rs140324259,
264 was associated with CB and prospectively ascertained AE in SPIROMICS and was also
265 associated with CB-related phenotypes in the UK Biobank. Additional studies are needed to
266 further evaluate whether rs140324259 may be a biomarker of CB and AE susceptibility in COPD
267 in other populations, and to determine how this variant influences MUC5B concentration in
268 sputum.
269
270 **MATERIALS AND METHODS**
271 **Ethics statement**
272 Subjects provided informed consent to participate in the studies described here. Details and
273 institutional review boards for each clinical site are provided in Supporting File 1.
274
275 **Data availability**
276 We will make as much subject-level data available as possible upon acceptance of the

277 manuscript, and GWAS summary statistics will be provided. We note, however, that not all
278 subjects consented to sharing their genotype data publicly. Thus, the data we share represents all
279 of the data than can be shared according to the consents provided. The distributions of
280 demographic, exposure, and clinical variables (age, sex, smoking status, disease severity) were
281 not significantly different between subjects who consented to genotype data sharing vs. those
282 that did not.
283
284 **Study subjects and genotype data**
285 The primary analyses presented here are based on study participants in SPIROMICS
286 (ClinicalTrials.gov Identifier: NCT01969344), and a schematic of the SPIROMICS datasets used
287 here is shown in Figure S1. The study design has been described previously [31]. SPIROMICS
288 participants were genotyped using the Illumina OmniExpress Human Exome Beadchip [32].
289 Quality controls included testing for sex concordance and removal of SNPs with high genotype
290 missing rates (>5%) and/or Hardy Weinberg $p < 1x10^{-6}$. Genotype imputation was performed
291 using the Michigan Imputation Server [33] using haplotypes from Phase3 of the 1000 Genomes
292 Project [34]. Study participants were categorized into either European ancestry (EA, N=576 or
293 African ancestry (AA, N=132) groups based on genotype data. We used an adaptive $R^2$ threshold
294 to filter imputed variants in each ancestry group based on the minor allele frequency (MAF). For
295 each MAF interval, the $R^2$ value was chosen such that the average $R^2$ for variants with values
296 larger than the threshold is at least 0.8 (Tables S14 and S15). We limited our analyses to SNPs
297 with minor allele counts >8, resulting in ~10 million SNPs in EA and 12 million variants in AA
298 subjects for association with total and specific mucin concentrations.
299

300 In COPDGene [35] (ClinicalTrials.gov Identifier: NCT00608764), genotype data for
301 rs140324259 was obtained from whole genome sequencing performed through the TOPMed
302 consortium [36]. Results from the UK Biobank data were obtained from the Pan-UK Biobank
303 analysis (see further description below) [37].
304
305 **Sputum Mucin Phenotype data**
306 Sputum mucin concentration: sputum induction and measurement methods have been previously
307 reported [8,9,38]. In brief, hypertonic saline was used to induce sputum, which was then placed
308 in a buffer containing 6 molar guanidine, and stored at 4 degrees. Total sputum mucin
309 concentration was determined using a size exclusion chromatography / differential refractometry
310 measurement approach. For a subset of subjects, MUC5AC and MUC5B concentration was
311 determined using stable isotope labeled mass spectrometry [38]. Data were generated in two
312 batches. In addition to SPIROMICS participants with COPD (n=439), two additional sets of
313 subjects were also included in the mucin analyses: non-smoking controls (n=50), and smokers
314 without COPD that are referred to as the "at-risk" group (n=219). These subjects were included
315 in genetic analysis of sputum mucin concentration but were not included in the analysis of
316 COPD outcomes.
317
318 **Clinical/Phenotype Data**
319 We analyzed data on two COPD phenotypes, namely CB and AE, in SPIROMICS. The CB
320 phenotype was ascertained at the first study visit ("baseline") and was categorized based on
321 participants' responses to questions regarding frequency of cough and mucus/phlegm production
322 in the St. George's Respiratory Questionnaire. The analysis of AE was based on previous work

323 from SPIROMICS [22,31], in which AE were defined as events that required health care
324 utilization (i.e., office visit, hospital admission, or emergency department visit for a respiratory
325 flare-up) involving the use of antibiotics or systemic corticosteroids, or both. In COPDGene,
326 chronic bronchitis was based on chronic cough and phlegm production for ≥ 3 mo/y for 2
327 consecutive years [5]. For AE, self-reported moderate-to-severe exacerbations in the year prior to
328 enrollment and the number of moderate-to-severe exacerbations ascertained prospectively from
329 longitudinal follow up data were examined. In the Pan-UK Biobank
330 (https://pan.ukbb.broadinstitute.org/), we evaluated results of association analyses for two CB-
331 related phenotypes, namely bringing up phlegm/sputum/mucus daily (yes vs. no, phenocode
332 22504) and coughing on most days (yes vs. no, phenocode 22502), as well as smoking history
333 variables, of which were assessed by questionnaire.
334
335 **Statistical models**
336 Sputum mucin concentration: Data on total and specific (MUC5AC and MUC5B) mucin
337 concentrations were log-transformed prior to analysis. GWAS analysis was performed using
338 version 0.5.0 of the ProbABEL software [39]. Analysis of total mucin concentration was
339 conducted in each ancestry group separately (N=576 EA and 132 AA), followed by a pooled
340 analysis of both ancestry groups. In ancestry-specific analyses, main effect SNP models of each
341 mucin phenotype included covariates for the top two principal components of ancestry (PC)
342 obtained from EIGENSTRAT [40], age, sex, batch of mucin quantitation analysis, current
343 smoking status, smoking pack-years, and CB. Results were not materially different when we
344 included up to 10 genotype PCs. For MUC5AC and MUC5B, GWAS was performed in EA
345 subjects only (N=215) with the same covariates used for total mucin concentration. SNP-based

346 heritability for MUC5AC and MUC5B was estimated on an LD-pruned set of markers from the
347 genotyped data (subsetting to individuals with the relevant phenotype data) using GCTA version
348 1.92.1. We performed exploratory genome-wide interaction studies of SNP × smoking
349 interactions in which we tested for the joint effects of SNP and SNP × smoking interactions (2
350 d.f. test) on mucin concentrations in models including the same covariates as above.
351
352 eQTL Analysis: Airway epithelia gene expression from 144 SPIROMICS participants was
353 analyzed to test whether rs140324259 is an eQTL for *MUC5B* by performing a genome-wide
354 eQTL mapping as described before in Kasela et al. [15]. Briefly, RNA-seq data from the airway
355 epithelium was normalized, filtered, and transformed using inverse normal transformation.
356 Genotype data was obtained from TOPMed (Freeze 9) [36]. The eQTL regression model for a
357 given gene included sex, four genotype PCs, and 15 PEER factors (probabilistic estimation of
358 expression residuals [41]) as covariates. eQTL mapping was performed using tensorQTL [42]
359 and 10,000 permutations were used to control for multiple testing at false discovery rate (FDR) <
360 0.05. To look up the eQTL association with *MUC5B*, we used the proxy SNP rs55680540
361 because rs140324259 did not pass variant filter quality control.
362
363 Clinical Phenotypes:
364 Chronic Bronchitis (CB): Following the analyses of sputum mucin concentration data, we tested
365 for an association between the lead variant for sputum MUC5B concentration (rs140324259) and
366 CB in the larger SPIROMICS population (N=1257). Logistic regression models were used for
367 CB, accounting for the top two genotype PCs, age, sex, current smoking status, pack-years of
368 smoking, and FEV1.

369

370 Acute Exacerbations (AE): Exacerbation outcomes were modeled using negative binomial
371 regression models including the same covariates as above. Additionally, in the analysis of
372 prospectively ascertained AE, we included AE in the year prior to enrollment as a predictor.
373 Because prior work showed that exacerbation frequency among subjects with COPD in
374 SPIROMICS is not stable [22], we leveraged a previously developed classification system which
375 categorized SPIROMICS participants as never, inconsistent or consistent exacerbators using
376 three years of follow up data [22]. Consistent exacerbators were subjects who experienced at
377 least one acute exacerbation in each of the three years; subjects who had an exacerbation during
378 some but not all of the three years of follow up where defined as inconsistent exacerbators. We
379 analyzed the association between rs140324259 genotype and these three exacerbation groups
380 using a proportional odds model, comparing (1) never exacerbators versus inconsistent and
381 consistent exacerbators, and (2) consistent exacerbators versus never and inconsistent
382 exacerbators. Based on the results of these analyses, we collapsed the exacerbation groups into
383 two categories: ever (combining inconsistent and consistent exacerbators) vs. never exacerbators,
384 then modeled this outcome using logistic regression with covariates for top two genotype PCs,
385 age, sex, current smoking status, pack-years of smoking, FEV1 (% predicted), and the number of
386 AE in the year prior to enrollment.

387

388 Mediation analysis: in the subset of SPIROMICS subjects for which there is complete phenotype
389 data on genotype, sputum MUC5B, and clinical outcomes (N=141), we tested for evidence of
390 that MUC5B mediates an association between rs140324259 and AE (i.e. rs140324259 →
391 MUC5B → AE), invoking the overall mediation analysis framework of Baron and Kenny [20].

392 We evaluated a direct path from rs140324259 → AE (c), and an indirect path from rs140324259
393 → MUC5B (a) and MUC5B → AE (b), while also examining the path from rs140324259 → AE
394 conditional on MUC5B (c'). All regression models included age, sex, two ancestry PCs, current
395 smoking status, pack-years of smoking, and FEV1 (% predicted) as predictors. For (a), we used a
396 linear model for MUC5B in which rs140324259 was coded linearly (0,1,2). For (b), (c), and (c'),
397 we used we used negative binomial regression models of AE that included rs140324259 (c),
398 MUC5B (b), or both (c'). To formally test for mediation, we leveraged the SNP mediation
399 intersection-union test (SMUT) [21] which jointly tests for non-zero parameter estimates from
400 models for (a) and (b), which is equivalent to testing that a x b is not equal to 0 in the Baron and
401 Kenny framework.
402
403
404

**Acknowledgements**

Y.L. was partially supported by NIH grants R01GM105765 and U01DA052713. S.K. and T.L. were supported by NIH grant R01HL142028. S.N.P.K. was partially supported by NIH grant R01 HL122711.

The authors thank Dr. Wesley Crouse for helpful discussions related to mediation analysis. The authors are grateful to SPIROMICS participants and participating physicians, investigators and staff, for making this research possible. More information about the study and how to access SPIROMICS data is available at www.spiromics.org. The authors would like to acknowledge the University of North Carolina at Chapel Hill BioSpecimen Processing Facility for sample processing, storage, and sample disbursements (http://bsp.web.unc.edu/).

We would like to acknowledge the following current and former investigators of the SPIROMICS sites and reading centers: Neil E Alexis, MD; Wayne H Anderson, PhD; Mehrdad Arjomandi, MD; Igor Barjaktarevic, MD, PhD; R Graham Barr, MD, DrPH; Patricia Basta, PhD; Lori A Bateman, MSc; Surya P Bhatt, MD; Eugene R Bleecker, MD; Richard C Boucher, MD; Russell P Bowler, MD, PhD; Stephanie A Christenson, MD; Alejandro P Comellas, MD; Christopher B Cooper, MD, PhD; David J Couper, PhD; Gerard J Criner, MD; Ronald G Crystal, MD; Jeffrey L Curtis, MD; Claire M Doerschuk, MD; Mark T Dransfield, MD; Brad Drummond, MD; Christine M Freeman, PhD; Craig Galban, PhD; MeiLan K Han, MD, MS; Nadia N Hansel, MD, MPH; Annette T Hastie, PhD; Eric A Hoffman, PhD; Yvonne Huang, MD; Robert J Kaner, MD; Richard E Kanner, MD; Eric C Kleerup, MD; Jerry A Krishnan, MD, PhD; Lisa M LaVange, PhD; Stephen C Lazarus, MD; Fernando J Martinez, MD, MS; Deborah A Meyers, PhD; Wendy C Moore, MD; John D Newell Jr, MD; Robert Paine, III, MD; Laura Paulin, MD, MHS; Stephen P Peters, MD, PhD; Cheryl Pirozzi, MD; Nirupama Putcha, MD, MHS; Elizabeth C Oelsner, MD, MPH; Wanda K O'Neal, PhD; Victor E Ortega, MD, PhD; Sanjeev Raman, MBBS, MD; Stephen I. Rennard, MD; Donald P Tashkin, MD; J Michael Wells, MD; Robert A Wise, MD; and Prescott G Woodruff, MD, MPH. The project officers from the Lung Division of the National Heart, Lung, and Blood Institute were Lisa Postow, PhD, and Lisa Viviano, BSN; SPIROMICS was supported by contracts from the NIH/NHLBI (HHSN268200900013C, HHSN268200900014C, HHSN268200900015C, HHSN268200900016C, HHSN268200900017C, HHSN268200900018C, HHSN268200900019C, HHSN268200900020C), grants from the NIH/NHLBI (U01 HL137880 and U24 HL141762), and supplemented by contributions made through the Foundation for the NIH and the COPD Foundation from AstraZeneca/MedImmune; Bayer; Bellerophon Therapeutics; Boehringer-Ingelheim Pharmaceuticals, Inc.; Chiesi Farmaceutici S.p.A.; Forest Research Institute, Inc.; GlaxoSmithKline; Grifols Therapeutics, Inc.; Ikaria, Inc.; Novartis Pharmaceuticals Corporation; Nycomed GmbH; ProterixBio; Regeneron Pharmaceuticals, Inc.; Sanofi; Sunovion; Takeda Pharmaceutical Company; and Theravance Biopharma and Mylan.

The COPDGene project described was supported by Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health. The COPDGene project is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Novartis,

Pfizer, Siemens and Sunovion. A full listing of COPDGene investigators can be found at: http://www.copdgene.org/directory

Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Genome sequencing for "NHLBI TOPMed: Genetic Epidemiology of COPD (COPDGene)" (phs000951.v4.p4) was performed at the Northwest Genomics Center (3R01HL089856-08S1) and Broad Institute Genomics Platform (HHSN268201500014C). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

## References


1. Diaz-Guzman E, Mannino DM. Epidemiology and prevalence of chronic obstructive pulmonary disease. Clin Chest Med. 2014;35: 7–16. doi:10.1016/j.ccm.2013.10.002
2. Ferkol T, Schraufnagel D. The global burden of respiratory disease. Ann Am Thorac Soc. 2014;11: 404–406. doi:10.1513/AnnalsATS.201311-405PS
3. Dunican EM, Elicker BM, Henry T, Gierada DS, Schiebler ML, Anderson W, et al. Mucus Plugs and Emphysema in the Pathophysiology of Airflow Obstruction and Hypoxemia in Smokers. Am J Respir Crit Care Med. 2021;203: 957–968. doi:10.1164/rccm.202006-2248oc
4. Ramos FL, Krahnke JS, Kim V. Clinical issues of mucus accumulation in COPD. Int J Chron Obstruct Pulmon Dis. 2014;9: 139–50. doi:10.2147/COPD.S38938
5. Kim V, Han MK, Vance GB, Make BJ, Newell JD, Hokanson JE, et al. The chronic bronchitic phenotype of COPD: An analysis of the COPDGene study. Chest. 2011;140: 626–633. doi:10.1378/chest.10-2948
6. Vestbo J, Prescott EVA, Lange P, Copenhagen T, Heart C, Group S. Association of Chronic Mucus Hypersecretion with FEV 1 Decline and Chronic Obstrurtive Pulmonary Disease Morbidity. Am J Respir Crit Care Med. 1996;153: 1530–1535.
7. Boucher RC. Muco-Obstructive Lung Diseases. N Engl J Med. 2019;380: 1941–1953. doi:10.1056/NEJMra1813799
8. Kesimer M, Ford AA, Ceppe A, Radicioni G, Cao R, Davis CW, et al. Airway Mucin Concentration as a Marker of Chronic Bronchitis. N Engl J Med. 2017;377: 911–922. doi:10.1056/NEJMoa1701632
9. Radicioni G, Ceppe A, Ford AA, Alexis NE, Barr RG, Bleecker ER, et al. 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;2600: 1–14. doi:10.1016/s2213-2600(21)00079-5
10. Shrine N, Portelli MA, John C, Soler Artigas M, Bennett N, Hall R, et al. Moderate-to-severe asthma in individuals of European ancestry: a genome-wide association study. Lancet Respir Med. 2019;7: 20–34. doi:10.1016/S2213-2600(18)30389-8
11. Altman MC, Flynn K, Rosasco MG, Dapas M, Kattan M, Lovinsky-Desir S, et al. Inducible expression quantitative trait locus analysis of the MUC5AC gene in asthma in urban populations of children. J Allergy Clin Immunol. 2021; S0091-6749. doi:10.1016/j.jaci.2021.04.035
12. Sajuthi SP, Everman JL, Jackson ND, Saef B, Rios CL, Moore CM, et al. Nasal airway transcriptome-wide association study of asthma reveals genetically driven mucus pathobiology. Nat Commun 2022 131. 2022;13: 1–17. doi:10.1038/s41467-022-28973-7
13. Seibold MA, Wise AL, Speer MC, Steele MP, Brown KK, Loyd JE, et al. A Common MUC5B Promoter Polymorphism and Pulmonary Fibrosis. N Engl J Med. 2011;364: 1503–1512.
14. Donoghue LJ, Livraghi-Butrico A, McFadden KM, Thomas JM, Chen G, Grubb BR, et al. Identification of trans Protein QTL for Secreted Airway Mucins in Mice and a Causal Role for Bpifb1. Genetics. 2017;207: 801–812. doi:10.1534/genetics.117.300211/-/DC1.1
15. Kasela S, Ortega VE, Martorella M, Garudadri S, Nguyen J, Ampleford E, et al. Genetic and non-genetic factors affecting the expression of COVID-19-relevant genes in the large airway epithelium. Genome Med. 2021;13: 1–17. doi:10.1186/S13073-021-00866-2

513 16. Polineni D, Dang H, Gallins PJ, Jones LC, Pace RG, Stonebraker JR, et al. Airway
514 mucosal host defense is key to genomic regulation of cystic fibrosis lung disease severity.
515 Am J Respir Crit Care Med. 2018;197: 79–93. doi:10.1164/rccm.201701-0134OC
516 17. Aguet F, Barbeira AN, Bonazzola R, Brown A, Castel SE, Jo B, et al. The GTEx
517 Consortium atlas of genetic regulatory effects across human tissues. Science (80- ).
518 2020;369: 1318–1330. doi:10.1126/SCIENCE.AAZ1776
519 18. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation,
520 and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids
521 Res. 2012;40: D930–D934. doi:10.1093/NAR/GKR917
522 19. Tsao P-N, Vasconcelos M, Izvolsky KI, Qian J, Lu J, Cardoso W V. Notch signaling
523 controls the balance of ciliated and secretory cell fates in developing airways.
524 Development. 2009;136: 2297–2307. doi:10.1242/dev.034884
525 20. Baron RM, Kenny DA. The moderator-mediator variable distinction in social
526 psychological research: Conceptual, strategic, and statistical considerations. J Pers Soc
527 Psychol. 1986;51: 1173–1182. doi:10.1037/0022-3514.51.6.1173
528 21. Zhong W, Spracklen CN, Mohlke KL, Zheng X, Fine J, Li Y. Multi-SNP mediation
529 intersection-union test. Bioinformatics. 2019;35: 4724–4729.
530 doi:10.1093/bioinformatics/btz285
531 22. Han MLK, Quibrera PM, Carretta EE, Barr RG, Bleecker ER, Bowler RP, et al.
532 Frequency of exacerbations in patients with chronic obstructive pulmonary disease: an
533 analysis of the SPIROMICS cohort. Lancet Respir Med. 2017;5: 619–626.
534 doi:10.1016/S2213-2600(17)30207-2
535 23. Chick JM, Munger SC, Simecek P, Huttlin EL, Choi K, Gatti DM, et al. Defining the
536 consequences of genetic variation on a proteome-wide scale. Nature. 2016;534: 500–505.
537 doi:10.1038/nature18270
538 24. Schafer S, Adami E, Heinig M, Rodrigues KEC, Kreuchwig F, Silhavy J, et al.
539 Translational regulation shapes the molecular landscape of complex disease phenotypes.
540 Nat Commun. 2015;6: 7200. doi:10.1038/ncomms8200
541 25. Wan ES. Examining genetic susceptibility in acute exacerbations of COPD. Thorax.
542 2018;73: 507–509. doi:10.1136/thoraxjnl-2017-211106
543 26. Keene JD, Jacobson S, Kechris K, Kinney GL, Foreman MG, Doerschuk CM, et al.
544 Biomarkers Predictive of Exacerbations in the SPIROMICS and COPDGene Cohorts. Am
545 J Respir Crit Care Med. 2017;195: 473–481. doi:10.1164/RCCM.201607-1330OC
546 27. Packer R, Shrine N, Hall R, Melbourne C, Thompson R, Williams AT, Paynton ML,
547 Davitte J, Hessel E, Michalovich D, Betts JC, Sayers I, Yeo A, Hall IP, Tobin MD WL.
548 Genome-wide association study of chronic sputum production implicates loci involved in
549 mucus production and infection. medRxiv. 2022. doi:10.1101/2022.01.11.22269075
550 28. Ash SY, Harmouche R, Putman RK, Ross JC, Martinez FJ, Choi AM, et al. Association
551 between acute respiratory disease events and the MUC5B promoter polymorphism in
552 smokers. Thorax. 2018;73: 1071–1074. doi:10.1136/thoraxjnl-2017-211208
553 29. Dijkstra AE, Smolonska J, van den Berge M, Wijmenga C, Zanen P, Luinge MA, et al.
554 Susceptibility to Chronic Mucus Hypersecretion, a Genome Wide Association Study.
555 PLoS One. 2014;9: e91621. doi:10.1371/journal.pone.0091621
556 30. Dijkstra AE, Boezen HM, Van Den Berge M, Vonk JM, Hiemstra PS, Barr RG, et al.
557 Dissecting the genetics of chronic mucus hypersecretion in smokers with and without
558 COPD. Eur Respir J. 2015;45: 60–75. doi:10.1183/09031936.00093314

559 31. Couper D, LaVange LM, Han M, Barr RG, Bleecker E, Hoffman EA, et al. Design of the
560 Subpopulations and Intermediate Outcomes in COPD Study (SPIROMICS). Thorax.
561 2014;69: 491–4. doi:10.1136/thoraxjnl-2013-203897
562 32. Sun W, Kechris K, Jacobson S, Drummond MB, Hawkins GA, Yang J, et al. Common
563 Genetic Polymorphisms Influence Blood Biomarker Measurements in COPD. PLoS
564 Genet. 2016;12: 1–33. doi:10.1371/journal.pgen.1006011
565 33. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation
566 genotype imputation service and methods. Nat Genet. 2016;48: 1284–1287.
567 doi:10.1038/ng.3656
568 34. The 1000 Genomes Project Consortium. A global reference for human genetic variation.
569 Nature. 2015;526: 68–74. doi:10.1038/nature15393
570 35. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, et al. Genetic
571 epidemiology of COPD (COPDGene) study design. COPD J Chronic Obstr Pulm Dis.
572 2010;7: 32–43. doi:10.3109/15412550903499522
573 36. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of
574 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590: 290–299.
575 doi:10.1038/s41586-021-03205-y
576 37. Pan UKBB. [cited 17 Jun 2022]. Available: https://pan.ukbb.broadinstitute.org/
577 38. Henderson AG, Ehre C, Button B, Abdullah LH, Cai LH, Leigh MW, et al. Cystic fibrosis
578 airway secretions exhibit mucin hyperconcentration and increased osmotic pressure. J Clin
579 Invest. 2014;124: 3047–3060. doi:10.1172/JCI73469
580 39. Aulchenko YS, Struchalin M V., van Duijn CM. ProbABEL package for genome-wide
581 association analysis of imputed data. BMC Bioinformatics. 2010;11: 134.
582 doi:10.1186/1471-2105-11-134
583 40. Price A, Patterson N, Plenge R, Weinblatt M, Shadick N, Reich D. Principal components
584 analysis corrects for stratification in genome-wide association studies. Nat Genet.
585 2006;38: 904–909. doi:10.1038/ng1847
586 41. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of
587 expression residuals (PEER) to obtain increased power and interpretability of gene
588 expression analyses. Nat Protoc. 2012;7: 500–507. doi:10.1038/nprot.2011.457
589 42. Taylor-Weiner A, Aguet F, Haradhvala NJ, Gosai S, Anand S, Kim J, et al. Scaling
590 computational genomics to millions of individuals with GPUs. Genome Biol. 2019;20: 1–
591 5. doi:10.1186/S13059-019-1836-7
592

593 **Table 1. Descriptive Statistics of SPIROMICS Participants in Mucin GWAS***

| | Non-smoking Controls | | At-risk | | GOLD 1 | | GOLD 2 | | GOLD 3 | |
|---|---|---|---|---|---|---|---|---|---|---|
| Study Population | Total Mucin | MUC5AC, MUC5B | Total Mucin | MUC5AC, MUC5B | Total Mucin | MUC5AC, MUC5B | Total Mucin | MUC5AC, MUC5B | Total Mucin | MUC5AC, MUC5B |
| N | 50 | 25 | 219 | 46 | 130 | 40 | 241 | 56 | 67 | 48 |
| Age, mean (range) | 57.9 (40-80) | 61.3 (42-80) | 60.4 (40-79) | 63.3 (40-79) | 66.7 (45-80) | 65.8 (49-79) | 65.0 (42-80) | 64.9 (48-80) | 66.5 (48-80) | 67.1 (52-80) |
| Males, n (%) | 26 (51.0) | 13 (52.0) | 117 (53.2) | 21 (45.7) | 91 (70.0) | 31 (77.5) | 142 (58.9) | 35 (62.5) | 35 (52.2) | 21 (43.8) |
| European Ancestry, n (%) | 36 (72.0) | 25 (100) | 155 (70.8) | 46 (100) | 112 (86.2) | 40 (100) | 214 (88.8) | 56 (100) | 59 (88.1) | 48 (100) |
| African Ancestry, n (%) | 14 (28.0) | NA† | 64 (29.2) | NA‡ | 18 (13.9) | NA‡ | 27 (11.2) | NA‡ | 8 (11.9) | NA‡ |
| Chronic Bronchitis†, n (%) | 3 (6.0) | 2 (8.0) | 89 (40.6) | 15 (32.6) | 52 (40.0) | 19 (47.5) | 134 (55.4) | 35 (62.5) | 29 (43.2) | 20 (41.7) |
| Current smoker, n (%) | 0 (0) | 0 (0) | 113 (51.6) | 21 (45.7) | 45 (34.6) | 17 (42.5) | 114 (47.1) | 26 (46.4) | 18 (26.9) | 11 (22.9) |
| Former smoker, n (%) | 0 (0) | 0 | 106 (48.4) | 25 (54.4) | 85 (65.4) | 23 (57.5) | 127 (52.5) | 30 (53.6) | 49 (73.3) | 37 (77.1) |
| Smoking, pack-years (range) | 0 | 0 (0) | 43.0 (20-150) | 44.0 (20-100) | 51.9 (20-160) | 54.7 (20-117) | 57.1 (20-270) | 55.8 (20-147) | 51.8 (20-126) | 50.3 (21-126) |

594
595 *One subject with a GOLD stage of 4 was included in the total mucin analysis but is not shown here.
596 † Based on St. George's Respiratory Questionnaire
597 ‡Due to limited sample size of SPIROMICS participants of African ancestry with MUC5AC/MUC5B data, only data on European
598 ancestry subjects were used in these analyses.

599 **Table 2. Logistic Regression Model of Chronic Bronchitis at Baseline and rs140324259**
600 **genotype (n=1257)**

| Parameter | Odds Ratio | 95% Confidence Interval |
|---|---|---|
| Age | 0.99 | 0.98-1.00 |
| Sex (M vs. F) | 1.40 | 1.09-1.80 |
| PC1 | 1.01 | 0.87-1.20 |
| PC2 | 1.10 | 0.95-1.30 |
| Smoking pack years | 1.00 | 1.00-1.00 |
| Current smoker | 5.55 | 4.17-7.40 |
| FEV1, % predicted | 0.99 | 0.99-1.00 |
| **rs140324259 (C vs. T)** | 1.42 | 1.10-1.80 |

601
602
603
604

605 **Table 3. Logistic Regression Model Comparing Exacerbators Versus Non-Exacerbators**
606 **Based on Prospectively Ascertained Exacerbation Count Over a Three-Year Period**
607 **(n=822)**

| **Parameter** | **Odds Ratio** | **95% Confidence Interval** |
|---|---|---|
| Age | 1.00 | 0.97-1.20 |
| Sex (M vs F) | 0.65 | 0.47-0.89 |
| PC1 | 0.95 | 0.78-1.15 |
| PC2 | 1.11 | 0.90-1.37 |
| Smoking, pack years | 1.01 | 1.00-1.01 |
| Current smoker | 1.22 | 0.86-1.74 |
| Exacerbations, year prior to enrollment | 1.96 | 1.54-2.49 |
| FEV1, % predicted | 0.97 | 0.96-0.98 |
| **rs140324259 (C vs. T)** | 1.41 | 1.02-1.94 |

608

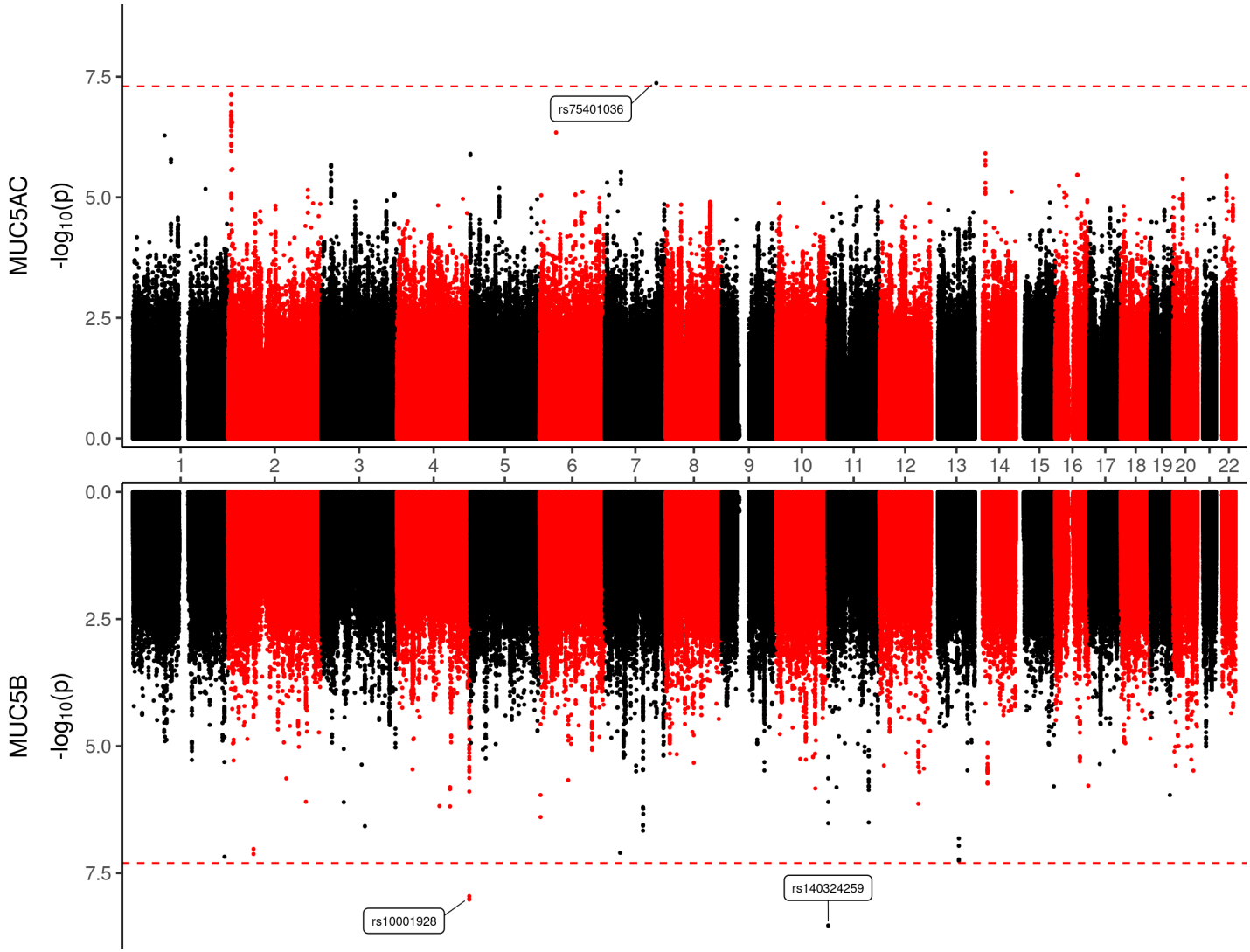
609 **Table 4. Association analysis results for lead MUC5B pQTL variant with CB-related phenotypes and smoking history in the**
610 **UK Biobank***

| UK Biobank Phenotype | # cases/controls | rs140324259 C Allele Frequency Cases | rs140324259 C Allele Frequency Controls | P-value |
|---|---|---|---|---|
| Bring up phlegm/sputum/mucus on most days | 9,169 / 97,072 | 0.150 | 0.143 | 2.37E-05 |
| Cough on most days | 14,606 / 91,635 | 0.149 | 0.143 | 3.94E-06 |
| Current Smoking Status | 43,192 / 375,625 | 0.141 | 0.143 | 4.33E-03 |
| Ever Smoked | 253,507 / 165,353 | 0.143 | 0.143 | 9.39E-01 |
| Smokes tobacco on all or most days | 2,447 / 103,281 | 0.137 | 0.144 | 2.27E-02 |

611
612 *analysis limited to subjects of European Ancestry.
613

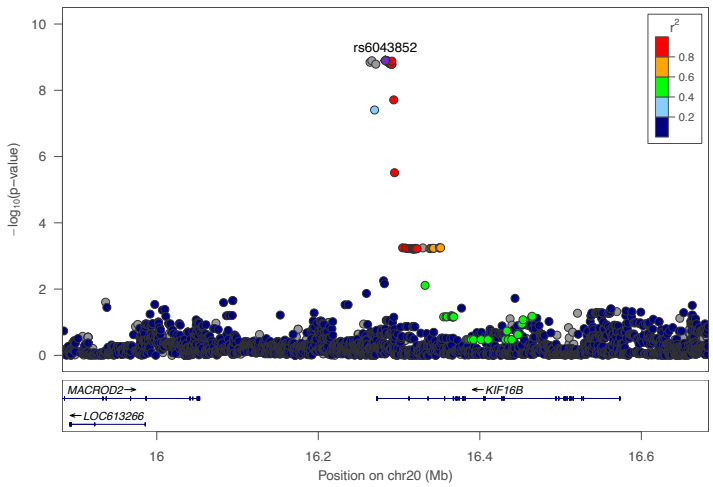
614 **Figure Legends**
615
616 **Figure 1. Distal and local pQTL for sputum MUC5AC and MUC5B.** Results of association
617 analysis using sputum mucin concentration data from 215 EA SPIROMICS participants are
618 shown. Dashed red line denotes genome-wide significance threshold.
619
620 **Figure 2. A Genotype x Smoking interaction locus (rs6043852) for sputum MUC5B on**
621 **Chromosome 20**. **A**. Locus zoom plot for the genotype x smoking locus (rs6043852). **B**.
622 MUC5B concentration as a function of both rs6043852 genotype and current smoking status.
623 The not current smoker category includes never smokers and former smokers. Note that while
624 we plot carriers of the minor allele here as one group, the regression model for MUC5B used
625 genotype dosages.
626
627 **Figure 3. The Chromosome 11 MUC5B pQTL. A**. Regional view of association test results.
628 Four genes were omitted due to small size. The lead variant, rs140324259, is approximately 100
629 kb upstream of *MUC5B*. **B.** Effect of rs140324259 genotype on sputum MUC5B. Numbers in
630 parentheses on x-axis denote sample size per genotype. Each C allele yields a 0.8 log (ln) unit
631 decreased in MUC5B, corresponding to 2.3 picomol/ml drop in MUC5B concentration.
632
633 **Figure 4. Mediation analysis reveals that rs140324259 exerts effects on exacerbations in the**
634 **year prior to enrollment through direct and indirect paths with contrasting allele effects.**
635 We leveraged the mediation analysis framework of Baron and Kenny [20] to examine whether
636 rs140324259 exerts effects on exacerbations through MUC5B. Using complete data on 142
637 subjects, in **(A)** we tested for the total effect of rs140324259 on acute exacerbations of COPD
638 ("c")**.** In (**B)**, the mediation analysis framework is shown in which the effect of rs140324259 on
639 acute exacerbations is modeled as the sum of direct (rs140324259 to exacerbations, c') and
640 indirect paths (rs140324259 to exacerbations via MUC5B (a, b)). Statistical evidence of the
641 indirect path assessed by jointly testing that both rs140324259 → MUC5B (a) and MUC5B →
642 exacerbations (b) are significant using an intersection union test (which is equivalent to testing
643 that $\gamma$ x $\beta_1$ is not equal to 0). $\beta_1$ (b) and $\beta_2$ (c') come from the same negative binomial regression
644 model including both rs140324259 and MUC5B as predictors of exacerbations. Note that in this
645 mediation analysis framework, the total effect (c) in part A is the sum of the direct (c') and
646 indirect paths (a→b) in part B, i.e., c = c' + (a x b). Thus, because the sign of path a is negative
647 while both b and c' are positive, the total effect c (in panel A) is necessarily weaker in
648 magnitude.
649
650

# Figure 1

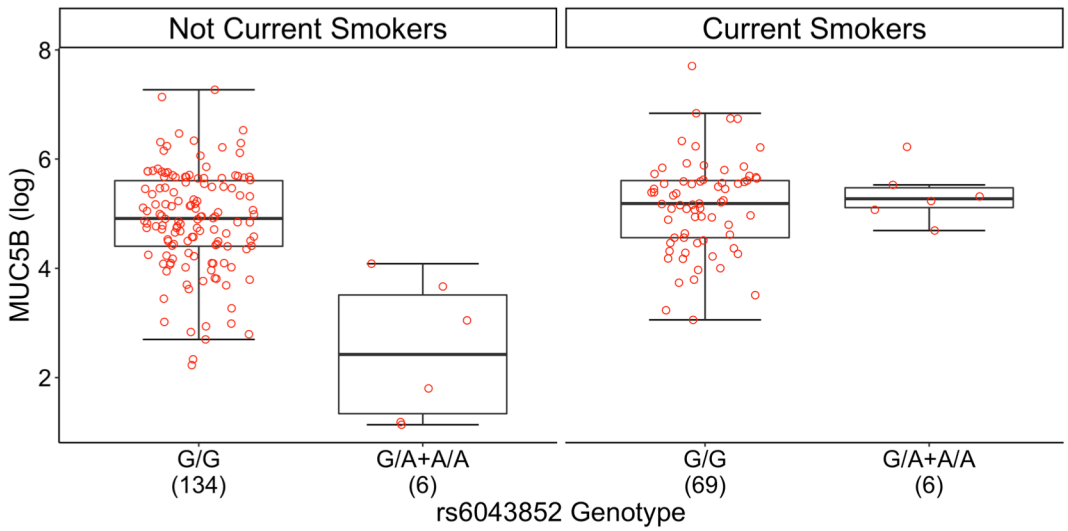

MUC5AC
-log10(p)
7.5
5.0
2.5
0.0
rs75401036
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
22
MUC5B
-log10(p)
0.0
2.5
5.0
7.5
rs10001928
rs140324259

# Figure 2

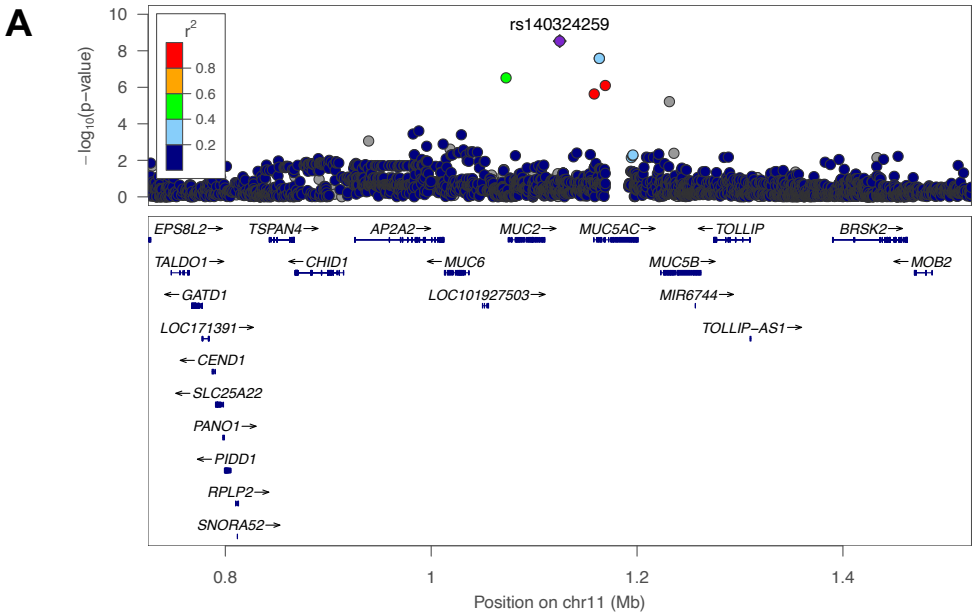
A


rs6043852
$r^2$
0.8
0.6
0.4
0.2
$-\log_{10}$(p-value)
10
8
6
4
2
0
MACROD2
KIF16B
LOC613266
16
16.2
16.4
16.6
Position on chr20 (Mb)

B


Not Current Smokers
Current Smokers
MUC5B (log)
8
6
4
2
G/G
(134)
G/A+A/A
(6)
G/G
(69)
G/A+A/A
(6)
rs6043852 Genotype

# Figure 3


A
rs140324259
r2
0.8
0.6
0.4
0.2
−log10(p−value)
10
8
6
4
2
0
EPS8L2
TSPAN4
AP2A2
MUC2
MUC5AC
TOLLIP
BRSK2
TALDO1
CHID1
MUC6
MUC5B
MOB2
GATD1
LOC101927503
MIR6744
LOC171391
TOLLIP−AS1
CEND1
SLC25A22
PANO1
PIDD1
RPLP2
SNORA52
0.8
1
1.2
1.4
Position on chr11 (Mb)

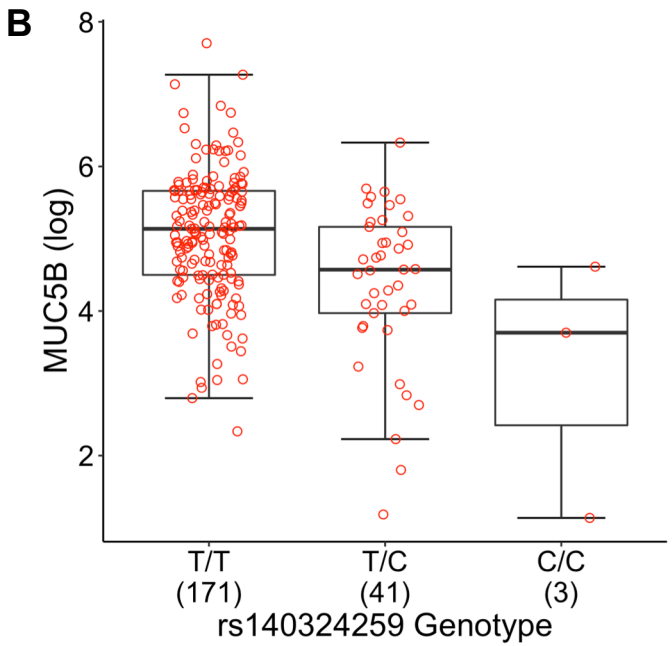
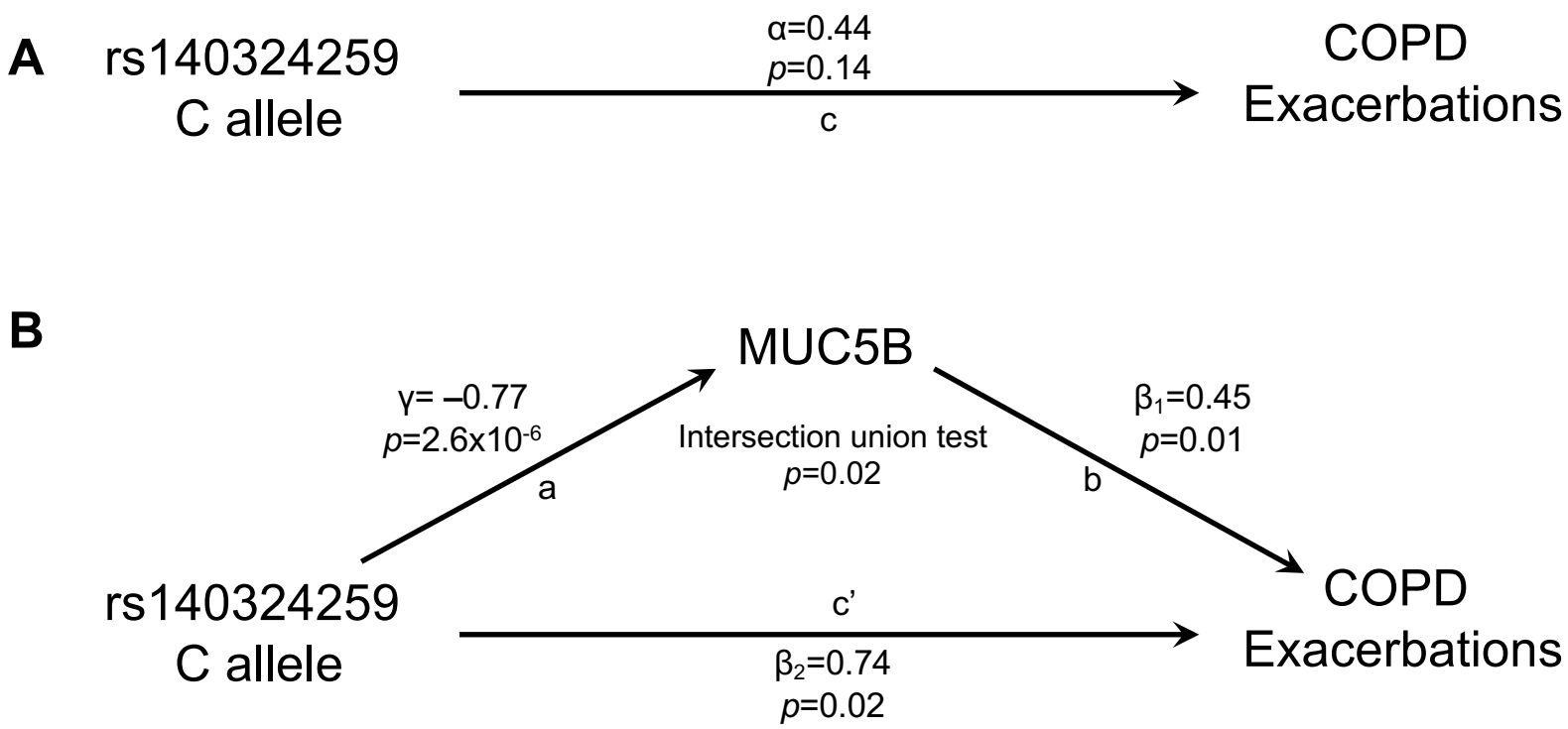

B
MUC5B (log)
8
6
4
2
T/T
(171)
T/C
(41)
C/C
(3)
rs140324259 Genotype

**Figure 4**


A
rs140324259
C allele
α=0.44
p=0.14
c
COPD
Exacerbations
B
MUC5B
γ= –0.77
p=2.6x10-6
a
Intersection union test
p=0.02
β1=0.45
p=0.01
b
rs140324259
C allele
c'
β2=0.74
p=0.02
COPD
Exacerbations