A Phenome-Wide Association Study of genes associated with COVID-19 severity reveals shared genetics with complex diseases in the Million Veteran Program

Anurag Verma, PhD[1,9,10¶]; Noah L. Tsao, BS[1,11¶]; Lauren Thomann, MPH[2]; Yuk-Lam Ho[7]; Sudha K. Iyengar, PhD[4,22], Shiuh-Wen Luoh, MD[3,24], Rotonya Carr, MD[1,10]; Dana Crawford, PhD[19,22,23]; Jimmy T. Efird, PhD[16]; Jennifer Huffman, PhD[2]; Adriana Hung, MD[17,20]; Kerry L. Ivey, PhD[9,21,22];; Michael Levin, MD[10], PhD[3,24]; Julie Lynch, PhD[5]; Pradeep Natarajan, MD, MMSc[2,15,31,32]; Saiju Pyarajan, PhD[2,12]; Alexander Bick MD, PhD[2,28]; Lauren Costa[7]; Giulio Genovese, PhD[32,33,34]; Richard Hauger, MD[35]; Ravi Madduri[36,37]; Gita A. Pathak PhD[6,42], Renato Polimanti PhD[6,42], Benjamin Voight, PhD[1,9,26,27]; Marijana Vujkovic, PhD[1,10]; Maryam Zekavat, BSc[2,38,39]; Hongyu Zhao, PhD[6,39,40]; Marylyn D. Ritchie, PhD[3]; *VA Million Veteran Program COVID-19 Science Initiative*, Kyong-Mi Chang, MD[1,10]; Kelly Cho, PhD[7]; Juan P. Casas, MD, PhD[7,13]; Philip S. Tsao, PhD[8,41]; J. Michael Gaziano, MD, MPH[7,13]; Christopher O'Donnell, MD, MPH [2,12,13]; Scott M. Damrauer, MD[1,11]*, Katherine P. Liao, MD, MPH[2,12,13]*

[1] Corporal Michael Crescenz VA Medical Center, Philadelphia, Philadelphia, USA
[2] VA Boston Healthcare System, Boston, Massachusetts, USA
[3] VA Portland Health Care System, Portland OR, USA
[4] Louis Stokes Cleveland VA Medical Center, Cleveland, OH , USA
[5] VA Informatics and Computing Infrastructure, VA Salt Lake City Health Care System, Salt Lake City, Utah, USA.
[6] VA Connecticut Healthcare System, West Haven, CT, USA
[7] Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Boston, Massachusetts, USA
[8] VA Palo Alto Health Care System, Palo Alto, California, USA
[9] Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA.
[10] Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[11] Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[12] Harvard Medical School, Boston, Massachusetts, USA.
[13] Brigham and Women's Hospital, Boston, Massachusetts, USA.
[14] Cardiovascular Research Center, Massachusetts General Hospital, Boston, Massachusetts, USA.
[15] Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA
[16] Cardiovascular Medicine Division, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA
[17] Division of Hematology and Medical Oncology, Knight Cancer Institute, Oregon Health and Science University, Portland, Oregon, USA
[18] College of Nursing and Health Sciences, University of Massachusetts, Lowell, Massachusetts, USA.
[19] Cleveland Institute for Computational Biology, Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, USA.
[20] Cooperative Studies Program Epidemiology Center, Health Services Research and Development, DVAHCS (Duke University Affiliate), Durham, North Carolina, USA.
[21] Tennessee Valley Healthcare System (Nashville VA) & Vanderbilt University, Nashville, Tennessee, USA.
[22] Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, USA.
[23] Department of Genetics and Genome Sciences, Case Western Reserve University, Cleveland, Ohio, USA
[24] Division of Hematology and Medical Oncology, Knight Cancer Institute, Oregon Health and Science University, Portland, Oregon, USA
[25] Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA.
[26] Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[27] Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[28] Vanderbilt University, Nashville, Tennessee, USA

[29] South Australian Health and Medical Research Institute, Infection and Immunity Theme, South Australia, Australia
[30] Harvard T.H. Chan School of Public Health, Department of Nutrition, Massachusetts, USA
[31] Cardiovascular Research Center, Massachusetts General Hospital, Boston, Massachusetts, USA.
[32] Program in Medical and Population Genetics and the Cardiovascular Disease Initiative, Broad Institute of Harvard & MIT, Cambridge, Massachusetts, USA
[33] Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[34] Department of Genetics, Harvard Medical School, Boston, MA, USA
[35] Department of Psychiatry, University of California, San Diego, La Jolla, CA; Center of Excellence for Stress and Mental Health, VA San Diego Healthcare System, San Diego, CA, USA
[36] University of Chicago Consortium for Advanced Science and Engineering, The University of Chicago, Chicago, Illinois, USA
[37] Data Science and Learning Division, Argonne National Laboratory, Lemont, Illinois.
[38] Broad Institute of MIT & Harvard, Cambridge, MA, USA
[39] Yale School of Medicine New Haven, CT, USA
[40] Computational Biology and Bioinformatics Program, Yale University, New Haven, CT, USA
[41] Department of Medicine (Cardiovascular Medicine), Stanford University School of Medicine, Stanford, CA, USA
[42] Department of Psychiatry, Yale School of Medicine, Connecticut, USA

Joint Authorship
¶These authors contributed equally to this work
*These authors jointly supervised this work
Correspondence
Anurag Verma: anurag.verma@pennmedicine.upenn.edu, Katherine Liao: kliao@bwh.harvard.edu

Abstract

The study aims to determine the shared genetic architecture between COVID-19 severity with existing medical conditions using electronic health record (EHR) data. We conducted a Phenome-Wide Association Study (PheWAS) of genetic variants associated with critical illness (n=35) or hospitalization (n=42) due to severe COVID-19 using genome-wide association summary from the Host Genetics Initiative. PheWAS analysis was performed using genotype-phenotype data from the Veterans Affairs Million Veteran Program (MVP). Phenotypes were defined by International Classification of Diseases (ICD) codes mapped to clinically relevant groups using published PheWAS methods. Among 658,582 Veterans, variants associated with severe COVID-19 were tested for association across 1,559 phenotypes. Variants at the *ABO* locus (rs495828, rs505922) associated with the largest number of phenotypes ($n_{rs495828}$= 53 and $n_{rs505922}$=59); strongest association with venous embolism, odds ratio ($OR_{rs495828}$ 1.33 ($p=1.32 \times 10^{-199}$), and thrombosis $OR_{rs505922}$ 1.33, $p=2.2 \times 10^{-265}$. Among 67 respiratory conditions tested, 11 had significant associations including *MUC5B* locus (rs35705950) with increased risk of idiopathic fibrosing alveolitis OR 2.83, $p=4.12 \times 10^{-191}$; *CRHR1 (*rs61667602) associated with reduced risk of pulmonary fibrosis, OR 0.84, $p=2.26 \times 10^{-12}$. The *TYK2* locus (rs11085727) associated with reduced risk for autoimmune conditions, e.g., psoriasis OR 0.88, $p=6.48 \times 10^{-23}$, lupus OR 0.84, $p=3.97 \times 10^{-06}$. PheWAS stratified by genetic ancestry demonstrated differences in genotype-phenotype associations across ancestry. *LMNA* (rs581342) associated with neutropenia OR 1.29 $p=4.1 \times 10^{-13}$ among Veterans of African ancestry but not European. Overall, we observed **a** shared genetic architecture between COVID-19 severity and conditions related to underlying risk factors for severe and poor COVID-19 outcomes. Differing associations between genotype-phenotype across ancestries may inform heterogenous outcomes

observed with COVID-19. Divergent associations between risk for severe COVID-19 with autoimmune inflammatory conditions both respiratory and non-respiratory highlights the shared pathways and fine balance of immune host response and autoimmunity and caution required when considering treatment targets.

Introduction

Coronavirus disease 2019 (COVID-19) first identified in December of 2019[1], became a global pandemic by March 2020. As of September 2021, COVID-19, transmitted by the SARS-CoV-2 virus, has resulted in the loss of over 4.6 million lives worldwide.[2] Identifying host genetic variants associated with severe clinical manifestations from COVID-19, can identify key pathways important in the pathogenesis of this condition. International efforts such as the COVID-19 Host Genetics Initiative (HGI)[3] have meta-analyzed genome-wide association study (GWAS) summary statistics at regular intervals to identify novel genetic associations with COVID-19 severity. Thus far, ten independent variants associated with COVID-19 severity at genome-wide significance have been identified, most notably at the *ABO* locus.[4] These GWASs have also identified variations in genes involving inflammatory cytokines and interferon signaling pathways such as *IFNAR2*, *TYK2*, and *DPP9*.[4]

The unprecedented availability of genome-wide data for COVID-19 provides an opportunity to study clinical conditions that share genetic risk factors for COVID-19 severity. Examining known conditions, each with a body of knowledge regarding important pathways and targets, may in turn improve our understanding of pathways relevant for COVID-19 severity and inform the development of novel treatments against this pathogen. The Phenome-Wide Association Study (PheWAS) is an approach for simultaneously testing genetic variants' association with a wide spectrum of conditions and phenotypes.[5] The Veteran's Affairs (VA) Million Veterans Program (MVP) has generated genotypic data on over 650,000 participants linked with electronic health record (EHR) data containing rich phenotypic data, enables large-scale PheWAS. Moreover, MVP has the highest racial and ethnic diversity of the major biobanks

157 worldwide affording an opportunity to compare whether associations are similar across genetic
158 ancestries.[6]

159 The objective of this study was to use existing clinical EHR data to identify conditions
160 that share genetic variants with COVID-19 severity using the disease-agnostic PheWAS
161 approach. Since COVID-19 is a new condition, identifying existing conditions which share
162 genetic susceptibility may allow us to leverage existing knowledge from these known conditions
163 to provide context regarding important pathways for COVID-19 severity, as well as how
164 pathways may differ across subpopulations.

165 **Methods**

166 **Data sources**

167 The VA MVP is a national cohort launched in 2011 designed to study the contributions of
168 genetics, lifestyle, and military exposures to health and disease among US Veterans.[6] Blood
169 biospecimens were collected for DNA isolation and genotyping, and the biorepository was
170 linked with the VA EHR, which includes diagnosis codes (International Classification of
171 Diseases ninth revision [ICD-9] and tenth revision [ICD-10]) for all Veterans followed in the
172 healthcare system up to September 2019. The single nucleotide polymorphism (SNP) data in the
173 MVP cohort was generated using a custom Thermo Fisher Axiom genotyping platform called
174 MVP 1.0. The quality control steps and genotyping imputation using 1000 Genomes
175 cosmopolitan reference panel on the MVP cohort has been described previously.[7]

176

177

Genetic variant selection

An overview of the analytic workflow is outlined in Fig 1. Variants were derived from the COVID-19 HGI GWAS meta-analysis release v6[3]. In this study, we analyzed the following HGI GWAS summary statistics: 1) hospitalized and critically ill COVID-19 vs. population controls denoted as "A2" in HGI, and referred to as "critical COVID" in this study, and 2) hospitalized because of COVID-19 vs. population controls, denoted as "B2" in HGI, referred to as "hospitalized COVID" in this study[3]. For each GWAS, variants with a Benjamini-Hochberg false discovery rate (FDR) corrected p-value < 0.01 were selected as candidate lead SNPs (3,502 associated with critical COVID, and 4,336 associated with hospitalized COVID). Variants with r^2 <0.1 were clustered within a 250 kb region according to 1000 Genomes phase 3 transethnic reference panel[8], resulting in 45 independent variants associated with critical COVID and 42 variants associated with hospitalized COVID summary statistics. The lead variants from each set of GWAS summary statistics are available in eTable 1.

Outcomes

For both MVP and UKBB, clinical data prior to the onset of the COVID-19 pandemic were used to reduce potential confounding bias from SARS-CoV-2 infection on existing conditions. Phenotypes were defined by phecodes from prior studies[5,9]. Each phecode represents ICD codes grouped into clinically relevant phenotypes for clinical studies. For example, the phecode "deep venous thrombosis" includes "venous embolism of deep vessels of the distal lower extremities," and "deep venous thrombosis of the proximal lower extremity," both of which have distinct ICD codes. Using this approach, all ICD codes for all Veterans in MVP were extracted and each

assigned a phenotype defined by a phecode. ICD-9 and ICD-10 codes were mapped to 1876 phecodes, as previously described.[5,9]

For each phecode, participants with ≥2 phecode-mapped ICD-9 or ICD-10 codes were defined as cases, whereas those with no instance of a phecode-mapped ICD-9 or ICD-10 code were defined as controls. Based on our previous simulation studies of ICD EHR data, populations where the phecode comprises < 200 cases were more likely to result in spurious results[10], and we thus applied this threshold in each ancestry group. In total, we analyzed 1,617 (EUR), 1304 (AFR), 993 (HIS), 294 (ASN) phecodes from the MVP cohort. The same process was applied to UKBB, resulting in 1,064 phecodes from 1,615 phecodes.

Phenome-wide association studies

The primary PheWAS analysis used SNPs identified from the HGI GWAS of critical and hospitalized COVID, and tested association of these SNPs with phenotypes extracted from the EHR using data prior to the COVID-19 pandemic. Logistic regression using PLINK2 to examine the SNP association with phecodes and firth regression was applied when logistic regression model failed to converge. Regression models were adjusted for sex, age (at enrollment), age squared, and the first 20 principal components. Genetic ancestry was determined using the HARE method for four major groups: African (AFR), Asian (ASN), Hispanic (HIS), and European (EUR) ancestry[11]. Ancestry-specific PheWAS was first performed in these four groups, and summary data were meta-analyzed using an inverse-variance weighted fixed-effects model implemented in the PheWAS R package[9]. We assessed heterogeneity using I^2 and excluded any results with excess heterogeneity ($I^2 > 40\%$).

To address multiple testing, an association between SNP and phecode with FDR $p < 0.01$ was considered significant. Thus, the threshold for significance was $p < 6.07 \times 10^{-05}$ for critical COVID lead variants, and $p < 4.13 \times 10^{-05}$ for hospitalized COVID lead variants. In the main manuscript we highlight PheWAS significant associations using FDR < 0.01 and an effect size associated with increased or reduced risk for a condition by 10%, with complete PheWAS results provided in S2 Table and S3 Table.

Results

We studied 658,582 MVP participants, with mean age 68 years (SD), 90% male, with 30% participants from non-European ancestry (Table 1). The PheWAS was performed on 35 genetic variants associated with critical COVID-19, and 42 genetic variants (S1 Table) associated with hospitalized COVID, across 1,559 phenotypes.

From the trans-ethnic meta-analysis, we identified 151 phenotypes significantly associated with critical COVID GWAS-identified variants, and 156 associations with hospitalized COVID GWAS-identified lead variants (FDR, $p<0.01$). Among these lead variants with significant PheWAS associations, 10 SNPs were associated with reduced risk of critical and hospitalized COVID-19 in HGI. Six variants were common to both severe and hospitalized COVID and had significant PheWAS associations, namely, variations nearest to the genes *ABO* (rs495828 and rs505922), *DPP9* (rs2277732), *MUC5B* (rs35705950), TYK2 (rs11085727), and *CCHCR1* (rs9501257) (S2 Table and S3 Table).

241 Association of ABO loci with known risk factors and outcomes related to COVID-19 severity

242 In the transethnic meta-analysis, the phenotypes with the strongest association with variants near
243 *ABO* locus (rs495828 and rs505922) was "hypercoagulable state" ($OR_{critical\_PheWAS}$ = 1.48 [1.42 -
244 1.54], $P_{critical\_PheWAS} = 1.84 \times 10^{-40}$; $OR_{hospitalized\_PheWAS}$ = 1.51 [1.46 - 1.56], $P_{hospitalized\_PheWAS}$ =
245 2.11×10^{-55}, Fig 2). The ABO loci had the largest number of significant PheWAS association
246 findings, accounting for 35% (53/151) of significant phenotype associations in the critical
247 COVID PheWAS, and 37% (59/156) in the hospitalized COVID PheWAS. The phenotypes with
248 the most significant associations and largest effect size were related to hypercoagulable states
249 and coagulopathies. As expected, conditions not related to coagulopathy associated with the
250 *ABO* locus, included type 2 diabetes and ischemic heart disease, have been reported as risk
251 factors for or are complications associated with COVID-19 severity and mortality (Fig 2, S2
252 Table and S3 Table).

253 Associations between variants associated with COVID-19 severity and respiratory conditions
254 and infections

255 Among 68 respiratory conditions, only 11 diseases had significant associations (FDR < 0.01)
256 shared with genetic variants associated with severe COVID-19. The most significant association
257 was observed between rs35705950 (*MUC5B*) and idiopathic fibrosing alveolitis (OR = 2.83
258 [2.76 – 2.90]; $P = 4.12 \times 10^{-191}$), also known as idiopathic pulmonary fibrosis (IPF). Similarly,
259 rs2277732 near *DPP9* was associated with IPF (OR = 1.16 [1.09 – 1.22]; $P = 5.84 \times 10^{-06}$), both
260 association between *MUC5B, DPP9* variants and IPF has been reported in previous studies.[12]
261 However, the association of genetic variants with other respiratory conditions may represent
262 novel findings: the association of intronic variant rs61667602 in *CRHR1* with reduced risk of

post inflammatory pulmonary fibrosis (OR = 0.84 [0.80 – 0.89]; $P = 2.26\times 10^{-12}$), "alveolar and parietoalveolar pneumonopathy" (OR = 0.80 [0.72 – 0.88]; $P = 1.58 \times 10^{-08}$) and IPF (OR = 0.87 [0.82 - 0.92], $P = 7.5 \times 10^{-07}$). We did not detect associations between any of the variants and other respiratory conditions which are known risk factors for COVID-19 such as COPD, cystic fibrosis, pulmonary hypertension. (S2 Table, S3 Table).

Associations between variants associated with COVID-19 severity and reduced risk for certain phenotypes

The rs11085727-T allele of *TYK2*, a lead variant from the both critically ill and hospitalized COVID GWAS was associated with a reduced risk for psoriasis (OR = 0.88 [0.86- 0.91], $P = 6.48 \times 10^{-23}$), psoriatic arthropathy (OR = 0.82 [0.76 - 0.87], $P = 6.97 \times 10^{-12}$), and lupus (OR = 0.84 [0.76 - 0.91], $P = 63.97 \times 10^{-06}$). This *TYK2* signal has been previously reported to be associated with reduced risk of psoriasis, psoriatic arthropathy, type 1 diabetes, systemic lupus erythematosus and RA as well as other autoimmune inflammatory conditions[13,14] (Table 2).

Ancestry specific PheWAS provide insights into disease risks across ancestries

The PheWAS analyses performed across four major genetic ancestry group in MVP observed similar findings as the overall meta-analysis with few associations unique to each ancestry. (Fig 3, S8 Table). SNP rs581342 (*LMNA*), associated with severe COVID-19, was a highly prevalent variant among subjects with AFR ancestry (MAF=0.53) and was associated with neutropenia (OR_{AFR} = 0.82 [0.76 - 0.87], $P_{AFR} = 4.09 \times 10^{-13}$); this association was not observed in larger population of EUR descent (S8 Table). Following up on this finding, we extracted data on laboratory values and observed a strong association between *LMNA* with lower white blood cell

count (beta = -0.34 [-0.35, -0.32], P_{AFR}= 1 x 10^{-300}) and lower median neutrophil fraction (beta = -1.84 [-1.94, -1.75], P_{AFR} = 1 x 10^{-300}) compared to those without this variant. This association in laboratory values was again more significant with a stronger effect size among subjects with AFR ancestry in comparison to EUR (P=0.005). Among AFR individuals, each allele was associated with a 1.84% lower neutrophil fraction, where among EUR individuals, each allele was associated with only a 0.04% reduction (S9 Table).

Similarly, associations between rs9268576 (*HL-DRA*) and thyrotoxicosis was only observed in AFR ancestry participants. The EUR ancestry specific PheWAS identified 39 significant associations which were not observed in other ancestry groups. One such association was between MUC5B variant and phecode for "dependence on respirator [Ventilator] or supplemental oxygen" (OR_{EUR} = 1.16 [1.11 – 1.12], P_{EUR} = 1.72× 10^{-10}) among EUR ancestry participants was not significant in other ancestry population (S8 Table). It is important to note that the conditions with significant association among EUR participants had similar prevalence among other ancestries. However, since there were overall fewer subjects in non-EUR ancestry groups, this likely resulted in lower statistical power to detect associations. All ancestry specific PheWAS results are available in supplementary tables (S4 Table, S5 Table, S6 Table, S7 Table).

Association with variation at sex chromosome

In the hospitalized COVID-19 GWAS, we identified rs4830964 as the only lead variant on chromosome X. The SNP is located near *ACE2* and was associated "non-healing surgical wound" (OR = 0.92 [0.89 – 0.96], P = 2.23× 10^{-05}). Notably, the SNP had nominal association ($p<0.05$) with type 2 diabetes and diabetes related complications that are previously reported

association with variation in *ACE2* (S3 Table). We did not observe any association with this variant in the ancestry specific PheWAS analysis.

Discussion

In this large-scale PheWAS, we identified the shared genetic architecture between variants associated with severe COVID-19 and other complex conditions using data from MVP, one of the largest and most diverse biobanks in the world. Broadly, these risk alleles identified conditions associated with risk factors for severe COVID-19 manifestations such as T2D, ischemic heart disease across all ancestries. Notably, the strongest associations with the highest effect size were related to coagulopathies, specifically, hypercoagulable state including deep venous thrombosis and other thrombotic complications, also shared variants associated with severe COVID-19. In contrast, among respiratory conditions, only idiopathic pulmonary fibrosis and chronic alveolar lung disease shared genetic risk factors, with the notable absence of an association with COPD, pulmonary hypertension, and other respiratory infections. When comparing findings from the two largest ancestry groups in MVP, AFR and EUR, we observed that a risk allele associated with severe COVID-19 that shares an association with neutropenia on among Veterans of AFR ancestry. Finally, we observed that variants associated with severe COVID-19 had an opposite association, or reduced odds with autoimmune inflammatory conditions, such as psoriasis, psoriatic arthritis, RA, and inflammatory lung conditions.

A classic GWAS tests the association between millions of genetic variants with the presence or absence of one phenotype, e.g., GWAS of deep venous thrombosis. In the COVID-19 HGI GWAS, the “phenotype” was patients hospitalized for or critically ill from COVID-19. Clinically, this population includes a mixture of patients with a complex list of medical

conditions at high risk for severe COVID complications and those who had actual complications from COVID-19. Thus, we would anticipate that many of the significant phenotypes would be associated with risk factors such as obesity and deep venous thrombosis. The clinical data used in this study pre-dates the emergence of COVID-19 to reduce potential confounding bias that can occur in a population infected with SARS-CoV-2, e.g., interaction between COVID-19 and type 2 diabetes. Additionally, our findings suggest that the PheWAS approach can be a useful tool to identify clinical factors related to emerging infectious diseases regarding severity or complications when genomic data are available.

The PheWAS results of SNPs in the *ABO* locus served as a positive control for this study. Genetic variations in *ABO* are an established risk factor for COVID-19 severity. Patients with blood group A have a higher risk of requiring mechanical ventilation and extended ICU stay compared with patients with blood group O.[15] These same variations at *ABO* had known associations with a spectrum of blood coagulation disorders identified in studies pre-dating COVID-19.[16–18] The PheWAS of ABO variants identified associations with increased risk of deep vein thrombosis, pulmonary embolism, and other circulatory disorders, in line with prior studies, and recent studies among patients hospitalized with COVID-19.[19–23]

Among the respiratory conditions, only idiopathic pulmonary fibrosis (IPF) and chronic alveoli lung disease had associations with the variants near genes *MUC5B*, *CRHR1*, and *NSF*. Located in the enhancer region of the *MUC5B*, rs35705950, is a known risk factor for IPF, and a high mortality rate was observed among the COVID-19 patients with pre-existing IPF.[24] However, the variant is associated with a reduced risk of severe COVID-19 (OR=0.89), revealing the risk allele's opposing effect for infection and pulmonary fibrosis. In a separate

study of MVP participants tested for COVID-19, we identified a significant mediating effect of the *MUC5B* variant in reducing risk for pneumonia due to COVID-19[25]. An intronic variation in *CRHR1* (rs61667602-T) had reduced risk for severe COVID-19 (OR= 0.91) as well as respiratory conditions such as IPF. *CRHR1* gene is a receptor that binds to the corticotropin-releasing hormone has a key role in immune, behavioral, autonomic, and neuroendocrine responses to stress. Depression and anxiety are the known conditions associated with variations in *CRHR1*, but variations in this gene have also shown associations enhanced improvement in pulmonary function in asthma patients taking inhaled corticosteroid[26]. This finding may inform results from the RECOVERY clinical trial of patients hospitalized with COVID-19 where a survival benefit was observed for dexamethasone use among those receiving respiratory support[27].

Several conditions shared genetic variants associated with severe COVID-19, however, the association was for reduced odds for these conditions. All except one, rosacea, have a known autoimmune etiology. The existing literature can help explain the dual association between reduced risk of autoimmune conditions such as psoriasis and RA and increased risk of severe COVID-19 via *TYK2*. *TYK2*, a member of the Janus Kinase (*JAK*) family of genes, plays a key role in cytokine signal transduction and the inflammatory response, particularly via IL-12, IL-23, and is also important for IL-6 and IL-10 signaling (Fig 3).[28] *TYK2* serves a central role in type 1 interferon signaling, part of the innate immune response blocking the spread of a virus from infected to uninfected cells. Partial loss of *TYK2* function is associated with reduced risk for several autoimmune disorders such as RA and psoriatic disease, conditions treated with immunosuppressive therapy.[13,29–32] Humans with complete *TYK2* loss of function have clinically significant immunodeficiency with increased susceptibility to mycobacterial and viral

infections.[28,33] In line with the *TYK2* findings is enhanced steroid responsiveness among patients with asthma carrying the *CRHR1* variant[34]. Here again, a variant associated with severe COVID-19 is associated with a non-COVID phenotype responsive to immunosuppressive therapy. In summary, reviewing the overall signal of opposing associations of variants with COVID-19 and autoimmune conditions, highlights the known fine balance between host immune response and autoimmunity.

While non-white populations are disproportionately affected by COVID-19, the current genetic studies of severe COVID-19 still predominantly consist of individuals from EUR ancestry. MVP has the most racial and ethnic diversity compared to other major biobanks. The availability of linked EHR data provide the opportunity to provide more in-depth studies of genotype-phenotype associations observed from the PheWAS. The GWAS from the HGI provides the most diverse genomic data of COVID-19 consisting of participants from over 25 countries EUR (33% non-EUR samples), enabling identification of variants more prevalent in non-EUR populations. In the present study, we observed that a variant located in the *LMNA* gene locus was associated with neutropenia in AFR ancestry but not in other ancestry groups, including EUR which would have been well powered to detect an association. Furthermore, examination of actual neutrophil percentages measured as part of routine care demonstrated stronger associations in Veterans of AFR ancestry compared to EUR.

LMNA variants are associated with a broad spectrum of cardiomyopathies such as dilated cardiomyopathies, familial atrial fibrillation. However, the association with neutropenia has not been previously reported. Neutropenia refers to an abnormally low number of neutrophils cell in the blood, and predisposes to increased risk of infection . Epidemiology studies have shown that

lower neutrophil counts are more common in individuals with African Ancestry[36] and are hypothesized to be a result of selection and generally considered benign. Whether low neutrophil levels may clinically impact COVID-19 outcomes remains to be seen and should be further studied.

Limitations

We note several limitations. First, the PheWAS was designed as a broad screen to test for potentially clinically relevant associations between genes and phenotypes, with limited power to detect associations among uncommon conditions, and when further stratified by genetic ancestry. Findings from this study suggest that variants associated with severe COVID-19 are also associated with reduced odds of having an autoimmune inflammatory condition. However, the results cannot provide information on the impact of actual SARS-CoV-2 infection in these individuals after diagnosis of an autoimmune disease

Conclusions

The PheWAS of genetic variants reported to associate with severe COVID-19 demonstrated shared genetic architecture between COVID-19 severity and known underlying risk factors for both severe COVID-19 and poor COVID-19 outcomes, rather than susceptibility to other viral infections. Overall, the associations observed were generally consistent across genetic ancestries, with the exception of a stronger association with neutropenia among Veterans of African ancestry than European ancestry. Notably, only few respiratory conditions had a shared genetic association with severe COVID-19. Among these, variants associated with a reduced risk for severe COVID-19 had an opposite association, with reduced risk for inflammatory and fibrotic pulmonary conditions. Similarly, other divergent associations were observed between severe

COVID-19 and autoimmune inflammatory conditions, shedding light on the concept of the fine balance between immune tolerance and immunodeficiency. This balance will be important when considering therapeutic targets for COVID-19 therapies where pathways may control both inflammation and the viral host response.

Funding

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 or the United States Government. R.M.C. is supported by NIH grants R01 AA026302 and P30 DK0503060. K.P.L. is supported by NIH P30 AR072577, and the Harold and Duval Bowen Fund.

Conflict of Interest

RMC has received research support from Intercept Pharmaceuticals, Inc and Merck & Co. MDR is on the scientific advisory board for Goldfinch Bio and Cipherome. 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.

Acknowledgements

We are grateful to our Veterans for their contributions to MVP. Full acknowledgements for the VA Million Veteran Program COVID-19 Science Initiative can be found in the supplementary

methods. We would like to thank the Host Genetic Initiative for making their data publicly available (https://www.covid19hg.org/acknowledgements/).

References

1. CDC. About COVID-19 - CDC. : https://www.cdc.gov/coronavirus/2019-ncov/cdcresponse/about-COVID-19.html

2. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis*. 2020;20(5):533-534. doi:10.1016/S1473-3099(20)30120-1

3. The COVID-19 Host Genetics Initiative, ganna andrea. *Mapping the Human Genetic Architecture of COVID-19 by Worldwide Meta-Analysis*. Genetic and Genomic Medicine; 2021. doi:10.1101/2021.03.10.21252820

4. The GenOMICC Investigators, The ISARIC4C Investigators, The COVID-19 Human Genetics Initiative, et al. Genetic mechanisms of critical illness in COVID-19. *Nature*. 2021;591(7848):92-98. doi:10.1038/s41586-020-03065-y

5. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*. 2010;26(9):1205-1210. doi:10.1093/bioinformatics/btq126

6. Gaziano JM, Concato J, Brophy M, et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol*. 2016;70:214-223. doi:10.1016/j.jclinepi.2015.09.016

7. Hunter-Zinck H, Shi Y, Li M, et al. Genotyping Array Design and Data Quality Control in the Million Veteran Program. *Am J Hum Genet*. 2020;106(4):535-548. doi:10.1016/j.ajhg.2020.03.004

8. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393

9. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics*. 2014;30(16):2375-2376. doi:10.1093/bioinformatics/btu197

10. Verma A, Bradford Y, Dudek S, et al. A simulation study investigating power estimates in phenome-wide association studies. *BMC Bioinformatics*. 2018;19(1):120. doi:10.1186/s12859-018-2135-0

11. Fang H, Hui Q, Lynch J, et al. Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. *Am J Hum Genet*. 2019;105(4):763-772. doi:10.1016/j.ajhg.2019.08.012

12. Allen RJ, Guillen-Guio B, Oldham JM, et al. Genome-Wide Association Study of Susceptibility to Idiopathic Pulmonary Fibrosis. *Am J Respir Crit Care Med*. 2020;201(5):564-574. doi:10.1164/rccm.201905-1017OC

13. Diogo D, Bastarache L, Liao KP, et al. TYK2 Protein-Coding Variants Protect against Rheumatoid Arthritis and Autoimmunity, with No Evidence of Major Pleiotropic Effects on Non-Autoimmune Complex Traits. Chiorini JA, ed. *PLOS ONE*. 2015;10(4):e0122271. doi:10.1371/journal.pone.0122271

14. Dendrou CA, Cortes A, Shipman L, et al. Resolving TYK2 locus genotype-to-phenotype differences in autoimmunity. *Sci Transl Med*. 2016;8(363):363ra149-363ra149. doi:10.1126/scitranslmed.aag1974

15. Hoiland RL, Fergusson NA, Mitra AR, et al. The association of ABO blood group with indices of disease severity and multiorgan dysfunction in COVID-19. *Blood Adv*. 2020;4(20):4981-4989. doi:10.1182/bloodadvances.2020002623

16. Zietz M, Zucker J, Tatonetti NP. Associations between blood type and COVID-19 infection, intubation, and death. *Nat Commun*. 2020;11(1):5761. doi:10.1038/s41467-020-19623-x

17. Paranjpe I, Fuster V, Lala A, et al. Association of Treatment Dose Anticoagulation With In-Hospital Survival Among Hospitalized Patients With COVID-19. *J Am Coll Cardiol*. 2020;76(1):122-124. doi:10.1016/j.jacc.2020.05.001

18. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA*. Published online February 24, 2020. doi:10.1001/jama.2020.2648

19. Matsunaga H, Ito K, Akiyama M, et al. Transethnic Meta-Analysis of Genome-Wide Association Studies Identifies Three New Loci and Characterizes Population-Specific Differences for Coronary Artery Disease. *Circ Genomic Precis Med*. 2020;13(3):e002670. doi:10.1161/CIRCGEN.119.002670

20. Plagnol V, Howson JMM, Smyth DJ, et al. Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. *PLoS Genet*. 2011;7(8):e1002216. doi:10.1371/journal.pgen.1002216

21. Reilly MP, Li M, He J, et al. Identification of ADAMTS7 as a novel locus for coronary atherosclerosis and association of ABO with myocardial infarction in the presence of coronary atherosclerosis: two genome-wide association studies. *Lancet Lond Engl*. 2011;377(9763):383-392. doi:10.1016/S0140-6736(10)61996-4

22. Trégouët D-A, Heath S, Saut N, et al. Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach. *Blood*. 2009;113(21):5298-5303. doi:10.1182/blood-2008-11-190389

23. Larson NB, Bell EJ, Decker PA, et al. ABO blood group associations with markers of endothelial dysfunction in the Multi-Ethnic Study of Atherosclerosis. *Atherosclerosis*. 2016;251:422-429. doi:10.1016/j.atherosclerosis.2016.05.049

24. Gallay L, Uzunhan Y, Borie R, et al. Risk Factors for Mortality after COVID-19 in Patients with Preexisting Interstitial Lung Disease. *Am J Respir Crit Care Med*. 2021;203(2):245-249. doi:10.1164/rccm.202007-2638LE

25. Verma A, Minnier J, Huffman JE, et al. A MUC5B *Gene Polymorphism, Rs35705950-T, Confers Protective Effects in COVID-19 Infection*. Infectious Diseases (except HIV/AIDS); 2021. doi:10.1101/2021.09.28.21263911

26. Kim WJ, Sheen SS, Kim T-H, et al. Association between *CRHR1* polymorphism and improved lung function in response to inhaled corticosteroid in patients with COPD. *Respirology*. 2009;14(2):260-263. doi:10.1111/j.1440-1843.2008.01425.x

27. The RECOVERY Collaborative Group. Dexamethasone in Hospitalized Patients with Covid-19. *N Engl J Med*. 2021;384(8):693-704. doi:10.1056/NEJMoa2021436

28. Nemoto M, Hattori H, Maeda N, et al. Compound heterozygous TYK2 mutations underlie primary immunodeficiency with T-cell lymphopenia. *Sci Rep*. 2018;8(1):6956. doi:10.1038/s41598-018-25260-8

29. Hellquist A, Järvinen TM, Koskenmies S, et al. Evidence for Genetic Association and Interaction Between the *TYK2* and *IRF5* Genes in Systemic Lupus Erythematosus. *J Rheumatol*. 2009;36(8):1631-1638. doi:10.3899/jrheum.081160

30. Sigurdsson S, Nordmark G, Göring HHH, et al. Polymorphisms in the Tyrosine Kinase 2 and Interferon Regulatory Factor 5 Genes Are Associated with Systemic Lupus Erythematosus. *Am J Hum Genet*. 2005;76(3):528-537. doi:10.1086/428480

31. , The Wellcome Trust Case–Control Consortium (WTCCC) and Alastair Compston, Ban M, Goris A, et al. Replication analysis identifies TYK2 as a multiple sclerosis susceptibility factor. *Eur J Hum Genet*. 2009;17(10):1309-1313. doi:10.1038/ejhg.2009.41

32. Cunninghame Graham DS, Morris DL, Bhangale TR, et al. Association of NCF2, IKZF1, IRF8, IFIH1, and TYK2 with Systemic Lupus Erythematosus. McCarthy MI, ed. *PLoS Genet*. 2011;7(10):e1002341. doi:10.1371/journal.pgen.1002341

33. Watford WT, O'Shea JJ. Human Tyk2 Kinase Deficiency: Another Primary Immunodeficiency Syndrome. *Immunity*. 2006;25(5):695-697. doi:10.1016/j.immuni.2006.10.007

34. Tantisira KG, Lake S, Silverman ES, et al. Corticosteroid pharmacogenetics: association of sequence variants in CRHR1 with improved lung function in asthmatics treated with inhaled corticosteroids. *Hum Mol Genet*. 2004;13(13):1353-1359. doi:10.1093/hmg/ddh149

35. Boxer L, Dale DC. Neutropenia: Causes and consequences. *Semin Hematol*. 2002;39(2):75-81. doi:10.1053/shem.2002.31911

36. Reich D, Nalls MA, Kao WHL, et al. Reduced Neutrophil Count in People of African Descent Is Due To a Regulatory Variant in the Duffy Antigen Receptor for Chemokines Gene. Visscher PM, ed. *PLoS Genet*. 2009;5(1):e1000360. doi:10.1371/journal.pgen.1000360

Tables and Figures

Table 1. Patient characteristics of Million Veteran Program participants

Characteristics	Million Veteran Program
	Number (%)
Total Patients	658,582
Male	592,516 (90)
Genetic Ancestry	
European	464,961 (70)
African	123,120 (19)
Hispanic	52,183 (8)
Asian	83,29 (1)
Other	99,89 (2)
Comorbidities	
Obesity (phecode = 278)	283,197 (43)
Hypertension (phecode = 401.1)	451,998 (69)
Type 2 Diabetes (phecode = 250.2)	227,575 (34)
Coronary Artery Disease (phecode = 411.4)	152,136 (23)
Chronic Kidney Disease (phecode = 585.2)	100,46 (15)

Table 2. Phenotypes sharing association with variants also associated with severe COVID-19 infection, with reduced odds of disease listed in order of p-value*.

Phenotype	OR (95% CI)	p-value	Gene	SNP	COVID-severity
Psoriasis	0.89 [0.86-0.91]	6.48E-23	*TYK2*	rs11085727	Both
Rosacea	0.84 [0.8-0.89]	7.54E-16	*HLA-DPB1*	rs9501257	Critical
Psoriatic arthropathy	0.82 [0.77-0.88]	6.97E-12	*TYK2*	rs11085727	Both
Post-inflammatory pulmonary fibrosis	0.87 [0.83-0.92]	4.54E-09	*NSF*	rs9896243	Critical
Vitiligo	0.69 [0.56-0.82]	3.03E-08	*CCHCR1*	rs111837807	Both
Sarcoidosis	0.74 [0.62-0.85]	1.80E-07	*CCHCR1*	rs111837807	Both
Lupus (localized and systemic)	0.84 [0.77-0.91]	3.97E-06	*TYK2*	rs11085727	Both
Cutaneous lupus erythematosus	0.79 [0.68-0.89]	6.21E-06	*TYK2*	rs11085727	Both
Post-inflammatory pulmonary fibrosis	0.85 [0.8-0.9]	2.26E-12	*CRHR1*	rs61667602	Hospitalized
Rheumatoid arthritis	0.84 [0.79-0.9]	4.20E-10	*HLA-DRA*	rs9268576	Hospitalized
Idiopathic fibrosing alveolitis	0.81 [0.73-0.88]	1.58E-08	*CRHR1*	rs61667602	Hospitalized
Rheumatoid arthritis and other inflammatory polyarthropathies	0.88 [0.84-0.93]	6.34E-08	*HLA-DRA*	rs9268576	Hospitalized
Other alveolar and parietoalveolar pneumonopathy	0.88 [0.83-0.93]	7.50E-07	*CRHR1*	rs61667602	Hospitalized

*OR<0.9 and P<10-5 shown in table, full results in supplementary; if multiple related conditions, e.g. psoriasis, psoriasis vulgaris, psoriasis and related disorders, description with lowest p-value selected shown in table.

Fig 1. Overview of variant selection and PheWAS analysis design.

Fig 2. PheWAS results of candidate SNPs from GWAS of critically ill and hospitalized COVID-19. Significant associations between 48 SNPs from critical ill COVID GWAS (A) and 39 SNPs from hospitalized COVID (C) and EHR derived phenotypes in the Million Veteran Program. The phenotypes are represented on the x-axis and ordered by broader disease categories. The red line denotes the significance threshold using false discovery rate of 1% using the Benjamini-Hochberg procedure. The description of phenotypes is highlighted for the associations with FDR < 0.1 and odds ratio < 0.90 or odds ratio > 1.10. B) and D) A heatmap plot of SNPs with at least one significant association (FDR < 0.1). The direction of effect disease risk is represented by

580 odds ratio. A red color indicates increased risk and blue color indicated reduced risk. The results
581 with odds ratio < 0.90 or odds ratio > 1.10 are shown.

582 **Fig 3**. PheWAS results of candidate SNPs from GWAS of Hospitalized COVID-19 in AFR
583 ancestry individuals. The plot highlights the association between rs581342 SNP and
584 Neutropenia, which was only observed in the AFR ancestry. The phenotypes are represented on
585 the x-axis and ordered by broader disease categories. The red line denotes the significance
586 threshold using false discovery rate of 1% using the Benjamini-Hochberg procedure. The table
587 on the top right of the plot shows the association results between rs581342 and neutropenia in
588 other ancestries. The association was not tested among participants of ASN ancestry due to low
589 case numbers.

590

COVID-19 GWAS Summary Statistics

- GWAS summary statistics (*Freeze 6*) analyzed for PheWAS:
  - COVID-19 critically ill vs Population (A2)
  - COVID-19 hospitalized vs Population (B2)

Selection of Lead Variants

- SNPs were clumped and pruned using the 1000G reference panel
  - FDR = 0.01
  - R^2 threshold = 0.1
  - Distance threshold = 250 kb
- **35** lead loci were identified from critically ill COVID-19 vs population GWAS
- **42** lead loci identified from hospitalized COVID-19 vs population GWAS

MVP Dataset

Genotype Data

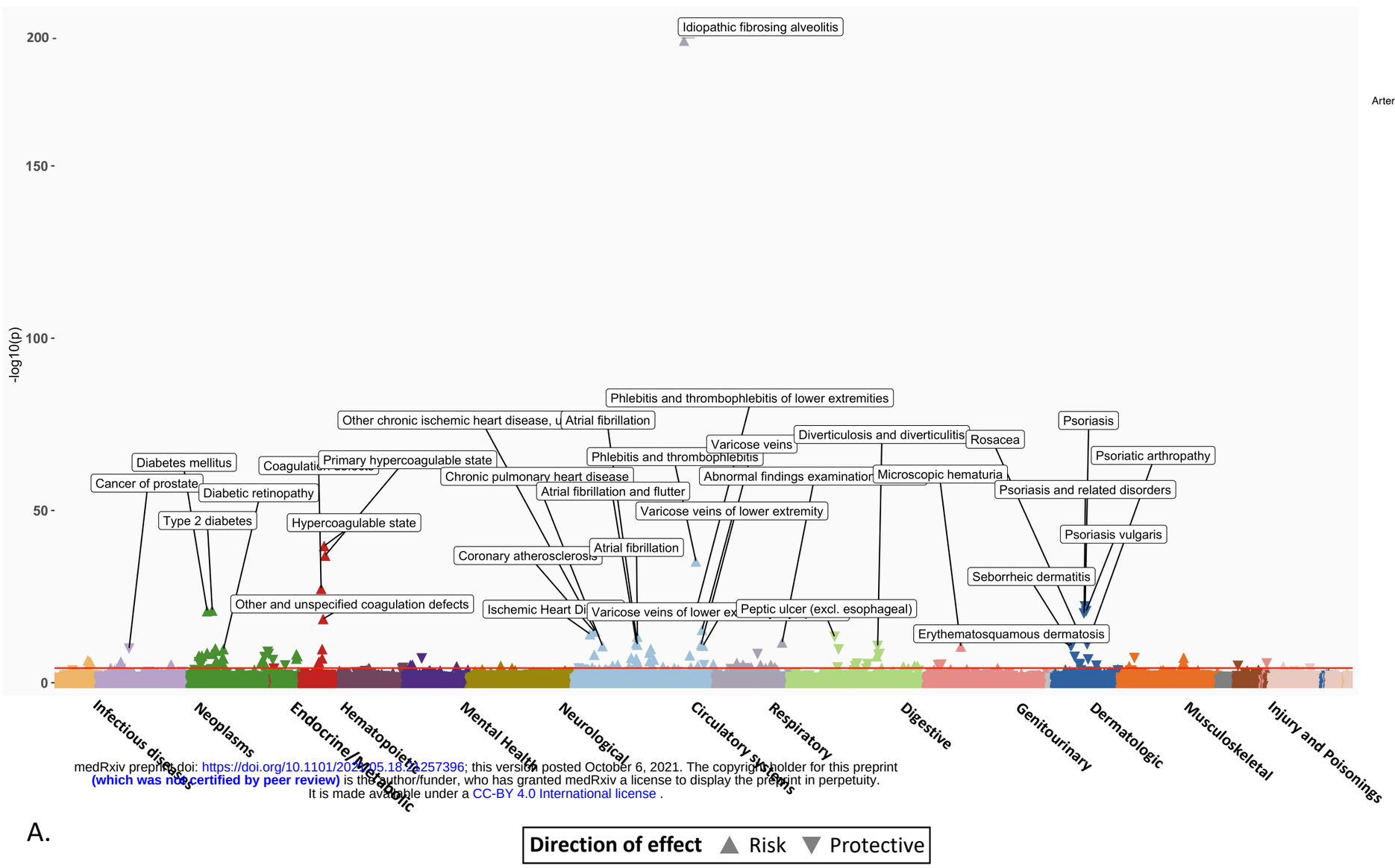
- 1000G imputed data
- Ancestry defined by HARE algorithm
- Total N=658,582
  - EUR = 464,961
  - AFR = 123,120
  - Hispanic = 52,183
  - Asian = 8,329

Phenotype Data

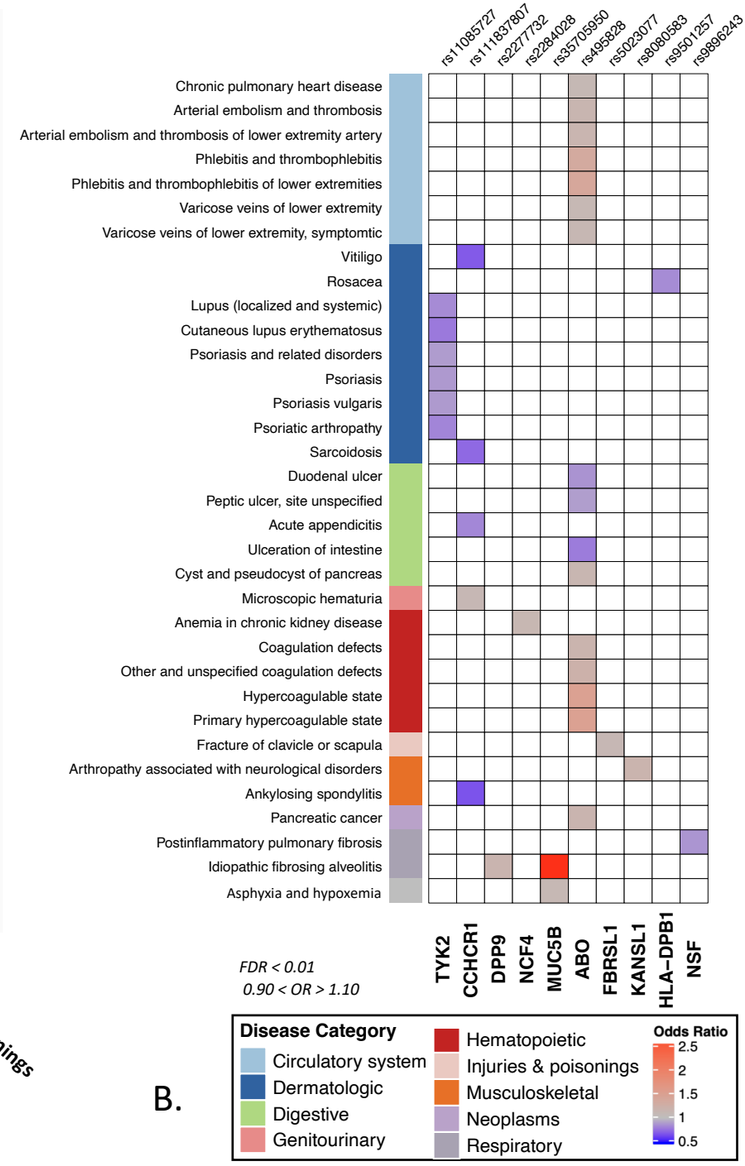
- ICD-9/10 codes mapped to phecodes
- Case-Control Definition:
  - Cases: >=2 phecodes
  - Controls: No phecodes
- Phecodes with cases < 200 were dropped
- Total phecodes : 1,688

PheWAS

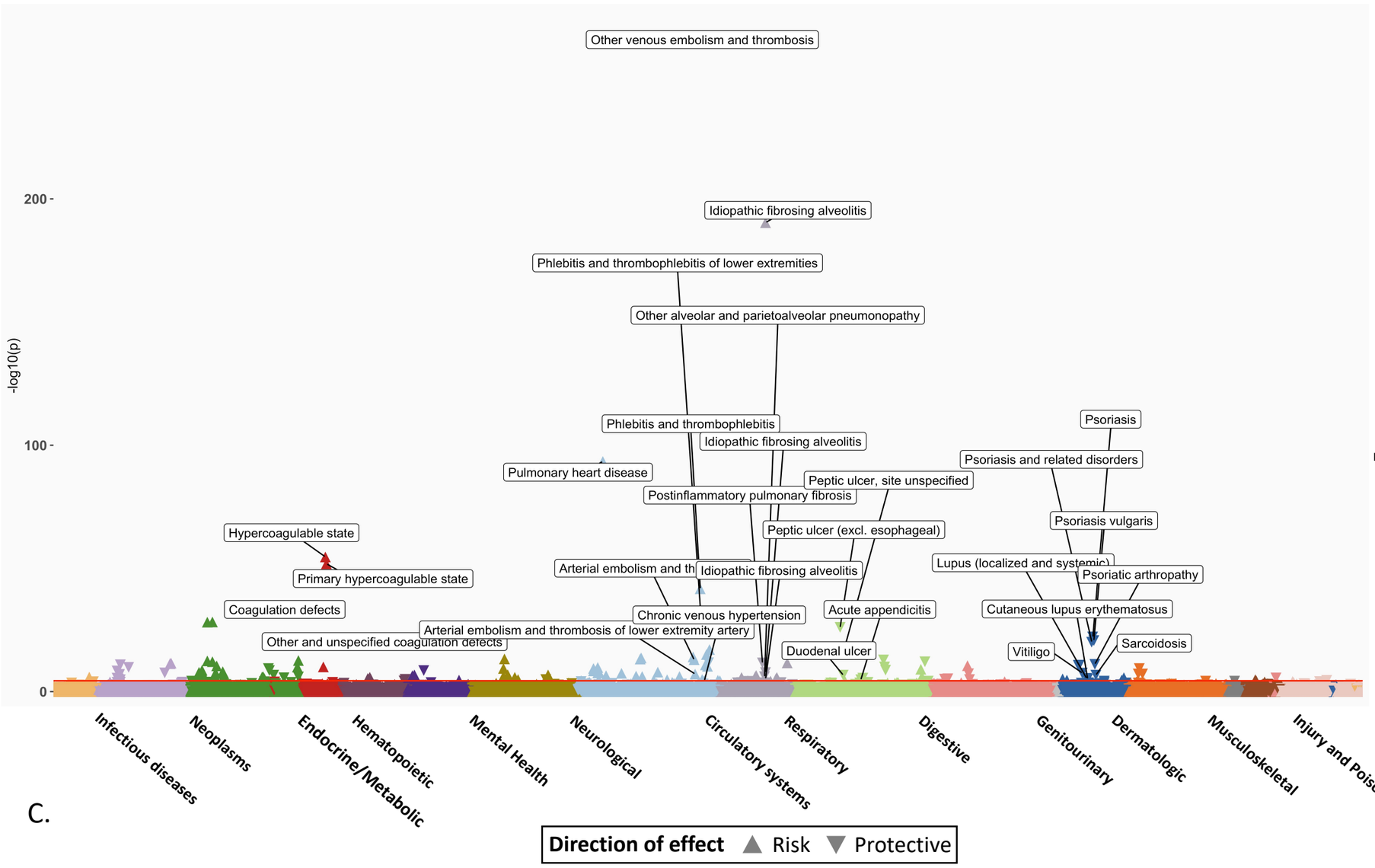
- Ancestry specific PheWAS
- Logistic or firth regression
- Covariates
  - 1st 20 PCs
  - Sex
  - Age
  - Age^2
- Transethnic Meta-analysis



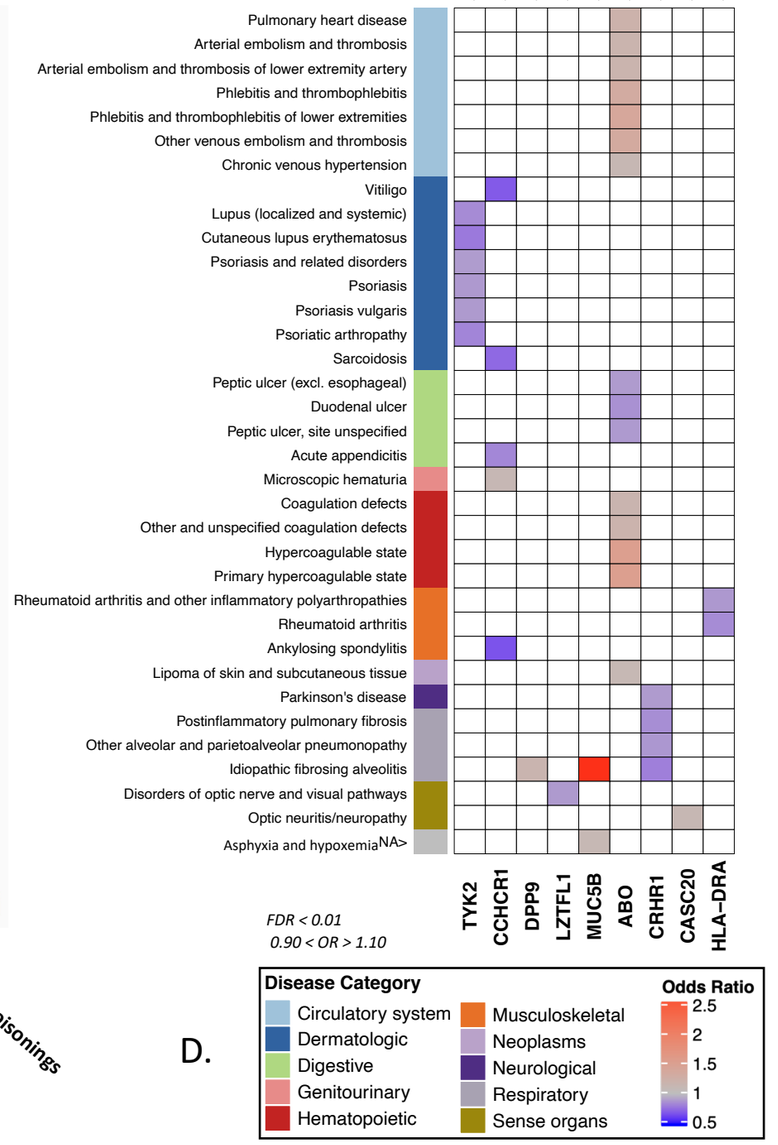
A.



B.



C.



D.

Ancestry	MAF	OR	P	Cases/Controls
AFR	0.56	1.29	3.1×10^{-13}	1,788/113,277
EUR	0.07	1.05	0.24	3,855/438,702
HIS	0.10	1.65	8.84×10^{-06}	308/49,162