A MUC5B gene polymorphism, rs3570950-T, confers protective effects in COVID-19 infection.


NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
USA, 45 VA Informatics and Computing Infrastructure (VINCI), VA Salt Lake City Healthcare System, Salt Lake City, UT, 46 Internal Medicine, Epidemiology, University of Utah School of Medicine, Salt Lake City, UT, 47 Case Western Reserve University, Cleveland, OH, 48 Vanderbilt University Medical Center, Nashville, TN, 49 Department of Psychiatry, Division of Human Genetics, Yale School of Medicine, New Haven, CT, 06511, USA, 50 VA Connecticut Healthcare System, West Haven, CT, 06516, USA, 51 Medicine, University of California, Los Angeles, Los Angeles, CA, 52 Infectious Disease Section, Louis Stokes Cleveland VA and Case Western Reserve University, Cleveland, OH, 44106, USA, 53 Data Science and Learning, Argonne National Laboratory, 9700 S Cass Ave, Lemont, IL, 60439, 54 Departments of Medicine, Biomedical Informatics, and Pharmacology, Vanderbilt University Medical Center, Nashville, TN, 37232, 55 USA, 56 Corporal Michael J Crescenz VA Medical Center, 57 Medicine, Rheumatology, VA Boston Healthcare System, Boston, MA, 02130, 58 Precision Medicine, VA Palo Alto Health Care System, 3801 Miranda Avenue, Palo Alto, CA, 59 94304, USA, 60 Emory University School of Medicine, Atlanta, GA, 30322, USA, 61 (No affiliation data provided), 62 Medicine, Cardiology, VA Boston Healthcare System, 1400 VFW Parkway, Boston, MA, 2132, USA, 63 VA Boston Healthcare System, Boston, MA, 64 Center of Excellence for Stress & Mental Health, VA San Diego Healthcare System, San Diego, CA, 92161, USA, 65 Center for Behavioral Genetics of Aging, University of California San Diego, La Jolla, CA, 92093, USA, 66 Case Western Reserve University, Cleveland, OH, 44106, USA, 67 Louis Stokes Cleveland VA Medical Center, Cleveland, OH, 44106, USA, 68 VA Portland Health Care System, Portland, OR, 97239, USA

Joint Authorship

† These authors contributed equally to this work
* These authors jointly supervised this work

Supported by MVP035 award from Million Veteran Program, Office of Research and Development, Veterans Health Administration, and Veteran Affairs of the United States Government BX 004831 (P.W/K.C). This publication does not represent the views of the Department of Veteran Affairs of the United States Government.

Author contributions: A.V, J.E.H, S.I, S.L, L.G and J.M and E.S.W. analyzed data. J.E.H., A.V., L.G., E.S.W., S.I., S.L. supervised data collection. A.V., J.E.H, L.G., J.M., E.S.W., S.I. and S.L. wrote the manuscript. A.V., J.E.H, L.G. J.M., E.S.W., S.I. and S.L. conceived the study design, supervised data collection and analysis, and wrote and edited the manuscript. All authors revised and approved the final version to be published.

Correspondence should be addressed to Shiu-Wen Luoh (Shiu-Wen.Luoh@va.gov), Sudha K Iyengar (ski@case.edu)

Manuscript word count: 2,896
Rationale: A common MUC5B gene polymorphism, rs35705950-T, is associated with idiopathic pulmonary fibrosis, but its role in the SARS-CoV-2 infection and disease severity is unclear.

Objectives: To assess whether rs35705950-T confers differential risk for clinical outcomes associated with COVID-19 infection among participants in the Million Veteran Program (MVP) and COVID-19 Host Genetics Initiative (HGI).

Methods: MVP participants were examined for an association between the incidence or severity of COVID-19 and the presence of a MUC5B rs35705950-T allele. Comorbidities and clinical events were extracted from the electronic health records (EHR). The analysis was performed within each ancestry group in the MVP, adjusting for sex, age, age$^2$, and first twenty principal components followed by a trans-ethnic meta-analysis. We then pursued replication and performed a meta-analysis with the trans-ethnic summary statistics from the HGI. A phenome-wide association study (PheWAS) of the rs35705950-T was conducted to explore associated pathophysiologic conditions.

Measurements and Main Results: A COVID-19 severity scale was modified from the World Health Organization criteria, and phenotypes derived from the International Classification of Disease-9/10 were extracted from EHR. Presence of rs35705950-T was associated with fewer hospitalizations ($N_{cases}=25353$, $N_{controls}=631,024$; OR=0.86 [0.80-0.93], $p=7.4 \times 10^{-5}$) in trans-ethnic meta-analysis within MVP and joint meta-analyses with the HGI ($N=1641311$; OR=0.89 [0.85-0.93], $p = 1.9 \times 10^{-6}$). Moreover, individuals of European Ancestry with at least one copy of rs35705950-T had fewer post-COVID-19 pneumonia events (OR=0.85 [0.76-0.96], $p =0.008$). PheWAS exclusively revealed pulmonary involvement.

Conclusions: The MUC5B variant rs35705950-T is protective in COVID-19 infection.

Keywords: coronavirus disease 2019; severe acute respiratory syndrome coronavirus 2; idiopathic pulmonary fibrosis; electronic health records; genetic association
Introduction

A respiratory disease caused by a novel coronavirus was first reported towards the end of 2019, now known as SARS-CoV-2 (COVID-19). Despite massive public health measures and vaccination initiatives, the COVID-19 pandemic remains a major global health threat. By September 2, 2021, the coronavirus disease-2019 (COVID-19) pandemic had caused more than 219 million confirmed infections and more than 4.5 million deaths worldwide(1).

Parenchymal fibrosis is a late complication of respiratory infections with COVID-19(2–4). Among chronic lung diseases, idiopathic pulmonary fibrosis (IPF)(5), a disorder characterized by progressive pulmonary scarring which is associated with a median survival of 2-3 years in the absence of lung transplantation(6), shares several risk factors with those for severe COVID-19 disease, including advanced age(7), cardiovascular disease, diabetes, and history of smoking(5). Thus, common pathological processes may be shared between the fibrotic response towards COVID-19 infection and those underlying IPF.

IPF likely develops from a multifaceted interaction between genetic and environmental factors, age-related mechanisms, and epigenetic profibrotic reprogramming(8, 9). One of the most robust genetic risk factors identified for IPF susceptibility is rs35705950-T, a common G to T transversion located approximately 3 kb upstream of the mucin 5B, oligomeric mucus/gel-forming MUC5B gene (10, 11). Laboratory evidence supports that rs35705950-T is: 1) a functional variant located within an enhancer subject to epigenetic programming and 2) contributes to pathologic mis-expression in IPF (12).

Given the high minor allele frequency (MAF) of rs35705950-T (~20% among individuals of European ancestry) and possible shared pathophysiological pathways between IPF and severe COVID-19 disease,
we examined the association between rs35705950-T and the clinical outcomes of COVID-19 infection in 143 the Million Veteran Program (MVP), a multi-ethnic cohort of over 650,000 U.S. Veterans with detailed 144 EHR and genotyping data(13). Following our primary analysis in the MVP, we validated our results with 145 a comparable analysis conducted in the Host Genetics Initiative (HGI), a global collaboration of over 160 146 genetic studies assembled to facilitate rapid discovery and dissemination of COVID-19 related science 147 (14).

Methods

Data Sources

Data from the MVP, a multi-ethnic genetic biobank sponsored by the United States Veterans Affairs 152 (VA), were analyzed (13). All protocols were approved by the VA Central Institutional Review Board and 153 all participants provided written informed consent. For detailed Materials and Methods, please see 155 methods in the online data supplement.

156

Demographic and pre-existing comorbidity data were collected from questionnaires and the VA EHR; 157 “pre-COVID” data was from the time of enrollment into the MVP to September 30, 2019. The cohort 158 demographics and a description of the clinical conditions for all tested patients in the two years 159 preceding the index dates are presented in a supplemental table (Table E1).

160

161 Genotyping was performed using a custom Thermo Fisher Axiom genotyping platform (MVP 1.0) which 162 included direct genotyping of rs35705950-T. Ancestry was defined using Harmonized ancestry, race, and 163 ethnicity (HARE) derived from self-report and genetic ancestry data(15).

164 COVID-19 outcome definitions
COVID-19 infection status from 02/2020 - 04/2021 was assessed by either self-report (if testing was performed outside the VA) or by a positive polymerase chain reaction (PCR)-based test (16, 17). For subgroup analyses of severity, only patients with confirmed PCR-based tests were examined. The index date was defined as a COVID-19 diagnosis date, i.e., specimen date, or a self-reported date of diagnosis; and for a hospitalized patient, the admission date up to 15 days prior to the COVID-19 case date.

Our analyses used harmonized definitions with the HGI to enable us to obtain larger sample sizes and consistent results. In accordance with the HGI definitions, the three following analyses were performed:

1. COVID Susceptibility: COVID-19 positive vs. population controls;
2. COVID Hospitalization-v1: COVID-19 positive and hospitalized vs. population controls;

Our other analyses focused on data present in MVP only and addressed the outcome severity and post-index events. For these sets of analyses, we only focused on patients who received their PCR-based COVID-19 testing within VA systems. COVID-19 severity scale was derived from the WHO COVID-19 Disease Progression Scale (18) as mild, moderate (hospitalization), severe (Intensive Care Unit-level care), or death within 30 days of PCR-confirmed COVID-19 infection. All data and variables were assessed centrally by the MVP data core's Shared Data Repository (SDR).

Post-index analytic constructs and study design

The ICD codes used to pull the pneumonia events within 60 days post-index (pneumonia60d) are presented in Table E2. Pre-index conditions were derived using natural language processing (NLP)-boosted unstructured notes, ICD and Current Procedural Terminology (CPT) codes, and medications are
taken 2 years prior. Post-index conditions, including pneumonia, were derived using ICD and CPT codes, and medications 60 days after the index date. Association with post-index pneumonia events (pneumonia60d) were performed among patients who received confirmatory COVID-19 PCR testing at VA sites.

**Statistical analysis**

Firth logistic regression\(^{(19, 20)}\) as implemented in the R (v3.6.1) package “brglm2” (version 0.7.1)\(^{(21)}\) was used to examine the association between COVID-19 outcomes and rs35705950-T (additive model) separately by ancestry, with adjustment for age, age\(^2\), sex, and ethnicity-specific principal components. Trans-ethnic meta-analyses were performed using random-effects models in “metafor” (version 2.4-0)\(^{(22)}\). Interactions between COVID-19 infection status and rs35705950-T on the outcome of post-index pneumonia at 60 days were assessed using a multiplicative interaction term followed by stratified analyses by COVID-19 infection status, with additional covariate adjustment of pre-index pneumonia.

**Phenome-wide and Laboratory-wide association studies (PheWAS and LabWAS)**

Associations between rs35705950-T allele and pre-existing comorbid conditions and laboratory values were examined using preclinical data prior to the COVID-19 era (Sept 2019). Individuals with ≥2 Phecodes\(^{(23)}\) were defined as cases. Phecodes with <200 cases within each ancestry group were excluded, resulting in 1618 (EUR), 1289 (AFR), 994 (HIS) Phecodes. LabWAS was conducted for 69 clinical tests; for individuals with repeated measures, the median of the individuals’ EHR record was used. Logistic/Firth regression and linear regression were used for Phecodes and laboratory measurements, respectively. Bonferroni-adjusted thresholds for significance (by ancestry) were: EUR = 3.09 x 10\(^{-5}\) (0.05/1618), AFR = 3.8 x 10\(^{-5}\) (0.05/1289), HIS = 5.03 x 10\(^{-5}\) (0.05/994). Analyses were performed using PLINK2\(^{(24)}\) (Additional details in supplemental methods).
Meta-analysis with HGI

Data from Release 5 (01/18/2021) of the COVID-19 Host Genetics Initiative (HGI) were utilized for replication via an inverse-variance weighted meta-analysis using plink2a (24) and GWAMA (25) (Additional details in supplemental methods).

Results

Elucidation of the shared genetics with the MUC5B rs35705950-T by PheWAS and LabWAS

In order to understand the pathophysiology associated with the MUC5B rs35705950-T allele, and more specifically how the presence of the MUC5B rs35705950-T allele(s) might impact the susceptibility and severity of COVID-19, we performed PheWAS and LabWAS to search for the MUC5B rs35705950-T allele associated conditions prior to COVID-19 infection. The sample sizes for MVP participants used for PheWAS and COVID-19 association studies, as well as HGI participants examined in this study, are shown in Table 1 (Figure E1). The results of the PheWAS are shown in Figure 1 and Table E3.

In the PheWAS analysis between this MUC5B variant and 1605 phenotypes (cases > 200) from participants of European ancestry, we found significant associations (P_{bonferroni} < 2.5 \times 10^{-6}) with 12 respiratory conditions. Consistent with the previous finding in IPF, rs35705950-T was associated with increased risk of Idiopathic fibrosing alveolitis (phecode = 504.1; OR = 2.85 [2.65 - 3.05], P = 8.90 \times 10^{-186}), other alveolar and parietoalveolar pneumonopathy (phecode = 504; OR = 2.64 [2.50 - 2.78], P = 7.07 \times 10^{-289}), and postinflammatory pulmonary fibrosis (phecode = 502; OR = 2.34 [2.23 - 2.45], P = 8.90 \times 10^{-186}). Additionally, we also observed significant associations with respiratory failure (Phecode: 509), ventilatory dependence (Phecode: 509.8), lung transplant (Phecode: 510.2) and pneumonia (Phecode: 480) (Figure 1, Table E3). Notably, we evaluated Phecodes for influenza infection (481) in our PheWAS analysis and did not observe an association with MUC5B rs35705950-T (p<0.05; the power to detect a
difference was >95% as there were 4728 cases of influenza in EUR).

We identified, as in EUR, a significant association of this \textit{MUC5B} variant with an increased risk of three pulmonary conditions in African ancestry participants: idiopathic alveolitis (Phecode: 504.1), other alveolar and parietoalveolar pneumonopathy (Phecode:504), and post-inflammatory fibrosis (Phecode: 502) (Figure 1, Table E3). Two of these associations, other alveolar and parietoalveolar pneumonopathy (Phecode:504) and post-inflammatory fibrosis (Phecode: 502), were also seen in HIS ancestry, suggesting shared etiology.

We performed a Laboratory-wide association study of the \textit{MUC5B} rs35705950-T with median values of clinical laboratory tests measured prior to the COVID-19 pandemic. We only included quantitative traits with 1000 or more individuals. Among EUR participants, we evaluated 63 lab measurements and 10 had a significant association with the rs35705950-T. Increased level of neutrophils (absolute count) had the most significant association (beta=0.05, \( p=6.24 \times 10^{-23} \)). This specific association has not been previously reported. Other significant associations with increased levels were white blood cell counts, neutrophil fraction, estimated glomerular filtration rate (eGFR), eosinophils (absolute count), monocytes (absolute count), and platelets (Figure 2, Table E4). The variant had an association with reduced levels of albumin, lymphocyte fraction, and creatinine (Figure 2, Table E4). There was no significant association with lab measurements in AFR or HIS, but among HIS monocytes (absolute count) were significant (beta =0.0078, \( p=1.66 \times 10^{-04} \)) in the same direction as in EUR.

\textbf{Association of the \textit{MUC5B} rs35705950-T allele with the COVID-19 infection or hospitalization in the MVP and meta-analysis with HGI}
We tested for association between MUC5B rs35705950-T with three COVID-19 phenotype definitions 1) COVID-19 positive as cases vs all the other participants in the MVP as controls 2) COVID-19 positive that required hospitalization for treatment vs all the other participants in the MVP as controls 3) COVID-19 positive that required hospitalization for treatment vs COVID-19 positives that didn’t require hospitalization as controls. First, we performed the analysis in three major ancestries separately (European, African, and Hispanic). Then, we meta-analyzed the summary statistics with the COVID-19 HGI (Freeze 5) using an inverse-variance weighted method (GWAMA)(25). Among the three COVID-19 phenotypes, the most significant association of rs35705950-T allele carriers was with fewer hospitalization events (OR = 0.89 [0.85-0.93], p=1.88 x 10^{-6}, Figure 3 and Table 2). 

Association of the MUC5B rs35705950-T allele with fewer pneumonia events within 60 days of COVID-19 infection in the MVP

In 9,216 COVID-19 infected MVP patients, the adjusted odds ratio for post-index pneumonia was 14.8% less with each additional MUC5B rs35705950-T allele (OR = 0.852 [0.757-0.958], p=0.008). In COVID-19 negative patients, the adjusted odds for post-index pneumonia was 7.8% higher with each additional MUC5B rs35705950-T allele (OR=1.078 [1.001-1.162], p=0.048). This differential effect of an additional MUC5B rs35705950-T allele on post-index pneumonia in COVID-19 positive vs. COVID-19 negative patients was statistically significant (p-value for interaction 0.0009) in EUR (Table 3, Table E5).

Association of the MUC5B rs35705950-T allele with severe outcomes of COVID-19 infection in the MVP

Presence of a MUC5B rs35705950-T allele was not associated with severe outcomes of COVID-19 infection in the MVP. The MUC5B rs35705950-T allele was not associated with severe outcomes with mortality (OR = 1.01 [0.58-1.20], p=0.72) nor mortality alone (OR = 0.91 [ 0.72-1.16], p=0.25) in EUR.
ancestry individuals (Table E6).

Discussion

The data herein establishes that the “T” allele of rs35705950-T in MUC5B, which has been associated with an increased risk for the development of IPF, confers a decreased risk of hospitalization and pneumonia following COVID-19 infection among MVP participants of European ancestry. The protective effect of the rs35705950-T, in addition to being counterintuitive, is in stark contrast to the increased risk of severe COVID-19 disease observed for other well-established causal variants or IPF, including variants located in the TERC, DEPTOR, and FAM13A(26).

The protein product of MUC5B is a major gel-forming mucin in the lung that plays a key role in mucociliary clearance (MCC) and host defense(27). MUC5B protein is secreted from proximal submucosal glands and distal airway secretory cells(28–30). Mucus traps inhaled particles, including bacteria, and transporting them out of the airways by ciliary and cough-driven forces. Mucin also helps remove endogenous debris including dying epithelial cells and leukocytes. MUC5AC and MUC5B are two major secreted forms of mucins in the lung.

The rs35705950-T is located within an enhancer region of MUC5B; the “T” allele demonstrates gain-of-function and is associated with enhanced expression of the MUC5B transcript in lung tissue from unaffected subjects and patients with IPF(31). In patients with IPF, excess MUC5B protein is especially observed in epithelial cells in the respiratory bronchiole and honeycomb cyst(29, 30, 32), regions of the lung involved in lung fibrosis.

Mouse models found that Muc5b is required for mucociliary clearance, for controlling bacterial
infections in the airways and middle ear, and for maintaining immune homeostasis in mouse lungs (33).

Muc5b deficiency caused materials to accumulate in the upper and lower airways. This defect led to chronic infection by multiple bacterial species, including *Staphylococcus aureus*, and to inflammation that failed to resolve normally. Apoptotic macrophages accumulated, phagocytosis was impaired, and interleukin-23 (IL-23) production was reduced in *Muc5b(-/-)* mice. By contrast, in transgenic mice that overexpress *Muc5b*, macrophage functions improved (33). *Muc5B* over-expressing transgenic mice have been shown to be more susceptible to the fibroproliferative effects of bleomycin (34), consistent with a role in IPF. Paradoxically, while the “T” allele of rs35705950-T increases susceptibility towards the development of IPF, the same allele has also been associated with decreased mortality among IPF patients (35).

Our analyses demonstrating a significant interaction between COVID-19 infection and the prospective development of pneumonia suggest a possible mechanism by which the protective effect of rs35705950-T is mediated. Whether enhanced pulmonary macrophage function or quantitative or qualitative changes in mucous production resulting from the minor allele of rs35705950-T are responsible for the observed protective effect should be explored in future work. Of note, the *MUC5B* rs35705950-T allele did not decrease the risk of pneumonia in COVID-19 tested negative participants (Table 3), suggesting that the protective effect may be specific to COVID-19 related pneumonia. More studies in the future are needed to further investigate this phenomenon.

No extrapulmonary association was noted on PheWAS analysis suggesting a very circumscribed molecular and clinical effect of this promoter variant. This supports the notion that the effect of rs35705950-T on COVID-19 infection is mediated in pulmonary tissues. The *Muc5b* over-expression in the distal airway may specifically or non-specifically affect the SARS-CoV-2 viral infection in the lung,
leading to decreased incidence of pneumonia and hospitalization in the infected individuals.

The human MUC5B rs35705950-T allele does not appear to be sufficient to cause pulmonary fibrosis. Although ~20% of the non-Hispanic white populations have a copy of the MUC5B rs35705950-T allele (31, 33), IPF is a rare disease with a population prevalence of less than 0.1% (36). Additional genetic and/or environmental insults are likely needed in the development of IPF in humans. Since the overwhelming number of individuals with the MUC5B rs35705950-T allele will not know their MUC5B status, it is unlikely that the reason for our observation is due to a change in health behaviors of participants that carry this variant.

The MUC5B rs35705950-T allele was associated with elevated neutrophil counts. This could be due in part to the association of this allele with an increased incidence of pneumonia. It is worth noting that neutrophils are a major source of alpha-defensin and elevated alpha-defensin levels were seen in the serum of IPF patients; the levels of alpha-defensin in the serum correlated with the lung function decline in the IPF patients (37, 38).

Longer follow-up of SARS-CoV-2 infected individuals with the MUC5B rs35705950-T allele is needed. One would need to be cautious regarding the longer-term outcome of COVID-19 in the MUC5B rs35705950-T allele positive individuals as a fibrotic response has been reported in the survivors of severe COVID-19. This is of particular importance if the manipulation of MUC5B expression is considered in the prevention/treatment of COVID-19.

The MUC5B rs35705950-T allele variant resides within an enhancer subject to lineage- and disease-dependent epigenetic remodeling. It was postulated that this G to T transversion in the MUC5B


rs35705950-T allele might lead to the removal of a binding site for the GCF transcription repressor (12, 39). A potential avenue for chromatin-based therapies in which MUC5B enhancer chromatin architecture serves as a target to block the MUC5B mis-expression was proposed (12, 39). Additional small molecule and signaling inhibitors targeting IPF are being studied as well (40). These strategies are generally aiming at reducing fibrosis or the effects associated with MUC5B over-expression. How these strategies or alternatives can be utilized to treat/prevent COVID-19 remains to be studied.

In conclusion, we show in this study a common MUC5B promoter variant leading to MUC5B over-expression is associated with fewer hospitalizations and pneumonia events after SARS-CoV-2 infection. Our study provides a strong rationale to stratify patient populations based on common and disease-related genetic polymorphism in order to better understand the mechanisms and their clinical implications in COVID-19. How the MUC5B rs35705950-T allele association may shed light on the pathogenesis and/or management of COVID-19 remains to be fully examined.

Strengths & Limitations

MVP is a large genomic medicine database with diverse ethnicity and geography. MVP participants are predominantly males but it represents a large multi-ethnic, prospective cohort available. Successful replication in the HGI and meta-analysis is a strength as well as our ability to investigate specific clinical events post-index. PheWAS was designed as a broad screen to test for potentially clinically relevant associations between genes and phenotypes and helped in the understanding of potential disease mechanisms but has limited power to detect associations among uncommon conditions, especially when further stratified by genetic ancestry.
Acknowledgments

This research is based on data from the Million Veteran Program, Office of Research and Development, Veterans Health Administration, and was supported by award MVP035. This publication does not represent the views of the Department of Veteran Affairs of the United States Government.

We are grateful to our Veterans for their contribution to MVP. Full acknowledgments for the VA Million Veteran Program COVID-19 Science Initiative can be found in the supplemental methods.

Conflict of Interest

CJO is an employee of Novartis Institute for Biomedical Research. PN reports grant support from Amgen, Apple, AstraZeneca, Boston Scientific, and Novartis, personal fees from Apple, AstraZeneca, Blackstone Life Sciences, Genentech, and Novartis, and spousal employment at Vertex, all unrelated to the present work.

References

1. Website. at <(https://coronavirus.jhu.edu/map.html)>.


12. Gally F, Sasse SK, Kurche JS, Gruca MA, Cardwell JH, Okamoto T, Chu HW, Hou X, Poirion OB,
Buchanan J, Preissl S, Ren B, Colgan SP, Dowell RD, Yang IV, Schwartz DA, Gerber AN. The MUC5B-associated variant rs35705950 resides within an enhancer subject to lineage- and disease-dependent epigenetic remodeling. *JCI Insight* 2021;6:


Table 1. Demographics for COVID-19 tested positive and all MVP participants examined in this study.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Million Veteran Program</th>
<th>COVID-19 Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number (%)</td>
<td>Number (%)</td>
</tr>
<tr>
<td>Total Patients</td>
<td>658,582</td>
<td>13,841</td>
</tr>
<tr>
<td>Male</td>
<td>592,516 (90)</td>
<td>12,320 (89)</td>
</tr>
<tr>
<td>Genetic Ancestry</td>
<td></td>
<td></td>
</tr>
<tr>
<td>European</td>
<td>464,961 (70)</td>
<td>8011 (58)</td>
</tr>
<tr>
<td>African</td>
<td>123,120 (19)</td>
<td>3749 (27)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>521,83 (8)</td>
<td>1903 (14)</td>
</tr>
<tr>
<td>Asian</td>
<td>8329 (1)</td>
<td>178 (1)</td>
</tr>
<tr>
<td>Other</td>
<td>9,989 (2)</td>
<td>0</td>
</tr>
<tr>
<td>Muc5B rs35705950</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 copy</td>
<td>10,604 (1.6)</td>
<td>353 (25)</td>
</tr>
<tr>
<td>1 copy</td>
<td>2161 (0.03)</td>
<td>75 (0.05)</td>
</tr>
<tr>
<td>2 copies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comorbidities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obesity (phecode = 278)</td>
<td>283,197 (43)</td>
<td>8905 (64)</td>
</tr>
<tr>
<td>Hypertension (phecode = 401.1)</td>
<td>451,998 (69)</td>
<td>10617 (77)</td>
</tr>
<tr>
<td>Type 2 Diabetes (phecode = 250.2)</td>
<td>227,757 (34)</td>
<td>10491 (76)</td>
</tr>
<tr>
<td>Coronary Artery Disease (phecode = 411.4)</td>
<td>152,136 (23)</td>
<td>4182 (30)</td>
</tr>
<tr>
<td>Chronic Kidney Disease (phecode = 585.2)</td>
<td>100,46 (15)</td>
<td>533 (38)</td>
</tr>
<tr>
<td>Outcomes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospitalized</td>
<td>-</td>
<td>4491 (32)</td>
</tr>
<tr>
<td>Severe</td>
<td>-</td>
<td>657 (47)</td>
</tr>
<tr>
<td>Deceased</td>
<td>-</td>
<td>644 (46)</td>
</tr>
</tbody>
</table>
Table 2. Association of rs35705950 in MUC5B with (i) COVID-19 Positive vs Population Controls, (ii) COVID-19 Positive, Hospitalized vs Population Controls, and (iii) COVID-19 Positive, Hospitalized vs COVID-19 Positive, not Hospitalized. Odds ratio (OR) and 95% confidence interval (95% CI) is reported for the minor (T) allele, and results are shown for VA Million Veteran Program (MVP) African Americans (AFR), European Americans (EUR), Hispanic/Latino Americans (HIS), and trans-ethnic meta-analysis (ALL), the COVID-19 Host Genetics Initiative (HGI) trans-ethnic release 5 meta-analysis excluding MVP and 23&Me, and the meta-analysis of MVP and HGI (META).

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Population</th>
<th>N Case</th>
<th>N Control</th>
<th>Total N</th>
<th>EAF</th>
<th>OR (95% CI)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive vs Population Control</strong></td>
<td>MVP (AFR)</td>
<td>6,411</td>
<td>114,781</td>
<td>121,192</td>
<td>0.02</td>
<td>0.99 [0.87, 1.12]</td>
<td>0.826</td>
</tr>
<tr>
<td></td>
<td>MVP (EUR)</td>
<td>15,814</td>
<td>443,428</td>
<td>459,242</td>
<td>0.11</td>
<td>0.96 [0.92, 0.99]</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>MVP (HIS)</td>
<td>3,128</td>
<td>47,462</td>
<td>50,590</td>
<td>0.07</td>
<td>0.95 [0.85, 1.05]</td>
<td>0.275</td>
</tr>
<tr>
<td></td>
<td>MVP (ALL)</td>
<td>25,353</td>
<td>605,671</td>
<td>631,024</td>
<td>0.09</td>
<td>0.96 [0.93, 1.00]</td>
<td>0.060</td>
</tr>
<tr>
<td></td>
<td>HGI (ALL)</td>
<td>25,652</td>
<td>1,282,972</td>
<td>1,308,624</td>
<td>0.11</td>
<td>0.98 [0.95, 1.01]</td>
<td>0.134</td>
</tr>
<tr>
<td></td>
<td>META</td>
<td>51,005</td>
<td>1,888,643</td>
<td>1,939,648</td>
<td>0.10</td>
<td>0.97 [0.95, 0.99]</td>
<td>4.57E-03</td>
</tr>
<tr>
<td><strong>Hospitalized vs Population Control</strong></td>
<td>MVP (AFR)</td>
<td>1,739</td>
<td>119,453</td>
<td>121,192</td>
<td>0.02</td>
<td>0.83 [0.64, 1.07]</td>
<td>0.147</td>
</tr>
<tr>
<td></td>
<td>MVP (EUR)</td>
<td>3,325</td>
<td>455,917</td>
<td>459,242</td>
<td>0.11</td>
<td>0.87 [0.80, 0.94]</td>
<td>5.43E-04</td>
</tr>
<tr>
<td></td>
<td>MVP (HIS)</td>
<td>657</td>
<td>49,933</td>
<td>50,590</td>
<td>0.07</td>
<td>0.86 [0.68, 1.07]</td>
<td>0.182</td>
</tr>
<tr>
<td></td>
<td>MVP (ALL)</td>
<td>5,721</td>
<td>625,303</td>
<td>631,024</td>
<td>0.09</td>
<td>0.86 [0.80, 0.93]</td>
<td>7.35E-05</td>
</tr>
<tr>
<td></td>
<td>HGI (ALL)</td>
<td>9,086</td>
<td>1,001,201</td>
<td>1,010,287</td>
<td>0.11</td>
<td>0.91 [0.85, 0.97]</td>
<td>4.12E-03</td>
</tr>
<tr>
<td></td>
<td>META</td>
<td>14,807</td>
<td>1,626,504</td>
<td>1,641,311</td>
<td>0.10</td>
<td>0.89 [0.85, 0.93]</td>
<td>1.88E-06</td>
</tr>
<tr>
<td><strong>Hospitalized vs Not Hospitalized</strong></td>
<td>MVP (AFR)</td>
<td>1,739</td>
<td>4,672</td>
<td>6,411</td>
<td>0.02</td>
<td>0.80 [0.59, 1.08]</td>
<td>0.141</td>
</tr>
<tr>
<td></td>
<td>MVP (EUR)</td>
<td>3,325</td>
<td>12,489</td>
<td>15,814</td>
<td>0.11</td>
<td>0.89 [0.81, 0.97]</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>MVP (HIS)</td>
<td>657</td>
<td>2,471</td>
<td>3,128</td>
<td>0.07</td>
<td>0.88 [0.68, 1.14]</td>
<td>0.319</td>
</tr>
<tr>
<td></td>
<td>MVP (ALL)</td>
<td>5,721</td>
<td>19,632</td>
<td>25,353</td>
<td>0.08</td>
<td>0.88 [0.81, 0.96]</td>
<td>2.64E-03</td>
</tr>
<tr>
<td></td>
<td>HGI (ALL)</td>
<td>4,420</td>
<td>11,093</td>
<td>15,513</td>
<td>0.16</td>
<td>0.97 [0.88, 1.08]</td>
<td>0.575</td>
</tr>
<tr>
<td></td>
<td>META</td>
<td>10,141</td>
<td>30,725</td>
<td>40,866</td>
<td>0.11</td>
<td>0.91 [0.86, 0.98]</td>
<td>7.20E-03</td>
</tr>
</tbody>
</table>
Table 3. Fewer pneumonia events developed within 60 days post COVID-19 infection for MVP EUR individuals with the presence of a MUC5B rs35705950 allele. Odds ratios are estimated from Firth logistic regression adjusting for pre-index pneumonia, age, age², and PC1-20, including an interaction between additive MUC5B rs35705950 allele and COVID-19 infection.

<table>
<thead>
<tr>
<th>MUC5B = 0</th>
<th>MUC5B = 1</th>
<th>MUC5B = 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR (95% CI) of COVID-19 positive status</td>
<td>OR (95% CI) of COVID-19 positive status</td>
</tr>
<tr>
<td>COVID-19 negative</td>
<td>1.08 (1.00, 1.16)</td>
<td>0.89 (0.76, 0.96)</td>
</tr>
<tr>
<td>p=0.048</td>
<td>p=0.008</td>
<td></td>
</tr>
<tr>
<td>10.00 (9.35, 10.70)</td>
<td>7.91 (6.97, 8.97)</td>
<td>6.26 (4.83, 8.08)</td>
</tr>
</tbody>
</table>

Figure 1. Phenome-Wide Association Study (PheWAS) of MUC5B rs35705950 allele in the Million Veteran Program. A PheWAS plot shows associations of rs35705950 and phenotypes derived from the electronic health records data prior to COVID-19 in MVP participants from A) European ancestry B) African ancestry and C) Hispanic ancestry. The phenotypes are shown on the x-axis and organized by disease categories. The p-value (-log10) of each association is shown on the y-axis the direction of the triangle represents the direction of effect of the associations - with the upward triangle as increased risk and the downward triangle as reduced risk. The red line indicates the significance threshold based on the Bonferroni correction. The forest plot of Bonferroni significant associations are shown within the right top corner of each PheWAS plot. The Bonferroni threshold for each ancestry group is shown in the forest plot.

Figure 2. Laboratory-Wide Association Study (PheWAS) of MUC5B rs35705950 allele in the Million Veteran Program. A LabWAS plot shows associations of rs35705950 and median values of laboratory measures extracted from electronic health records data prior to COVID-19 in MVP participants. The bottom panel shows the -log10 (p-value) on the y-axis and laboratory test descriptions on the x-axis. Triangles points up have increasing effects and points down have decreasing effects. The colors represent the different ancestry groups. The top panel shows beta from the regression model for each laboratory measure. The significant results are highlighted in the color corresponding to ancestry groups and other results are plotted in grey.

Figure 3. Forest plot association of rs35705950 in MUC5B with (i) COVID-19 Positive vs Population Controls, (ii) COVID-19 Positive, Hospitalized vs Population Controls, and (iii) COVID-19 Positive,
Hospitalized vs COVID-19 Positive, not Hospitalized. Odds ratio (OR) and 95% confidence interval (95% CI) is reported for the minor (T) allele, and results are shown for VA Million Veteran Program (MVP) African Americans (AFR), European Americans (EUR), Hispanic/Latino Americans (HIS), and trans-ethnic meta-analysis (ALL), the COVID-19 Host Genetics Initiative (HGI) trans-ethnic release 5 meta-analysis excluding MVP and 23&Me, and the meta-analysis of MVP and HGI (META).
Figure X: Forest plot association of rs35705950 in MUC5B with (i) COVID-19 Positive vs Population Controls, (ii) COVID-19 Positive, Hospitalized vs Population Controls, and (iii) COVID-19 Positive, Hospitalized vs COVID-19 Positive, not Hospitalized. Odds ratio (OR) and 95% confidence interval (95% CI) is reported for the minor (T) allele, and results are shown for VA Million Veteran Program (MVP) African Americans (AFR), European Americans (EUR), Hispanic/Latino Americans (HIS), and trans-ethnic meta-analysis (ALL), the COVID-19 Host Genetics Initiative (HGI) trans-ethnic release 4 meta-analysis excluding MVP and 23&Me, and the meta-analysis of MVP and HGI (META).