A *MUC5B* gene polymorphism, rs35705950-T, confers protective effects in COVID-19 infection.

Anurag Verma PhD[1,2†], Jessica Minnier PhD[3,4,5†], Jennifer E Huffman PhD[6†], Emily S Wan MD, MPH[7,8†], Lina Gao PhD[9,10], Jacob Joseph MBBS, MD[11,12], Yuk-Lam Ho MPH[13], Wen-Chih Wu MD, MPH[14,15], Kelly Cho PhD, MPH[16,17], Bryan R Gorman PhD[18], Nallakkandi Rajeevan PhD[19,20], Saiju Pyarajan[21,22], Helene Garcon MD[6], James B Meigs MD, MPH[23], Yan V Sun PhD[24,25], Peter D Reaven MD[26,27], John E McGeary PhD[28,29], Ayako Suzuki MD, PhD, MSc[30,31], Joel Gelernter MD[32,33], Julie A Lynch PhD, RN[34,35], Jeffrey M Petersen MD[36,37], Seyedeh Maryam Zekavat BS[38,39], Pradeep Natarajan MD, MMSc[40,39,41], Cecelia J Madison MBI[42], Sharvari Dalal MD[43,37], Darshana N Jhala MD, B Mus[44,45], Mehrdad Arjomandi MD[46], Elise Gatsby MPH[47], Kristine E Lynch MS PhD[47,48], Robert A Bonomo[49], Mat Freiberg[50], Gita A Pathak PhD[51,52], Jin J Zhou PhD[53,54], Curtis J Donskey MD[55], Ravi K Madduri MS[56], Quinn S Wells MD, PharmD[57], Rose DL Huang MS[6], Renato Polimanti PhD[51,52], Kyong-Mi Chang MD[58], Katherine P Liao MD, MPH[59], Philip S Tsao PhD[60], Peter W.F. Wilson MD[25,61], Adriana Hung[62], Christopher J O'Donnell MD, MPH[63], John M Gaziano[64], Richard L Hauger MD[65,66], Sudha K. Iyengar PhD[67,68*] and Shiuh-Wen Luoh MD PhD[69*]

[1]Corporal Michael Crescenz VA Medical Center, Philadelphia, PA, 19104, USA, [2]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA, [3]OHSU-PSU School of Public Health, Oregon Health & Science University, Portland, OR, 97239, USA, [4]Knight Cancer Institute, Biostatistics Shared Resource, Oregon Health & Science University, Portland, OR, 97239, USA, [5]VA Portland Health Care System, OR, 97239, USA, [6]MAVERIC, VA Boston Healthcare System, Boston, MA, 02130, USA, [7]Department of Medicine, Pulmonary, Critical Care, Sleep, and Allergy Section, VA Boston Healthcare System, Boston, MA, 02115, USA, [8]Channing Division of Network Medicine, Brigham & Women's Hospital, Boston, MA, 02115, USA, [9]Knight Cancer Institute, Biostatistics Shared Resource, Oregon Health & Science University, [10]VA Portland Health Care System, [11]Department of Medicine, VA Boston Healthcare System, Boston, MA 02115, USA, [12]Brigham & Women's Hospital, [13]MAVERIC, VA Boston Healthcare System, Boston, MA, 2130, USA, [14]Department of Medicine, Cardiology, Providence VA Healthcare System, Providence, RI, 02908, USA, [15]Alpert Medical School & School of Public Health, Brown University, Providence, RI, 02903, USA, [16]MAVERIC, VA Boston Healthcare System, 150 S Huntington Ave, Boston, MA, 02130, USA, [17]Medicine, Aging, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, 02130, USA, [18]VA Boston Healthcare System, 150 S Huntington Ave, Boston, MA, 02131, USA, [19]Yale Center for Medical Informatics, Yale School of Medicine, New Haven, CT, 06511', USA, [20]Clinical Epidemiology Research Center (CERC), VA Connecticut Healthcare System, West Haven, CT, 06516', USA, [21]VA Boston Healthcare System, Boston, MA, 2115, [22]Department of Medicine, Harvard Medical School, [23]Medicine, General Internal Medicine, Massachusetts General Hospital, 55 Fruit St, Boston, MA, 02115, USA, [24]Epidemiology, Emory University School of Public Health, 1518 Clifton Rd. NE, Atlanta, GA, 30322, USA, [25]Atlanta VA Health Care System, 1670 Clairmont Road, Decatur, GA, 30033, USA, [26]Deptartment of Medicine, Phoenix VA Healthcare System, Phoenix, AZ, 85012, USA, [27]Univ. of AZ, [28]Department of Psychiatry and Human Behavior, Providence VA Medical Center, Providence, RI, 2908, USA, [29]Brown University Medical School, [30]Department of Medicine, Gastroenterology, Durham VA Medical Center, 508 Fulton St, Durham, NC, 27705, USA, [31]Department of Medicine, Gastroenterology, Duke University, Durham, NC, 27710, USA, [32]Psychiatry, Human Genetics, Yale Univ. School of Medicine, 950 Campbell Avenue, West Haven, CT, 06516, USA, [33]VA CT Healthcare Center, [34]VA Informatics & Computing Infrastructure, VA Salt Lake City Health Care System, Salt Lake City, UT, USA, [35]Department of Medicine, University of Utah School of Medicine, Salt Lake City, UT, [36]Pathology and Laboratory Medicine, Corporal Michael Crescenz VA Medical Center, Philadelphia, PA, 19104, USA, [37]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA, [38]Computational Biology & Bioinformatics, Yale School of Medicine, 333 Cedar St, New Haven, CT, 6510, USA, [39]Program in Medical and Population Genetics, Cardiovascular Disease Initiative, Broad Institute of Harvard and MIT, Cambridge, MA, 02142, USA, [40]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, 02114, USA, [41]Department of Medicine, Harvard Medical School, Boston, MA, 02115, USA, [42]Clinical Data Science Research Group, ORD, Portland VA Medical Center, Portland, OR, 97239, USA, [43]Pathology and Laboratory Medicine, Corporal Michael J Crescenz VA Medical Center, Philadelphia, PA, 19104, USA, [44]Pathology and Laboratory Medicine, Corporal Michael Crescenz VA Medical Center, Philadelphia, PA, 19104, [45]Perelman School of Medicine, University of Pennsylvania, 19104, [46]Medicine, Pulmonary and Critical Care, San Francisco VA Healthcare System; University of California San Francisco, San Francisco, CA, 94121,

USA, [47]VA Informatics and Computing Infrastructure (VINCI), VA Salt Lake City Healthcare System, Salt Lake City, UT, [48]Internal Medicine, Epidemiology, University of Utah School of Medicine, Salt Lake City, UT, [49]Case Western Reserve University, Cleveland, OH, [50]Vanderbilt University Medical Center, Nashville, TN, [51]Department of Psychiatry, Division of Human Genetics, Yale School of Medicine, New Haven, CT, 06511, USA, [52]VA Connecticut Healthcare System, West Haven, CT, 06516, USA, [53]Medicine, University of California, Los Angeles, Los Angeles, CA, 90024, USA, [54]Epidemiology and Biostatistics, University of Arizona, AZ, 85724, USA, [55]Infectious Disease Section, Louis Stokes Cleveland VA and Case Western Reserve University, Cleveland, OH, 44106, USA, [56]Data Science and Learning, Argonne National Laboratory, 9700 S Cass Ave, Lemont, IL, 60439, [57]Departments of Medicine, Biomedical Informatics, and Pharmacology, Vanderbilt University Medical Center, Nashville, TN, 37232, USA, [58]Corporal Michael J Crescenz VA Medical Center, [59]Medicine, Rheumatology, VA Boston Healthcare System, Boston, MA, 02130, [60]Precision Medicine, VA Palo Alto Health Care System, 3801 Miranda Avenue, Palo Alto, CA, 94304, USA, [61]Emory University School of Medicine, Atlanta, GA, 30322, USA, [62](No affiliation data provided), [63]Medicine, Cardiology, VA Boston Healthcare System, 1400 VFW Parkway, Boston, MA, 2132, USA, [64]VA Boston Healthcare System, Boston, MA, [65]Center of Excellence for Stress & Mental Health, VA San Diego Healthcare System, San Diego, CA, 92161, USA, [66]Center for Behavioral Genetics of Aging, University of California San Diego, La Jolla, CA, 92093, USA, [67]Case Western Reserve University, Cleveland, OH, 44106, USA, [68]Louis Stokes Cleveland VA Medical Center, Cleveland, OH, 44106, USA, [69]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 Shiuh-Wen Luoh (Shiuh-Wen.Luoh@va.gov), Sudha K Iyengar (ski@case.edu)

Manuscript word count: **2,896**

Abstract

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 x 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 x 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 the Million Veteran Program (MVP), a multi-ethnic cohort of over 650,000 U.S. Veterans with detailed EHR and genotyping data(13). Following our primary analysis in the MVP, we validated our results with a comparable analysis conducted in the Host Genetics Initiative (HGI), a global collaboration of over 160 genetic studies assembled to facilitate rapid discovery and dissemination of COVID-19 related science (14).

Methods

Data Sources

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

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

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

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; (3) COVID-19 hospitalization-v2: COVID-19 positive and hospitalized vs. COVID-19 positive but not hospitalized.

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 = 3.09×10^{-05} (0.05/1618), AFR = 3.8×10^{-05} (0.05/1289), HIS = 5.03×10^{-05}(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 *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 *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 x 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 x 10^{-04}) in the same direction as in EUR.

Association of the *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

311 infections in the airways and middle ear, and for maintaining immune homeostasis in mouse lungs(33).
312 Muc5b deficiency caused materials to accumulate in the upper and lower airways. This defect led to
313 chronic infection by multiple bacterial species, including *Staphylococcus aureus*, and to inflammation
314 that failed to resolve normally. Apoptotic macrophages accumulated, phagocytosis was impaired, and
315 interleukin-23 (IL-23) production was reduced in *Muc5b(-/-)* mice. By contrast, in transgenic mice that
316 overexpress *Muc5b*, macrophage functions improved (33). *Muc5B* over-expressing transgenic mice have
317 been shown to be more susceptible to the fibroproliferative effects of bleomycin (34), consistent with a
318 role in IPF. Paradoxically, while the "T" allele of rs35705950-T increases susceptibility towards the
319 development of IPF, the same allele has also been associated with *decreased* mortality among IPF
320 patients(35).
321
322 Our analyses demonstrating a significant interaction between COVID-19 infection and the prospective
323 development of pneumonia suggest a possible mechanism by which the protective effect of
324 rs35705950-T is mediated. Whether enhanced pulmonary macrophage function or quantitative or
325 qualitative changes in mucous production resulting from the minor allele of rs35705950-T are
326 responsible for the observed protective effect should be explored in future work. Of note, the *MUC5B*
327 rs35705950-T allele did not decrease the risk of pneumonia in COVID-19 tested negative participants
328 (**Table 3**), suggesting that the protective effect may be specific to COVID-19 related pneumonia. More
329 studies in the future are needed to further investigate this phenomenon.
330
331 No extrapulmonary association was noted on PheWAS analysis suggesting a very circumscribed
332 molecular and clinical effect of this promoter variant. This supports the notion that the effect of
333 rs35705950-T on COVID-19 infection is mediated in pulmonary tissues. The *Muc5b* over-expression in
334 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)>.
2. Zhao Y-M, Shang Y-M, Song W-B, Li Q-Q, Xie H, Xu Q-F, Jia J-L, Li L-M, Mao H-L, Zhou X-M, Luo H, Gao Y-F, Xu A-G. Follow-up study of the pulmonary function and related physiological characteristics of COVID-19 survivors three months after recovery. *EClinicalMedicine* 2020;
3. Yan X, Huang H, Wang C, Jin Z, Zhang Z, He J, Yin S, Fan M, Huang J, Chen F, Zeng Y, Han X, Zhu Y. Follow-up study of pulmonary function among COVID-19 survivors 1 year after recovery. *J Infect* 2021;doi:10.1016/j.jinf.2021.05.034.
4. McGroder CF, Zhang D, Choudhury MA, Salvatore MM, D'Souza BM, Hoffman EA, Wei Y, Baldwin

406 MR, Garcia CK. Pulmonary fibrosis 4 months after COVID-19 is associated with severity of illness
407 and blood leucocyte telomere length. *Thorax* 2021;doi:10.1136/thoraxjnl-2021-217031.
408 5. King CS, Nathan SD. Idiopathic pulmonary fibrosis: effects and optimal management of
409 comorbidities. *Lancet Respir Med* 2017;5:72–84.
410 6. George PM, Patterson CM, Reed AK, Thillai M. Lung transplantation for idiopathic pulmonary
411 fibrosis. *Lancet Respir Med* 2019;7:271–282.
412 7. Ley B, Collard HR, King TE Jr. Clinical course and prediction of survival in idiopathic pulmonary
413 fibrosis. *Am J Respir Crit Care Med* 2011;183:431–440.
414 8. Selman M, Pardo A. Revealing the pathogenic and aging-related mechanisms of the enigmatic
415 idiopathic pulmonary fibrosis. an integral model. *Am J Respir Crit Care Med* 2014;189:1161–1172.
416 9. Selman M, Pardo A. The leading role of epithelial cells in the pathogenesis of idiopathic
417 pulmonary fibrosis. *Cell Signal* 2020;66:109482.
418 10.Zhang Y, Noth I, Gibson KF, Ma S-F, Richards TJ, Bon JM, Lindell KO, Branch RA, Nicolae D,
419 Sciurba F, Garcia AN, Kaminski N. A Variant In The Promoter Of MUC5B Is Associated With
420 Idiopathic Pulmonary Fibrosis And Not Chronic Obstructive Pulmonary Disease. *B102 INTERSTITIAL*
421 *LUNG DISEASE: NOVEL MANAGEMENT AND OUTCOME STRATEGIES* 2011;doi:10.1164/ajrccm-
422 conference.2011.183.1_meetingabstracts.a6395.
423 11.Moore C, Blumhagen RZ, Yang IV, Walts A, Powers J, Walker T, Bishop M, Russell P, Vestal B,
424 Cardwell J, Markin CR, Mathai SK, Schwarz MI, Steele MP, Lee J, Brown KK, Loyd JE, Crapo JD,
425 Silverman EK, Cho MH, James JA, Guthridge JM, Cogan JD, Kropski JA, Swigris JJ, Bair C, Kim DS, Ji
426 W, Kim H, *et al.* Resequencing Study Confirms That Host Defense and Cell Senescence Gene
427 Variants Contribute to the Risk of Idiopathic Pulmonary Fibrosis. *Am J Respir Crit Care Med*
428 2019;200:199–208.
429 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.:

13.Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, Whitbourne S, Deen J, Shannon C, Humphries D, Guarino P, Aslan M, Anderson D, LaFleur R, Hammond T, Schaa K, Moser J, Huang G, Muralidhar S, Przygodzki R, O'Leary TJ. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of Clinical Epidemiology* 2016;

14.COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature* 2021;doi:10.1038/s41586-021-03767-x.

15.Fang H, Hui Q, Lynch J, Honerlaw J, Assimes TL, Huang J, Vujkovic M, Damrauer SM, Pyarajan S, Gaziano JM, DuVall SL, O'Donnell CJ, Cho K, Chang K-M, Wilson PWF, Tsao PS, VA Million Veteran Program, Sun YV, Tang H. Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. *Am J Hum Genet* 2019;105:763–772.

16.Ioannou GN, Locke E, Green P, Berry K, O'Hare AM, Shah JA, Crothers K, Eastment MC, Dominitz JA, Fan VS. Risk Factors for Hospitalization, Mechanical Ventilation, or Death Among 10 131 US Veterans With SARS-CoV-2 Infection. *JAMA Netw Open* 2020;3:e2022310.

17.Obeidat M, Frank AR, Icardi MS, Klutts JS. VA-Wide, Multicenter Verification Study of the Cepheid Xpert SARS-CoV-2 Assay. *Acad Pathol* 2021;8:23742895211011911.

18.WHO Working Group on the Clinical Characterisation and Management of COVID-19 infection. A minimal common outcome measure set for COVID-19 clinical research. *Lancet Infect Dis* 2020;20:e192–e197.

19.Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993;

20.Kosmidis I, Firth D. Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika* 2021;

454 21.Kosmidis I. brglm2: Bias reduction in generalized linear models. *R package version 0 1* 2017;5.:

455 22.Viechtbauer W. Conducting Meta-Analyses inRwith themetaforPackage. *Journal of Statistical*

456 *Software* 2010;

457 23.Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR,

458 Roden DM, Crawford DC. PheWAS: demonstrating the feasibility of a phenome-wide scan to

459 discover gene-disease associations. *Bioinformatics* 2010;26:1205–1210.

460 24.Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to

461 the challenge of larger and richer datasets. *GigaScience* 2015;

462 25.Mägi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. *BMC*

463 *Bioinformatics* 2010;11:288.

464 26.Fadista J, Kraven LM, Karjalainen J, Andrews SJ, Geller F, COVID-19 Host Genetics Initiative,

465 Baillie JK, Wain LV, Jenkins RG, Feenstra B. Shared genetic etiology between idiopathic pulmonary

466 fibrosis and COVID-19 severity. *EBioMedicine* 2021;65:103277.

467 27.Evans CM, Fingerlin TE, Schwarz MI, Lynch D, Kurche J, Warg L, Yang IV, Schwartz DA. Idiopathic

468 Pulmonary Fibrosis: A Genetic Disease That Involves Mucociliary Dysfunction of the Peripheral

469 Airways. *Physiol Rev* 2016;96:1567–1591.

470 28.Rose MC, Voynow JA. Respiratory tract mucin genes and mucin glycoproteins in health and

471 disease. *Physiol Rev* 2006;86:245–278.

472 29.Seibold MA, Smith RW, Urbanek C, Groshong SD, Cosgrove GP, Brown KK, Schwarz MI, Schwartz

473 DA, Reynolds SD. The idiopathic pulmonary fibrosis honeycomb cyst contains a mucocilary

474 pseudostratified epithelium. *PLoS One* 2013;8:e58658.

475 30.Nakano Y, Yang IV, Walts AD, Watson AM, Helling BA, Fletcher AA, Lara AR, Schwarz MI, Evans

476 CM, Schwartz DA. MUC5BPromoter Variant rs35705950 Affects MUC5B Expression in the Distal

477 Airways in Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care*

Medicine 2016;

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

32.Conti C, Montero-Fernandez A, Borg E, Osadolor T, Viola P, De Lauretis A, Stock CJ, Bonifazi M, Bonini M, Caramori G, Lindahl G, Blasi FB, Nicholson AG, Wells AU, Sestini P, Renzoni E. Mucins MUC5B and MUC5AC in Distal Airways and Honeycomb Spaces: Comparison among Idiopathic Pulmonary Fibrosis/Usual Interstitial Pneumonia, Fibrotic Nonspecific Interstitial Pneumonitis, and Control Lungs. *Am J Respir Crit Care Med* 2016;193:462–464.

33.Roy MG, Livraghi-Butrico A, Fletcher AA, McElwee MM, Evans SE, Boerner RM, Alexander SN, Bellinghausen LK, Song AS, Petrova YM, Tuvim MJ, Adachi R, Romo I, Bordt AS, Bowden MG, Sisson JH, Woodruff PG, Thornton DJ, Rousseau K, De la Garza MM, Moghaddam SJ, Karmouty-Quintana H, Blackburn MR, Drouin SM, Davis CW, Terrell KA, Grubb BR, O'Neal WK, Flores SC, *et al.* Muc5b is required for airway defence. *Nature* 2014;505:412–416.

34.Hancock LA, Hennessy CE, Solomon GM, Dobrinskikh E, Estrella A, Hara N, Hill DB, Kissner WJ, Markovetz MR, Grove Villalon DE, Voss ME, Tearney GJ, Carroll KS, Shi Y, Schwarz MI, Thelin WR, Rowe SM, Yang IV, Evans CM, Schwartz DA. Muc5b overexpression causes mucociliary dysfunction and enhances lung fibrosis in mice. *Nat Commun* 2018;9:5363.

35.Peljto AL, Zhang Y, Fingerlin TE, Ma S-F, Garcia JGN, Richards TJ, Silveira LJ, Lindell KO, Steele MP, Loyd JE, Gibson KF, Seibold MA, Brown KK, Talbert JL, Markin C, Kossen K, Seiwert SD, Murphy E, Noth I, Schwarz MI, Kaminski N, Schwartz DA. Association between the MUC5B promoter polymorphism and survival in patients with idiopathic pulmonary fibrosis. *JAMA* 2013;309:2232–

2239.

36.Raghu G, Remy-Jardin M, Myers JL, Richeldi L, Ryerson CJ, Lederer DJ, Behr J, Cottin V, Danoff SK, Morell F, Flaherty KR, Wells A, Martinez FJ, Azuma A, Bice TJ, Bouros D, Brown KK, Collard HR, Duggal A, Galvin L, Inoue Y, Jenkins RG, Johkoh T, Kazerooni EA, Kitaichi M, Knight SL, Mansour G, Nicholson AG, Pipavath SNJ, *et al.* Diagnosis of Idiopathic Pulmonary Fibrosis. An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline. *Am J Respir Crit Care Med* 2018;198:e44–e68.

37.Mukae H, Iiboshi H, Nakazato M, Hiratsuka T, Tokojima M, Abe K, Ashitani J, Kadota J, Matsukura S, Kohno S. Raised plasma concentrations of alpha-defensins in patients with idiopathic pulmonary fibrosis. *Thorax* 2002;57:623–628.

38.Konishi K, Gibson KF, Lindell KO, Richards TJ, Zhang Y, Dhir R, Bisceglia M, Gilbert S, Yousem SA, Song JW, Kim DS, Kaminski N. Gene expression profiles of acute exacerbations of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2009;180:167–175.

39.Korfei M, Stelmaszek D, MacKenzie B, Skwarna S, Chillappagari S, Bach AC, Ruppert C, Saito S, Mahavadi P, Klepetko W, Fink L, Seeger W, Lasky JA, Pullamsetti SS, Krämer OH, Guenther A. Comparison of the antifibrotic effects of the pan-histone deacetylase-inhibitor panobinostat versus the IPF-drug pirfenidone in fibroblasts from patients with idiopathic pulmonary fibrosis. *PLoS One* 2018;13:e0207915.

40.Montesi SB, Fisher JH, Martinez FJ, Selman M, Pardo A, Johannson KA. Update in Interstitial Lung Disease 2019. *Am J Respir Crit Care Med* 2020;202:500–507.

Figures and Tables

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

Characteristics	Million Veteran Program	COVID-19 Positive
	Number (%)	Number (%)
Total Patients	658,582	13,841
Male	592516 (90)	12,320 (89)
Genetic Ancestry		
European	464961 (70)	8011 (58)
African	123120 (19)	3749 (27)
Hispanic	52183 (8)	1903 (14)
Asian	8329 (1)	178 (1)
Other	9989 (2)	0
Muc5B rs35705950		
0 copy	10604 (1.6)	353 (25)
1 copy	2161 (0.03)	75 (0.05)
2 copies		
Comorbidities		
Obesity (phecode = 278)	283197 (43)	8905 (64)
Hypertension (phecode = 401.1)	451998 (69)	10617 (77)
Type 2 Diabetes (phecode = 250.2)	227575 (34)	10491 (76)
Coronary Artery Disease (phecode = 411.4)	152136 (23)	4182 (30)
Chronic Kidney Disease (phecode = 585.2)	10046 (15)	533 (38)
Outcomes		
Hospitalized	-	4491 (32)
Severe	-	657 (47)
Deceased	-	644 (46)

535

536 **Table 2**. Association of rs35705950 in *MUC5B* with (i) COVID-19 Positive vs Population Controls, (ii)
537 COVID-19 Positive, Hospitalized vs Population Controls, and (iii) COVID-19 Positive, Hospitalized vs
538 COVID-19 Positive, not Hospitalized. Odds ratio (OR) and 95% confidence interval (95% CI) is reported
539 for the minor (T) allele, and results are shown for VA Million Veteran Program (MVP) African Americans
540 (AFR), European Americans (EUR), Hispanic/Latino Americans (HIS), and trans-ethnic meta-analysis (ALL),
541 the COVID-19 Host Genetics Initiative (HGI) trans-ethnic release 5 meta-analysis excluding MVP and
542 23&Me, and the meta-analysis of MVP and HGI (META).

Analysis	Population	N Case	N Control	Total N	EAF	OR (95% CI)	P
Positive vs Population Control	MVP (AFR)	6,411	114,781	121,192	0.02	0.99 [0.87, 1.12]	0.826
	MVP (EUR)	15,814	443,428	459,242	0.11	0.96 [0.92, 0.99]	0.019
	MVP (HIS)	3,128	47,462	50,590	0.07	0.95 [0.85, 1.05]	0.275
	MVP (ALL)	25,353	605,671	631,024	0.09	0.96 [0.93, 1.00]	0.060
	HGI (ALL)	25,652	1,282,972	1,308,624	0.11	0.98 [0.95, 1.01]	0.134
	META	**51,005**	**1,888,643**	**1,939,648**	**0.10**	**0.97 [0.95, 0.99]**	**4.57E-03**
Hospitalized vs Population Control	MVP (AFR)	1,739	119,453	121,192	0.02	0.83 [0.64, 1.07]	0.147
	MVP (EUR)	3,325	455,917	459,242	0.11	0.87 [0.80, 0.94]	5.43E-04
	MVP (HIS)	657	49,933	50,590	0.07	0.86 [0.68, 1.07]	0.182
	MVP (ALL)	5,721	625,303	631,024	0.09	0.86 [0.80, 0.93]	7.35E-05
	HGI (ALL)	9,086	1,001,201	1,010,287	0.11	0.91 [0.85, 0.97]	4.12E-03
	META	**14,807**	**1,626,504**	**1,641,311**	**0.10**	**0.89 [0.85, 0.93]**	**1.88E-06**
Hospitalized vs Not Hospitalized	MVP (AFR)	1,739	4,672	6,411	0.02	0.80 [0.59, 1.08]	0.141
	MVP (EUR)	3,325	12,489	15,814	0.11	0.89 [0.81, 0.97]	0.012
	MVP (HIS)	657	2,471	3,128	0.07	0.88 [0.68, 1.14]	0.319
	MVP (ALL)	5,721	19,632	25,353	0.08	0.88 [0.81, 0.96]	2.64E-03
	HGI (ALL)	4,420	11,093	15,513	0.16	0.97 [0.88, 1.08]	0.575
	META	**10,141**	**30,725**	**40,866**	**0.11**	**0.91 [0.86, 0.98]**	**7.20E-03**

543

544

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, age^2, and PC1-20, including an interaction between additive MUC5B rs35705950 allele and COVID-19 infection.

	COVID-19 negative	**COVID-19 positive**	COVID-19 & *MUC5B* p-value for interaction
	OR (95% CI) of a *MUC5B* allele		p=0.0009
	1.08 (1.00, 1.16) p=0.048	0.89 (0.76, 0.96) p=0.008	
***MUC5B*=0**	***MUC5B*=1**	***MUC5B*=2**	
OR (95% CI) of COVID-19 positive status			
10.00 (9.35, 10.70) p<0.0001	7.91 (6.97, 8.97) p<0.0001	6.26 (4.83, 8.08) p<0.0001	

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.

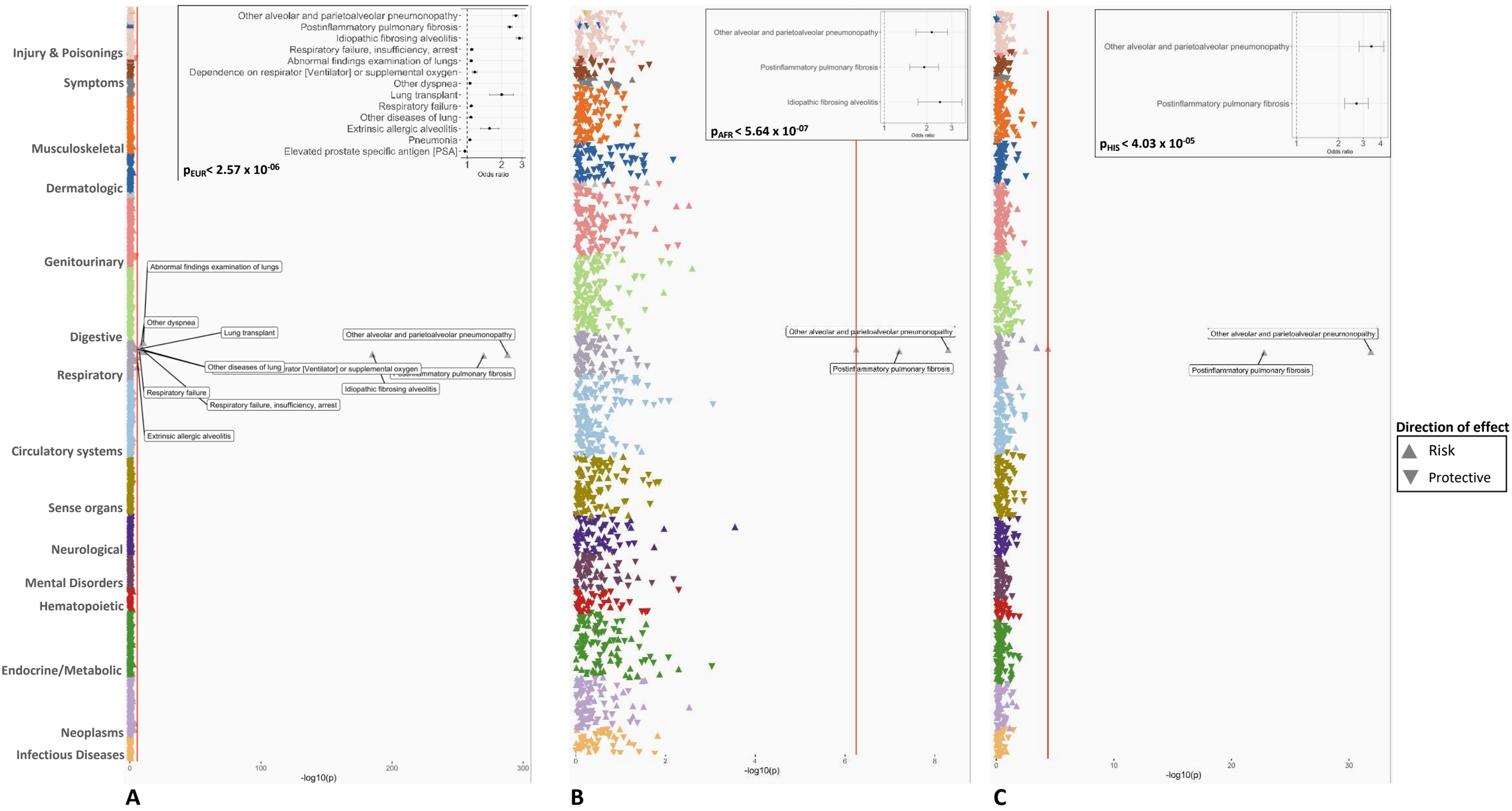
Figure 2. **Laboratory-Wide Association Study (PheWAS) of *MUC5B* rs35705950 allele in the Million Veteran Program.** A LabWAS plot shows associations of rs35705950 and median values of laboratory measures extracted from electronic health records data prior to COVID-19 in MVP participants. The bottom 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.

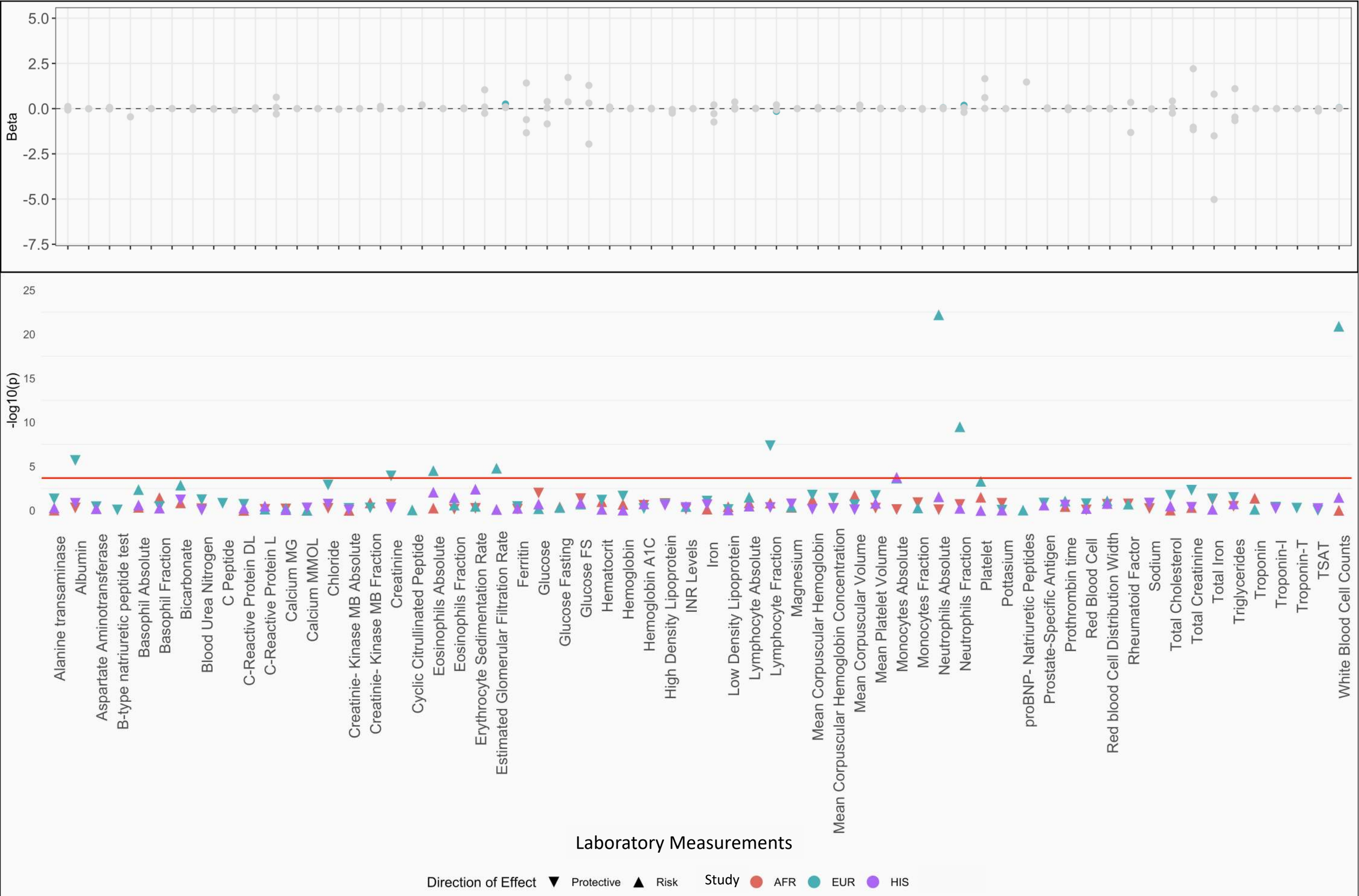
Figure 3. Forest plot association of rs35705950 in *MUC5B* with (i) COVID-19 Positive vs Population Controls, (ii) COVID-19 Positive, Hospitalized vs Population Controls, and (iii) COVID-19 Positive,

576 Hospitalized vs COVID-19 Positive, not Hospitalized. Odds ratio (OR) and 95% confidence interval (95%
577 CI) is reported for the minor (T) allele, and results are shown for VA Million Veteran Program (MVP)
578 African Americans (AFR), European Americans (EUR), Hispanic/Latino Americans (HIS), and trans-ethnic
579 meta-analysis (ALL), the COVID-19 Host Genetics Initiative (HGI) trans-ethnic release 5 meta-analysis
580 excluding MVP and 23&Me, and the meta-analysis of MVP and HGI (META).

581

582


Injury & Poisonings
Symptoms
Musculoskeletal
Dermatologic
Genitourinary
Digestive
Respiratory
Circulatory systems
Sense organs
Neurological
Mental Disorders
Hematopoietic
Endocrine/Metabolic
Neoplasms
Infectious Diseases
Other alveolar and parietoalveolar pneumonopathy
Postinflammatory pulmonary fibrosis
Idiopathic fibrosing alveolitis
Respiratory failure, insufficiency, arrest
Abnormal findings examination of lungs
Dependence on respirator [Ventilator] or supplemental oxygen
Other dyspnea
Lung transplant
Respiratory failure
Other diseases of lung
Extrinsic allergic alveolitis
Pneumonia
Elevated prostate specific antigen [PSA]
Odds ratio
$p_{EUR} < 2.57 \times 10^{-06}$
Abnormal findings examination of lungs
Other dyspnea
Lung transplant
Other alveolar and parietoalveolar pneumonopathy
Other diseases of lung
Idiopathic fibrosing alveolitis
Respiratory failure
Respiratory failure, insufficiency, arrest
Extrinsic allergic alveolitis
-log10(p)
A
Other alveolar and parietoalveolar pneumonopathy
Postinflammatory pulmonary fibrosis
Idiopathic fibrosing alveolitis
Odds ratio
$p_{AFR} < 5.64 \times 10^{-07}$
Other alveolar and parietoalveolar pneumonopathy
Postinflammatory pulmonary fibrosis
-log10(p)
B
Other alveolar and parietoalveolar pneumonopathy
Postinflammatory pulmonary fibrosis
Odds ratio
$p_{HIS} < 4.03 \times 10^{-05}$
Other alveolar and parietoalveolar pneumonopathy
Postinflammatory pulmonary fibrosis
-log10(p)
C
Direction of effect
Risk
Protective


Beta
5.0
2.5
0.0
-2.5
-5.0
-7.5
-log10(p)
25
20
15
10
5
0
Alanine transaminase
Albumin
Aspartate Aminotransferase
B-type natriuretic peptide test
Basophil Absolute
Basophil Fraction
Bicarbonate
Blood Urea Nitrogen
C Peptide
C-Reactive Protein DL
C-Reactive Protein L
Calcium MG
Calcium MMOL
Chloride
Creatinie- Kinase MB Absolute
Creatinie- Kinase MB Fraction
Creatinine
Cyclic Citrullinated Peptide
Eosinophils Absolute
Eosinophils Fraction
Erythrocyte Sedimentation Rate
Estimated Glomerular Filtration Rate
Ferritin
Glucose
Glucose Fasting
Glucose FS
Hematocrit
Hemoglobin
Hemoglobin A1C
High Density Lipoprotein
INR Levels
Iron
Low Density Lipoprotein
Lymphocyte Absolute
Lymphocyte Fraction
Magnesium
Mean Corpuscular Hemoglobin
Mean Corpuscular Hemoglobin Concentration
Mean Corpuscular Volume
Mean Platelet Volume
Monocytes Absolute
Monocytes Fraction
Neutrophils Absolute
Neutrophils Fraction
Platelet
Pottasium
proBNP- Natriuretic Peptides
Prostate-Specific Antigen
Prothrombin time
Red Blood Cell
Red blood Cell Distribution Width
Rheumatoid Factor
Sodium
Total Cholesterol
Total Creatinine
Total Iron
Triglycerides
Troponin
Troponin-I
Troponin-T
TSAT
White Blood Cell Counts
Laboratory Measurements
Direction of Effect
Protective
Risk
Study
AFR
EUR
HIS