Genetic association analysis of SARS-CoV-2 infection in 455,838 UK Biobank participants

J. A. Kosmicki[†], J. E. Horowitz[†], N. Banerjee, R. Lanche, A. Marcketta, E. Maxwell, Xiaodong Bai, D. Sun, J. Backman, D. Sharma, C. O'Dushlaine, A. Yadav, A. J. Mansfield, A. Li, J. Mbatchou, K. Watanabe, L. Gurski, S. McCarthy, A. Locke, S. Khalid, O. Chazara, Y. Huang, E. Kvikstad, A. Nadkar, A. O'Neill, P. Nioi, M. M. Parker, S. Petrovski, H. Runz, J. D. Szustakowski, Q. Wang, Regeneron Genetics Center*, UKB Exome Sequencing Consortium*, M. Jones, S. Balasubramanian, W. Salerno, A. Shuldiner, J. Marchini, J. Overton, L. Habegger, M. N. Cantor, J. Reid, A. Baras[‡], G. R. Abecasis[‡], M. A. Ferreira[‡]

From:

Regeneron Genetics Center, 777 Old Saw Mill River Rd., Tarrytown, NY 10591, USA (JK, JEH, NB, RL, AM, EM, XB, DS, JB, DSh, CO'D, AY, AJM, AL, JM, KW, LG, SM, AL, SK, MJ, SB, WS, AS, JM, JO, LH, MNC, JR, AB, GRA, MAF)

Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK (OC, AO'N, SP, QW)

Biogen, 300 Binney St, Cambridge, MA 02142, USA (YH, HR)

Alnylam Pharmaceuticals, 675 West Kendall St, Cambridge, MA 02142, USA (MMP, PN).

Bristol Myers Squibb, Route 206 and Province Line Road, Princeton, NJ 08543 (EK, AN, JDS)

*A complete list of investigators is provided in the Supplementary Appendix.

[†]J. A. Kosmicki and J. E. Horowitz contributed equally to this manuscript.

[‡]A. Baras, G. R. Abecasis and M. A. Ferreira jointly supervised this work.

Correspondence to: manuel.ferreira@regeneron.com and goncalo.abecasis@regeneron.com

This research has been conducted using the UK Biobank Resource (Project 26041)

ABSTRACT

Background. Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) causes Coronavirus disease-19 (COVID-19), a respiratory illness with influenza-like symptoms that can result in hospitalization or death. We investigated human genetic determinants of COVID-19 risk and severity in 455,838 UK Biobank participants, including 2,003 with COVID-19.

Methods. We defined eight COVID-19 phenotypes (including risks of infection, hospitalization and severe disease) and tested these for association with imputed and exome sequencing variants.

Results. We replicated prior COVID-19 genetic associations with common variants in the 3p21.31 (in *LZTFL1*) and 9q34.2 (in *ABO*) loci. The 3p21.31 locus (rs11385942) was associated with disease severity amongst COVID-19 cases (OR=2.2, $P=3\text{x}10^{-5}$), but not risk of SARS-CoV-2 infection without hospitalization (OR=0.89, $P=0.25$). We identified two loci associated with risk of infection at $P<5\text{x}10^{-8}$, including a missense variant that tags the ε4 haplotype in *APOE* (rs429358; OR=1.29, $P=9\text{x}10^{-9}$). The association with rs429358 was attenuated after adjusting for cardiovascular disease and Alzheimer's disease status (OR=1.15, $P=0.005$). Analyses of rare coding variants identified no significant associations overall, either exome-wide or with (i) 14 genes related to interferon signaling and reported to contain rare deleterious variants in severe COVID-19 patients; (ii) 36 genes located in the 3p21.31 and 9q34.2 GWAS risk loci; and (iii) 31 additional genes of immunologic relevance and/or therapeutic potential.

Conclusions. Our analyses corroborate the association with the 3p21.31 locus and highlight that there are no rare protein-coding variant associations with effect sizes detectable at current sample sizes. Our full analysis results are publicly available, providing a substrate for meta-analysis with results from other sequenced COVID-19 cases as they become available. Association results are available at https://rgc-covid19.regeneron.com .

INTRODUCTION

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was discovered in Wuhan, China in late 2019 [1] and causes coronavirus disease 2019 (COVID-19) [2]. COVID-19 symptoms range from flu-like symptoms such as fever, cough and headaches to respiratory failure, acute immune responses and death [3]. It is estimated that most infected individuals display few, if any, symptoms [4, 5]. As of October 2020, SARS-CoV-2 has been reported in >41 million individuals and to be associated with >1.1 million deaths worldwide. Known risk factors include male sex, older age, ancestry, obesity, cardiovascular and kidney disease, chronic obstructive pulmonary disease (COPD) and dementia [6-9], among others.

Studying host genetic variation among individuals infected with SARS-CoV-2 holds the potential to identify mechanisms that influence disease severity and outcomes. Akin to *IFNGR1*, *STAT1, TLR7* and other genetic immune deficiencies that predispose to early-onset severe infections [10-15], this information may help identify individuals at high risk of SARS-CoV-2 infection who should be prioritized for disease prevention strategies, including vaccination or monoclonal antibody treatments [16, 17]. Further, understanding host mechanisms that provide protection from SARS-CoV-2 infection or that modulate disease severity might guide the development of treatment efforts, in the same way that *CCR5* variation and HIV infection [18], or *FUT2* variation and infection by certain strains of norovirus [19], helped identify therapeutic strategies and targets for these diseases.

Since the start of the SARS-CoV-2 pandemic, host genetic analysis of common genetic variation among SARS-CoV-2 patients identified two genome-wide significant loci, one at 3p21.31 spanning at least six genes (including *SLC6A20* and *LZTFL1*) and a second at 9q34.2 in the *ABO* locus [20, 21]. The first locus has been consistently replicated in additional studies [21,

22], while the association at the *ABO* locus remains contentious. In addition to these genome-wide association studies (GWAS), two studies suggest that rare deleterious variants in genes related to interferon signaling may be implicated in more extreme clinical outcomes [14, 23]. However, to date, there has been no assessment of the contribution of rare genetic variation to COVID-19 disease susceptibility or severity through large population-based exome-wide association analyses.

To identify rare coding variants associated with COVID-19 susceptibility and severity, we evaluated clinical data derived from quantitative polymerase chain reaction (qPCR) tests for SARS-CoV-2, together with anonymized electronic health records and death registry data for both COVID-19 patients and other individuals in the UK Biobank study. We first analyzed imputed data for 455,838 individuals (2,003 with COVID-19), including a deep dive into the unequivocal 3p21.31 locus as a positive control, to calibrate our susceptibility and severity phenotypes with those used in other COVID-19 GWAS. We then analyzed exome sequencing data for a subset of 424,183 individuals (1,865 with COVID-19) to investigate disease associations with individual rare variants and rare variant-aggregated gene-burden tests. In addition to an agnostic exome-wide search for genetic risk factors, we also focused on 81 specific genes of interest (i) with a known role in interferon signaling and recently observed to contain rare deleterious variants in patients with severe COVID-19 [14, 23]; (ii) near two common risk variants for COVID-19 identified by GWAS [20]; or (iii) of immunologic relevance and/or therapeutic potential.

This study represents the largest exome-sequencing study of COVID-19 performed to date. Expanded analyses, particularly among individuals disproportionally affected by SARS-CoV-2, are essential to help identify human genetic determinants of disease risk and identify therapeutic avenues for the treatment of COVID-19.

METHODS

Study participants

We studied the host genetics of SARS-CoV-2 infection in participants of the UK Biobank study, which took place between 2006 and 2010 and includes approximately 500,000 adults aged 40-69 at recruitment [24]. In collaboration with UK health authorities, the UK Biobank has made available regular updates on COVID-19 status for all participants, including results from four main data types: qPCR test for SARS-CoV-2, anonymized electronic health records, primary care and death registry data. We report results based on the 12 September 2020 data refresh and excluded from the analysis 28,547 individuals with a death registry event prior to 2020.

COVID-19 phenotypes used for genetic association analyses

Using the data types outlined above, we grouped UK Biobank participants into three broad COVID-19 disease categories (**Table 1**): (i) positive – those with a positive qPCR test for SARS-CoV-2 or a COVID-19-related ICD10 code (U07), hospitalization or death; (ii) negative – those with only negative qPCR test results for SARS-CoV-2 and no COVID-19-related ICD10 code (U07), hospitalization or death; and (iii) unknown – those with no qPCR test result and no COVID-19-related ICD10 code (U07), hospitalization or death. We then used these broad COVID-19 disease categories, in addition to hospitalization and disease severity information, to create eight COVID-19-related phenotypes for genetic association analyses, as detailed in **Table 2**.

Array genotyping and imputation

DNA samples from participants of the UK Biobank study were genotyped as described previously

[24] using the Applied Biosystems UK BiLEVE Axiom Array (N=49,950) or the closely related Applied Biosystems UK Biobank Axiom Array (N=438,427). Genotype data for variants not included in the arrays were then inferred using three reference panels (Haplotype Reference Consortium, UK10K and 1000 Genomes Project phase 3) as described previously [24].

Exome sequencing

Sample Preparation and Sequencing. Genomic DNA samples normalized to approximately 16 ng/ul were transferred to the Regeneron Genetics Center from the UK Biobank in 0.5ml 2D matrix tubes (Thermo Fisher Scientific) and stored in an automated sample biobank (LiCONiC Instruments) at -80°C prior to sample preparation. Exome capture was completed using a high-throughput, fully-automated approach developed at the Regeneron Genetics Center. Briefly, DNA libraries were created by enzymatically shearing 100ng of genomic DNA to a mean fragment size of 200 base pairs using a custom NEBNext Ultra II FS DNA library prep kit (New England Biolabs) and a common Y-shaped adapter (Integrated DNA Technologies) was ligated to all DNA libraries. Unique, asymmetric 10 base pair barcodes were added to the DNA fragment during library amplification with KAPA HiFi polymerase (KAPA Biosystems) to facilitate multiplexed exome capture and sequencing. Equal amounts of sample were pooled prior to overnight exome capture, approximately 16 hours, with a slightly modified version of IDT's xGen probe library; supplemental probes were added to capture regions of the genome well-covered by a previous capture reagent (NimbleGen VCRome), but poorly covered by the standard xGen probes (design bed file available by request). Captured fragments were bound to streptavidin-coupled Dynabeads (Thermo Fisher Scientific) and non-specific DNA fragments removed through a series of stringent washes using the xGen Hybridization and Wash kit according to the manufacturer's recommended

protocol (Integrated DNA Technologies). The captured DNA was PCR amplified with KAPA HiFi and quantified by qPCR with a KAPA Library Quantification Kit (KAPA Biosystems). The multiplexed samples were pooled and then sequenced using 75 base pair paired-end reads with two 10 base pair index reads on the Illumina NovaSeq 6000 platform using S2 or S4 flow cells.

Variant calling and quality control. Sample read mapping and variant calling, aggregation and quality control were performed via the SPB protocol described in Van Hout et al. [25] (https://www.ukbiobank.ac.uk/wp-content/uploads/2019/08/UKB-50k-Exome-Sequencing-Data-Release-July-2019-FAQs.pdf). Briefly, for each sample, NovaSeq WES reads are mapped with BWA MEM to the hg38 reference genome. Small variants are identified with WeCall and reported as per-sample gVCFs. These gVCFs are aggregated with GLnexus into a joint-genotyped, multi-sample VCF (pVCF). SNV genotypes with read depth less than seven (DP < 7) and indel genotypes with read depth less than ten (DP < 10) are changed to no-call genotypes. After the application of the DP genotype filter, a variant-level allele balance filter is applied, retaining only variants that meet either of the following criteria: (i) at least one homozygous variant carrier or (ii) at least one heterozygous variant carrier with an allele balance greater than the cutoff (AB >= 0.15 for SNVs and AB >= 0.20 for indels).

Identification of low-quality variants from exome-sequencing using machine learning. Briefly, we defined a set of positive control and negative control variants based on: (i) concordance in genotype calls between array and exome sequencing data; (ii) mendelian inconsistencies in the exome sequencing data; (iii) differences in allele frequencies between exome sequencing batches; (iv) variant loadings on 20 principal components derived from the analysis of variants with a

MAF<1%. The model was then trained on 30 available WeCall/GLnexus site quality metrics, including, for example, allele balance and depth of coverage. We split the data into training (80%) and test (20%) sets. We then performed a grid search with 5-fold cross-validation on the training set and applied the model with highest accuracy to the test set. Out of 15 million variants in the exome target region, 1 million (6.5%) were identified as low-quality and excluded from the analysis. Similarly, we identified and removed 6 million out of 21 million variants (28.6%) in the buffer region.

Gene burden masks. Briefly, for each gene region as defined by Ensembl [26], genotype information from multiple rare coding variants was collapsed into a single burden genotype, such that individuals who were: (i) homozygous reference (Ref) for all variants in that gene were considered homozygous (RefRef); (ii) heterozygous for at least one variant in that gene were considered heterozygous (RefAlt); (iii) and only individuals that carried two copies of the alternative allele (Alt) of the same variant were considered homozygous for the alternative allele (AltAlt). We did not phase rare variants; compound heterozygotes, if present, were considered heterozygous (RefAlt). We did this separately for four classes of variants: (i) predicted loss of function (pLoF), which we refer to as an "M1" burden mask; (ii) pLoF or missense ("M2"); (iii) pLoF or missense variants predicted to be deleterious by 5/5 prediction algorithms ("M3"); (iv) pLoF or missense variants predicted to be deleterious by 1/5 prediction algorithms ("M4"). The five missense deleterious algorithms used were SIFT [27], PolyPhen2 (HDIV), PolyPhen2 (HVAR) [28], LRT [29], and MutationTaster [30]. For each gene, and for each of these four groups, we considered five separate burden masks, based on the frequency of the alternative allele of the variants that were screened in that group: <1%, <0.1%, <0.01%, <0.001% and singletons

only. Each burden mask was then tested for association with the same approach used for individual variants (see below).

Genetic association analyses

Association analyses in the UK Biobank study were performed using the Firth logistic regression test implemented in REGENIE [31], separately for variants derived from array-based imputation and exome sequencing. In this test, Firth's approach is applied when the p-value from the standard logistic regression score test is below 0.05. As the Firth penalty (*i.e.* Jeffrey's invariant prior) corresponds to a data augmentation procedure where each observation is split into a case and a control with different weights, it can handle variants with no minor alleles among cases. With no covariates, this corresponds to adding 0.5 in every cell of a 2x2 table of allele counts versus case-control status.

We included in step 1 of REGENIE (*i.e.* prediction of individual trait values based on the genetic data) variants that were directly genotyped, had a minor allele frequency (MAF) >1%, <10% missingness, Hardy-Weinberg equilibrium test *P*-value>10^{-15} and after linkage-disequilibrium (LD) pruning (1000 variant windows, 100 sliding windows and r^2<0.9). The association model used in step 2 of REGENIE included as covariates age, age^2, sex, age-by-sex, age^2-by-sex, and the first 10 ancestry-informative principle components (PCs) released by the UK Biobank. For the analysis of exome variants, we also included as covariates an indicator for exome sequencing batch and 20 PCs derived from the analysis of exome variants with a MAF between $2.6x10^{-5}$ (roughly corresponding to a minor allele count [MAC] of 20) and 1%. We did this because previous studies have found that PCs derived from common variants do not adequately correct for fine-scale population structure [32, 33].

For imputed variants, we retained association results for variants with both an imputation information score ≥0.3 and MAC ≥5, and either (i) MAF>0.5% or (ii) a protein-altering consequence (*i.e.* pLOF, missense or splice variants). For exome sequencing variants, we retained association results for variants with a MAC≥5. Association analyses were performed separately for three different ancestries defined based on the array data (African [AFR], European [EUR] and South Asian [SAS]), with results subsequently combined across ancestries using an inverse variance-weighed fixed-effects meta-analysis.

Results availability

All genotype-phenotype association results reported in this study are available for browsing using the RGC's COVID-19 Results Browser (https://rgc-covid19.regeneron.com). Data access and use is limited to research purposes in accordance with the Terms of Use (https://rgc-covid19.regeneron.com/terms-of-use). The COVID-19 Results Browser provides a user-friendly interface to explore genetic association results, enabling users to query summary statistics across multiple cohorts and association studies using genes, variants or phenotypes of interest. Results are displayed in an interactive tabular view ordered by p-value – enabling filtering, sorting, grouping and viewing additional statistics – with link outs to individual GWAS reports, including interactive Manhattan and QQ plots. LocusZoom views of LD information surrounding variants of interest are also available, with LD calculated using the respective source genetic datasets.

The data resource supporting the COVID-19 Results Browser is built using a processed version of the raw association analysis outputs. Using the RGC's data engineering toolkit based in Apache Spark and Project Glow (https://projectglow.io/), association results are annotated, enriched and partitioned into a distributed, columnar data store using Apache Parquet. Processed

Parquet files are registered with AWS Athena, enabling efficient, scalable queries on unfiltered association result datasets. Additionally, “filtered” views of associations significant at a threshold of p-value < 0.001 are stored in AWS RDS Aurora databases for low latency queries to service primary views of top associations. APIs into RDS and Athena are managed behind the scenes such that results with a p-value>0.001 are pulled from Athena as needed.

RESULTS

Demographics and health characteristics of study participants

Among 473,977 participants of the UK Biobank study who were alive in January 2020, 2,118 were COVID-19 positive, 16,331 were COVID-19 negative and 455,528 had unknown COVID-19 status (**Table 1**). Relative to participants who were COVID-19 negative or unknown (**Table 1**), COVID-19 positive individuals were more likely to be male, to have African or South Asian ancestry and to have cardiovascular or respiratory co-morbidities (**Table 3**). These co-morbidities were also observed in analyses stratified by ancestry group (**Table 4**).

Genome-wide association study (GWAS) of imputed variants

We performed ancestry-specific GWAS for eight COVID-19-related phenotypes, using imputed variants available for a subset of 455,838 individuals (**Table 5**). These phenotypes captured a spectrum of disease severity, from COVID-19 cases who did not require hospitalization to those with severe disease (respiratory support or death). Association results are publicly available at https://rgc-covid19.regeneron.com and main findings summarized below. The genomic inflation factor (λ_{GC}) was close to 1 for most analyses (**Supplementary Table 1**).

Association with variants reported in previous COVID-19 GWAS. Recently, Ellinghaus et al. [20] performed a GWAS comparing 1,610 cases with a PCR-positive test for SARS-CoV-2 and respiratory failure, against 2,205 controls with unknown SARS-CoV-2 status (mostly blood donors), all from Spain or Italy. Two loci reached genome-wide significance in that study: (i) 3p21.31, near the *LZTFL1* gene (rs11385942, OR=1.77 for the GA allele; 95% CI=1.48-2.11;

P=1.1x10^{-10}); and (ii) 9q34.2, near the *ABO* gene (rs657152, OR=1.39 for the A allele; 95% CI=1.20-1.47; P=4.9x10^{-8}) [20]. Both loci were recently replicated in a larger GWAS [21], with the former also replicated in a GWAS of severe COVID-19 patients in the UK [22]. We found a nominally significant and directionally consistent association with both variants in the European-specific analysis of the phenotype COVID-19 positive vs. COVID-19 negative or unknown (**Figure 1**). For the 3p21.31 locus (**Figure 1A**), we observed the largest effect with risks of hospitalization (OR=1.69; 95% CI=1.25-2.28; P=6x10^{-4}) and severe disease (OR=2.29; 95% CI=1.56-3.35; P=2x10^{-5}) amongst COVID-19 cases. In contrast, there was no association with the phenotype COVID-19 positive and not hospitalized vs. COVID-19 negative or unknown (OR=0.87; 95% CI=0.71-1.08; P=0.21). These results suggest that variants in this 3p21.31 locus influence COVID-19 severity and not risk of SARS-CoV-2 infection.

Significant associations with common variants in ancestry-specific GWAS. Across the eight phenotypes tested, we identified two loci with an association P<5x10^{-8}, both found in the European-specific analysis of the phenotype COVID-19 positive (N=1,797) vs. COVID-19 negative or unknown (N=434,038). The first locus was on chromosome 19q13.32; the lead variant was rs429358 (MAF=15%, OR=1.29, CI=1.18-1.40, P=8.9x10^{-9}), a common missense variant (Cys130Arg) that tags the epsilon (ε) 4 haplotype in *APOE* (**Figure 2A)**. This variant has established associations with both Alzheimer's disease (AD) and coronary artery disease (CAD). In addition, AD and CAD are known risk factors associated with COVID-19, and we observed an enrichment of both diseases amongst COVID-19 positive individuals (**Supplementary Table 2**). Therefore, we tested if the association between the *APOE* locus and susceptibility to COVID-19 could be confounded by AD or CAD case-control status. When both diseases were added as

covariates to the model, we found that the association with rs429358 was significantly attenuated (OR=1.15; 95% CI=1.04-1.26; *P*=0.005). These results suggest that the association between rs429358 in *APOE* and COVID-19 risk likely arose because of the enrichment of AD and CAD amongst COVID-19 cases.

The second locus was on chromosome 19p13.11, also associated with the phenotype COVID-19 positive vs. COVID-19 negative or unknown. The lead variant was rs117336466 (MAF=0.9%; OR=2.16, 95% CI=1.64-2.85, $P=4.5\text{x}10^{-8}$), located in the first intron of *TMEM161A* (**Figure 2B**). This variant was not associated with risks of hospitalization (OR=0.60, 95% CI=0.33-1.09, *P*=0.094) or severe disease (OR=0.58, 95% CI=0.27-1.24, *P*=0.161) amongst COVID-19 positive cases.

Genome-wide significant associations in trans-ancestry meta-analysis. Seven of the eight phenotypes were tested in two or more ancestries. For these, we combined results across ancestries using a fixed-effects meta-analysis, but no new loci were identified at $P<5\text{x}10^{-8}$.

Exome-wide association study of sequenced variants

We tested the association between the same eight COVID-19-related phenotypes and exome sequencing variants available for a subset of 424,183 individuals from the UKB study. We tested both single variants and a burden of rare variants in protein-coding genes (see Methods).

Exome-wide association results. The λ_{GC} for common variants (MAF>0.5%) was close to 1 for most analyses (**Supplementary Table 3**), while for rare variants (MAF<0.5%) we observed a considerable deflation of test statistics, caused by a large proportion of variants having a MAC of

0 in cases (*e.g.* 89% of variants in the European-only analysis of COVID-19 positive and hospitalized [N=1,065] vs. COVID-19 negative or unknown [N=403,700]). Overall, when considering both trans- and single-ancestry association analyses, we did not identify any associations with rare coding variants at a $P<5x10^{-8}$.

Association results for 14 genes in the anti-viral interferon signaling pathway. Two recent exome sequencing studies of COVID-19 suggested that rare deleterious variants in 14 genes related to interferon signaling may be implicated in more extreme clinical outcomes [14, 23]. Given our larger sample size, we examined whether there was any evidence for association between the COVID-19 hospitalization phenotype (1,184 cases vs. 422,318 controls) and a burden of rare (MAF<0.1%) pLoF variants (M1 burden test) or pLoF plus deleterious missense variants (M3 burden test) in these 14 genes. We found no nominal significant associations ($P<0.05$) with any of the 14 genes (**Table 6**). Further, these results were unchanged when testing COVID-19 severe cases (N=471), or when restricting the burden tests to include variants with a MAF<1% or singleton variants (**Supplementary Table 4**). Therefore, in our analysis of the UK Biobank data, we found no evidence for an association between the 14 specific interferon signaling genes and COVID-19 outcomes.

Association results for 36 genes located in two risk loci for COVID-19 identified by Ellinghaus et al. [20]. Associations with rare protein-coding variants might help pinpoint target genes of common risk variants identified in GWAS of COVID-19. To address this possibility, we focused on 36 protein-coding genes located within 500 kb of the two common risk variants identified by Ellinghaus et al. [20]: rs11385942 (locus 3p21.31) and rs657152 (locus 9q34.2). Of the 72 gene

burden tests performed (36 genes x 2 burden tests, considering variants with MAF<1%), four had a nominal significant association (**Supplementary Table 5**), including two protective (*CCR9* and *TSC1*) and two predisposing (*SARDH* and *XCR1*) associations. However, these associations did not remain significant after correcting for the number of tests performed (all with $P>0.05/72=0.0007$).

Association results for 31 additional genes of interest. Lastly, we performed the same analysis for 31 genes that are involved in the etiology of SARS-CoV-2 infection (*e.g. ACE2, TMPRSS2*), encode therapeutic targets (*e.g. IL6R, JAK2*) or have been implicated in other immune or infectious diseases through GWAS (*e.g. IL33*). After correcting for multiple testing, there were also no significant associations with a burden of rare deleterious variants for this group of genes (**Supplementary Table 6**).

DISCUSSION

Eleven months since the first reported cases of "pneumonia of unknown cause" to the World Health Organization and six months since the declaration of the COVID-19 pandemic [34], >41 million individuals have been infected with SARS-CoV-2 worldwide. Epidemiological studies have identified groups of individuals at high risk for severe disease, clinical complications and death [8, 9, 35-38]. More recently, studies focusing on host genetics have begun to identify common variants that contribute to heterogeneity in COVID-19 risk and severity [20-22].

Our analysis of COVID-19 in the UK Biobank indicates that, consistent with observational studies in the same UK participants [35, 37], COVID-19-related hospitalizations and deaths skew towards older, male individuals of non-European ancestry. Hypertension, obesity, CAD, type-2 diabetes and dementia are among the most frequently reported COVID-19 disease comorbidities [8, 9, 35]. Similarly, after adjusting for age, we observed a 1.7-fold enrichment in both cardiovascular disease and Alzheimer's disease among COVID-19 cases in the UK Biobank study.

Previous GWAS reported an association between risk of SARS-CoV-2 infection and common variants in the 3p21.31 locus [20-22]. We confirmed this association and further showed that this locus affects disease severity but not (or less so) risk of infection. We note, as have others, that the lead variant rs35652899 is in high LD with a lead expression quantitative trait locus (eQTL) for *SCL6A20* in lung tissue [39]. The *SLC6A20* gene encodes SIT1, a proline transporter expressed in the small intestine, lung, and kidney [40]. SIT1 expression and function is increased via interaction with angiotensin-converting enzyme 2 (ACE2), which is the SARS-CoV-2 receptor [41]. One intriguing hypothesis is that increased expression of *SLC6A20* in the gastrointestinal tract, lung or kidney might promote viral uptake, thus leading to increased risk of severe disease due to pathology in these tissues. Other candidate genes in the region include *LZTFL1*, which

366 encodes a cytoplasmic ciliary transport protein with expression in the lung and implicated in
367 recessive ciliopathies with renal dysfunction as one feature, *CXCR6* and *CCR9*, chemokine
368 receptors which mediate trafficking of T lymphocytes to the lung and GI tract, respectively, and
369 *XCR1* on plasmacytoid dendritic cells, which mediates antigen cross presentation, potentially
370 implicating dysregulation of immune cell trafficking and function in severe COVID-19, but further
371 work is required to attribute the purported biological mechanisms of these genes with SARS-CoV-
372 2 infection and disease progression of COVID-19.

373 Ellinghaus et al. [20] first reported an association between common variants in the *ABO*
374 locus and risk of SARS-CoV-2 infection. Furthermore, ABO blood groups have been associated
375 with severe COVID-19 [42, 43], with blood group A being associated with increased disease risk.
376 These observations raise the possibility that genes in the *ABO* locus play a role in COVID-19
377 susceptibility. However, genetic associations at the *ABO* locus can be confounded by population
378 stratification [44, 45]. Furthermore, the analysis reported by Ellinghaus et al [20] used blood
379 donors (which skew toward type O) as controls, which might have biased the association results
380 at the *ABO* locus. As such, it is important to determine if the association with the *ABO* locus is
381 reproducible in independent studies. First, we found no difference in representation of blood types
382 among COVID-19 cases and controls (not shown). Second, although we did observe a directionally
383 consistent and nominally significant association between risk of infection and the published lead
384 variant, when we combined results from the UK Biobank with those from the discovery cohort
385 [20], the association with this variant did not reach genome-wide significance (not shown). Third,
386 we found no evidence for an association between this locus and disease severity. Therefore, it
387 remains unclear whether variants in the *ABO* locus represent bona fide risk factors for COVID-19.

In our GWAS of imputed variants, we identified a genome-wide significant association between risk of SARS-CoV-2 infection and a variant that tags the ε4 haplotype in *APOE*. Common variants in *APOE* have been previously associated with SARS-CoV-2 infection, independent of CAD, dementia and other comorbidities [46]. However, in contrast to these findings, we found that the association with *APOE* was significantly attenuated after adjusting for AD and CAD. Similar results were obtained after conditioning on AD alone (not shown). This suggests that the observed association between risk of SARS-CoV-2 infection and *APOE* in our analysis of the UK Biobank was, at least partly, confounded with AD status.

We also identified a putative new association between common variants on chromosome 19p13.11 and risk of SARS-CoV-2 infection. However, this locus was not associated with increased risks of hospitalization or severe disease amongst COVID-19 positive individuals. Replication in independent studies is required to validate the association between 19p13.11 and risk of SARS-CoV-2 infection.

Lastly, we analyzed exome sequence data for a subset of 424,183 individuals in the UK Biobank to test the association between COVID-19 phenotypes and rare variants not captured by array genotyping or imputation. We found no associations at a $P<5x10^{-8}$ with pLoF variants, missense variants or in gene-burden analyses. We then concentrated on 81 genes of interest, including 14 genes related to interferon signaling [14, 23], 36 genes in two GWAS loci [20] and 31 additional genes of immunologic relevance and/or therapeutic potential. After correcting for the number of tests performed, there were no significant associations between the COVID-19 hospitalization phenotype and a burden of rare deleterious variants in any of these genes. We are expanding our analysis of exome sequence data to include additional studies and will update results accordingly.

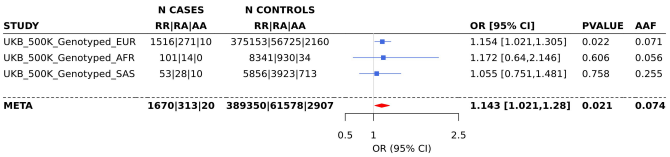
411

412 At the outset of the pandemic, testing for SARS-CoV-2 was restricted to symptomatic individuals
413 and often performed exclusively at inpatient/outpatient care sites. Thus, this current analysis is
414 likely weighted toward cases with demonstrable COVID-19 symptoms or clinical presentation.
415 Broader analysis of seropositive individuals who were asymptomatic or had infections mild
416 enough to resolve at home will be critical to identify genetic factors that might protect from severe
417 disease, particularly among high-risk groups with comorbidities. Regardless, further genetic
418 studies across ancestry groups will shed more light on human genetic risk factors associated with
419 susceptibility to SARS-CoV-2 and may point to pathways and approaches for the treatment of
420 COVID-19.

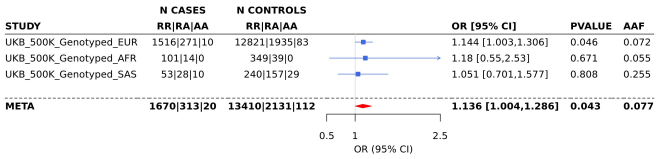
FIGURES

A. Locus 3p21.31 (rs11385942:GA, near the *LZTFL1* gene)

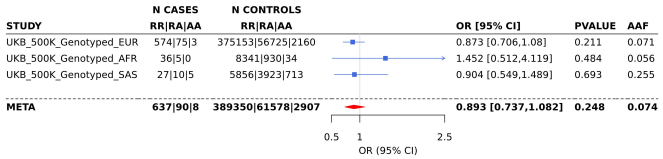
Covid-19 positive
vs.
Covid-19 negative or unknown



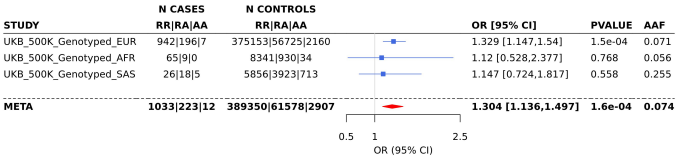
Covid-19 positive
vs.
Covid-19 negative



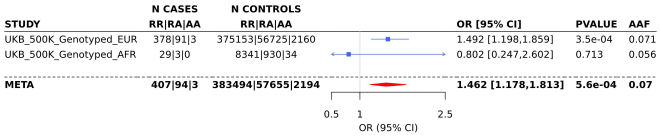
Covid-19 positive and not hospitalized
vs.
Covid-19 negative or unknown



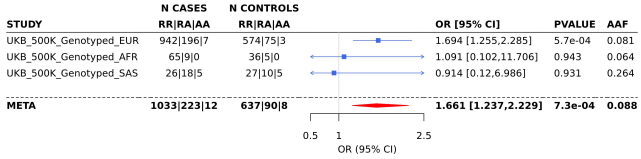
Covid-19 positive and hospitalized
vs.
Covid-19 negative or unknown



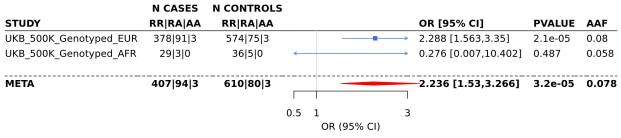
Covid-19 positive and severe
vs.
Covid-19 negative or unknown



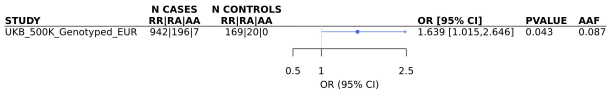
Covid-19 positive and hospitalized
vs.
Covid-19 positive and not hospitalized



Covid-19 positive and severe
vs.
Covid-19 positive and not hospitalized



Covid-19 positive and hospitalized
vs.
Covid-19 positive, not hospitalized, comorbidities and age>60



B. Locus 9q34.2 (rs8176719:TC, in the *ABO* gene)

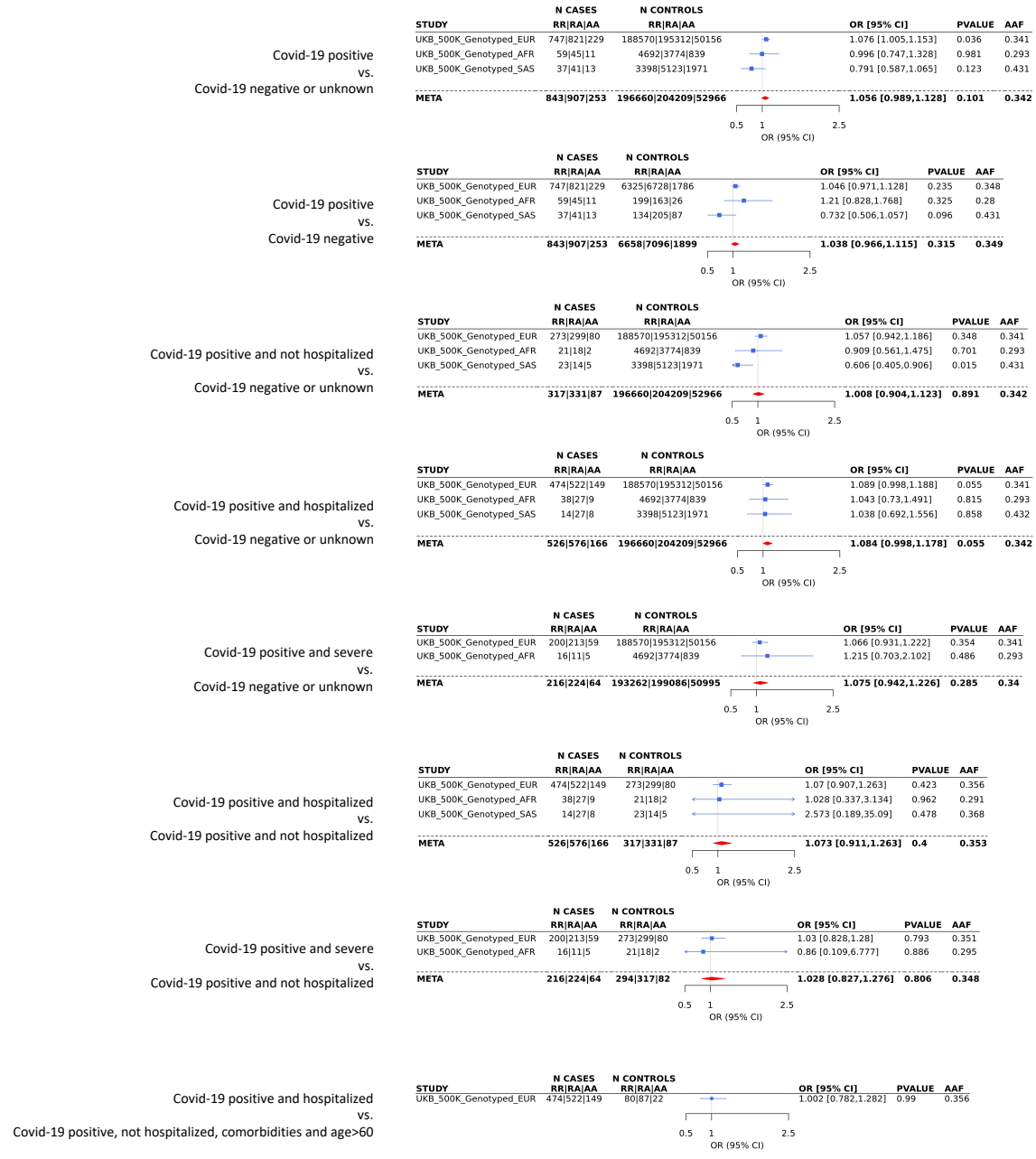
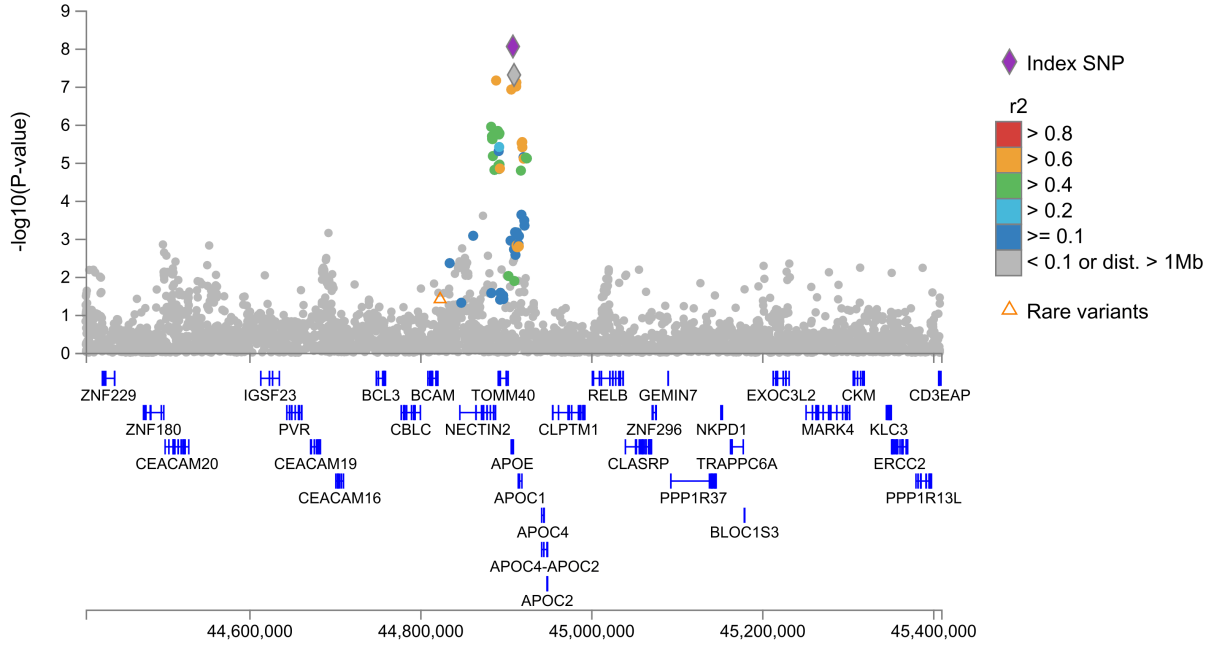


Figure 1. Results in the UKB for variants in two loci reported recently by Ellinghaus et al. [20] to associate with risk of hospitalization with severe COVID-19. For the chromosome 9 locus, we used rs8176719 as a proxy (r^2=0.94) for the lead variant reported by Ellinghaus et al. (rs657152).

A. Locus 19q13.32 (rs429358 in *APOE*)



B. Locus 19p13.11 (rs117336466 in *TMEM161A*)

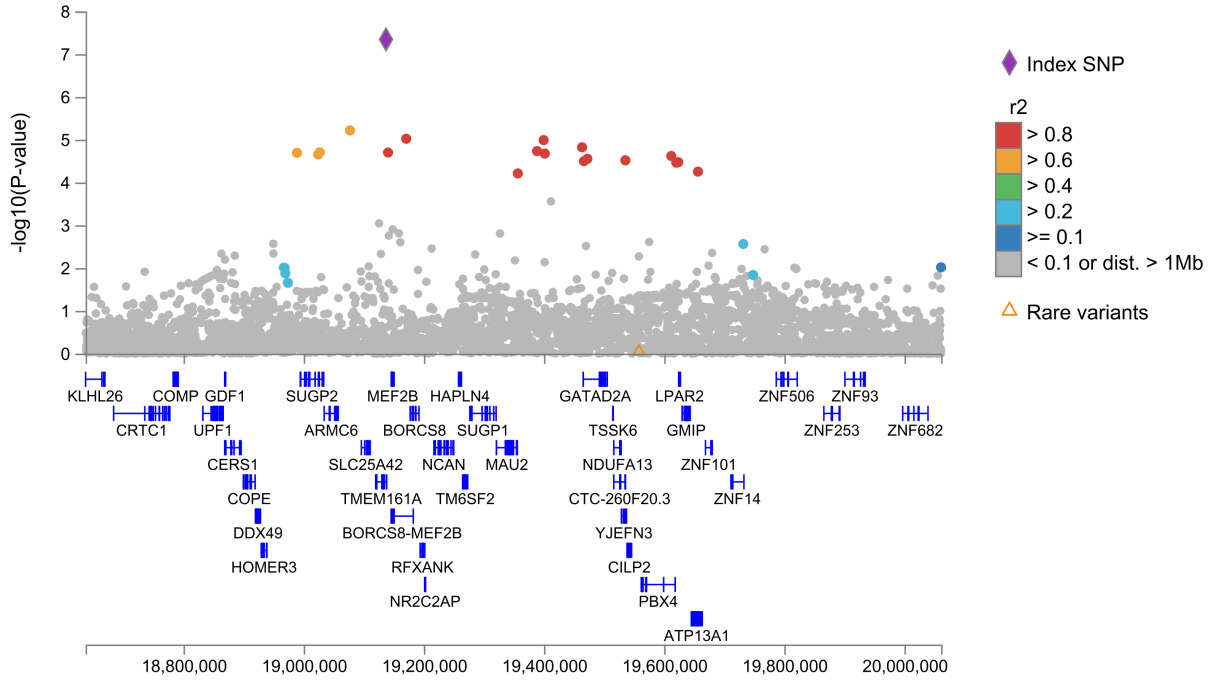


Figure 2. Regional association results for the three loci with common variants associated with COVID-19 phenotypes at $P<5x10^{-8}$. (A) *APOE* locus (lead variant rs429358), associated with COVID-19 positive vs. COVID-19 negative or unknown. (B) *TMEM161A* locus (lead variant rs117336466), associated with COVID-19 positive vs. COVID-19 negative or unknown. The lead variant in each locus is shown by the purple diamond. Linkage disequilibrium (LD) in these figures was estimated using genetic data from European individuals of the HapMap3 project.

TABLES

Table 1. Breakdown of COVID-19 status in participants of the UK Biobank study as of September 12, 2020.

COVID-19 status	Positive PCR for SARS-CoV-2	ICD10 U07 diagnosis	COVID-19-related death	Negative PCR test for SARS-CoV-2	N
Positive	Yes	Yes	Yes	-	251
	Yes	Yes	-	-	642
	Yes	-	Yes	-	42
	Yes	-	-	-	777
	-	Yes	Yes	-	16
	-	Yes	-	-	92
	-	-	Yes	-	122
	-	Yes	Yes	Yes	15
	-	Yes	-	Yes	150
	-	-	Yes	Yes	11
					Total = 2118
Negative	-	-	-	Yes	16331
Unknown	-	-	-	-	455528

Table 2. Criteria used to define COVID-19 phenotypes for genetic association analysis.

Phenotype	Case/ control group	COVID-19 status	Hospitalized	Severe disease*
COVID-19 positive vs. COVID-19 negative or unknown	Cases	Positive	-	-
	Controls	Negative or unknown	No or NA	No or NA
COVID-19 positive vs. COVID-19 negative	Cases	Positive	-	-
	Controls	Negative	No or NA	No or NA
COVID-19 positive and not hospitalized vs. COVID-19 negative or unknown	Cases	Positive	No	No
	Controls	Negative or unknown	No or NA	No or NA
COVID-19 positive and hospitalized vs. COVID-19 negative or unknown	Cases	Positive	Yes (or death**)	-
	Controls	Negative or unknown	No or NA	No or NA
COVID-19 positive and severe vs. COVID-19 negative or unknown	Cases	Positive	-	Yes
	Controls	Negative or unknown	No or NA	No or NA
COVID-19 positive and hospitalized vs. COVID-19 positive and not hospitalized	Cases	Positive	Yes (or death**)	-
	Controls	Positive	No	No
COVID-19 positive and severe vs. COVID-19 positive and not hospitalized	Cases	Positive	-	Yes
	Controls	Positive	No	No
COVID-19 positive and hospitalized vs. COVID-19 positive, not hospitalized, comorbidities and age>60	Cases	Positive	Yes (or death**)	-
	Controls	Positive with co-morbidities and age>60	No	No

* Severe disease: respiratory support (oxygen, ventilation) or death. ** A total of 175 individuals had a record of death due to COVID-19 but had no record of hospitalization. These individuals were included as cases. A hyphen ("-") indicates that the variable (*i.e.* Hospitalized and Severe Disease) was not considered as inclusion criteria.

453 **Table 3.** Demographics for participants of the UK Biobank study included in the analysis.

Demographic and clinical characteristics	**COVID-19 positive**	**COVID-19 negative**	**COVID-19 unknown**
N	2118	16331	455528
Female, n (%)	995 (46.9)	8547 (52.3)	252361 (55.4)
Median age at assessment, years (95% CI)	59 (51,67)	60 (53-66)	57 (50, 63)
Median body mass index, kg/m^2 (95% CI)	28 (24, 31)	27 (24, 30)	26 (23, 29)
Median C-reactive protein levels (95% CI)	1.6 (0.35, 2.9)	1.5 (0.3, 2.7)	1.3 (0.3, 0.3)
Number of current/past smokers, n (%)	1098 (51.8)	8228 (50.3)	199923 (43.9)
Median number of inpatient ICD10 3D codes (95% CI)	12 (3, 21)	12 (3, 20)	8 (2, 14)
Median Townsend deprivation index (95% CI)	-1.14 (-3.7, 1.46)	-1.9 (-4.2, 0.4)	-2.17 (-4.2, -0.12)
Ancestry			
African, n (%)	115 (5.4)	388 (2.4)	8921 (1.9)
East Asian, n (%)	17 (0.8)	65 (0.4)	2213 (0.4)
South Asian, n (%)	92 (4.3)	428 (2.6)	10125 (2.2)
European, n (%)	1798 (84.9)	14864 (91.0)	420007 (92.2)
Co-morbidities			
Hypertension, n (%)	1169 (55.2)	8917 (54.6)	200277 (43.9)
Coronary Disease, n (%)	229 (10.8)	1743 (10.6)	26015 (5.7)
Heart Failure, n (%)	100 (4.7)	563 (3.4)	5317 (1.1)
Type 2 Diabetes, n (%)	320 (15.1)	2013 (12.3)	30581 (6.7)
Chronic kidney disease, n (%)	92 (4.3)	573 (3.5)	6615 (1.4)
Asthma, n (%)	340 (16.0)	2895 (17.7)	64315 (14.1)
COPD, n (%)	159 (7.5)	1042 (6.4)	10661 (2.3)
Alzheimer's disease, n (%)	42 (1.9)	69 (0.42)	359 (0.07)

454

455 **Table 4.** Prevalence of co-morbidities, stratified by ancestry.

Disease	COVID-19 positive			Covid-19 negative or unknown
	All	Hospitalized*	Not hospitalized	
African ancestry				
Total N	115	74	41	9309
Hypertension, n (%)	76 (66.1)	51 (68.9)	25 (60.9)	4561 (48.9)
Coronary Disease, n (%)	5 (4.4)	4 (5.4)	1 (2.4)	329 (3.5)
Heart Failure, n (%)	6 (5.2)	6 (8.1)	0	98 (1.1)
Type 2 Diabetes, n (%)	31 (26.9)	26 (35.1)	5 (12.2)	1325 (14.2)
Chronic kidney disease, n (%)	4 (3.4)	4 (5.4)	0	179 (1.9)
Asthma, n (%)	16 (13.9)	10 (13.5)	6 (14.6)	1332 (14.3)
COPD, n (%)	3 (2.6)	3 (4.1)	0	98 (1.1)
Alzheimer's disease, n (%)	0	0	0	7 (0.07)
European ancestry				
Total N	1798	1145	653	434871
Hypertension, n (%)	993 (55.2)	704 (61.4)	289 (44.2)	191906 (44.1)
Coronary Disease, n (%)	205 (11.4)	166 (14.5)	39 (5.9)	25512 (5.8)
Heart Failure, n (%)	85 (4.7)	75 (6.5)	10 (1.5)	5396 (1.2)
Type 2 Diabetes, n (%)	267 (14.8)	211 (18.4)	56 (8.5)	27611 (6.3)
Chronic kidney disease, n (%)	78 (4.3)	64 (5.6)	14 (2.1)	6563 (1.5)
Asthma, n (%)	295 (16.4)	189 (16.5)	106 (16.2)	61755 (14.2)
COPD, n (%)	151 (8.4)	131 (11.4)	20 (3.1)	10949 (2.5)
Alzheimer's disease, n (%)	41 (2.3)	26 (2.7)	15 (2.3)	385 (0.08)
South Asian ancestry				
Total N	92	50	42	10553
Hypertension, n (%)	47 (51.1)	26 (52)	21 (50)	5149 (48.7)
Coronary Disease, n (%)	10 (10.8)	6 (12)	4 (9.5)	994 (9.4)
Heart Failure, n (%)	5 (5.4)	3 (6)	2 (4.7)	167 (1.6)
Type 2 Diabetes, n (%)	25 (27.1)	16 (32)	9 (21.4)	2308 (21.9)
Chronic kidney disease, n (%)	6 (6.5)	6 (12)	0	202 (1.9)
Asthma, n (%)	17 (18.4)	10 (20)	7 (16.6)	1691 (16.0)
COPD, n (%)	3 (3.2)	2 (4)	1 (2.3)	185 (1.7)
Alzheimer's disease, n (%)	0	0	0	9 (0.08)

456 *A total of 175 individuals had a record of death due to COVID-19 but had no record of hospitalization. These individuals were

457 included in the "Hospitalized" group in this analysis.

Table 5. Case-control sample size for eight COVID-19-related phenotypes tested in genetic association analyses in the UK Biobank study.

Phenotype	Ancestry	N cases		N controls	
		Imputed	Exome	Imputed	Exome
COVID-19 positive vs. COVID-19 negative or unknown	Combined	2003	1865	453,835	422,318
	AFR	115	110	9305	8599
	EUR	1797	1673	434,038	404,300
	SAS	91	82	10,492	9419
COVID-19 positive vs. COVID-19 negative	Combined	2003	1865	15,653	14,519
	AFR	115	110	388	361
	EUR	1797	1673	14,839	13,765
	SAS	91	82	426	393
COVID-19 positive and not hospitalized vs. COVID-19 negative or unknown	Combined	734	681	453,835	422,318
	AFR	41	39	9305	8599
	EUR	652	605	434,038	404,300
	SAS	42	37	10,492	9419
COVID-19 positive and hospitalized vs. COVID-19 negative or unknown	Combined	1268	1184	453,835	422,318
	AFR	74	71	9305	8599
	EUR	1145	1068	434,038	404,300
	SAS	49	45	10,492	9419
COVID-19 positive and severe vs. COVID-19 negative or unknown	Combined	-	471	443,343	412,899
	AFR	32	32	9305	8599
	EUR	-	439	434,038	404,300
COVID-19 positive and hospitalized vs. COVID-19 positive and not hospitalized	Combined	1268	1068	735	605
	AFR	74	-	41	-
	EUR	1145	1068	652	605
	SAS	49	-	42	-
COVID-19 positive and severe vs. COVID-19 positive and not hospitalized	Combined	504	439	693	605
	AFR	32	-	41	-
	EUR	472	439	652	605
COVID-19 positive and hospitalized vs. COVID-19 positive, not hospitalized, comorbidities and age>60	EUR	1145	1068	189	174

SAS case and control sample sizes for both severe COVID-19 sample sizes fell below the minimum case threshold to properly analyze. Dashes (-) in cells indicate the sample sizes were also too small for analysis.

Table 6. Association between the phenotype COVID-19 positive and hospitalized (N=1,184) vs COVID-19 negative or unknown (N=422,318) and 14 genes related to interferon signaling that were recently reported to contain rare (MAF<0.1%), deleterious variants in patients with severe COVID-19 [14, 23].

Gene	Burden test	Odds Ratio [95% CI]	P-value	Cases RR|RA|AA	Controls RR|RA|AA	AAF
STAT2	M3	2.34 (1.0, 5.5)	0.050	1176|8|0	421086|1231|1	1.47E-03
UNC93B1	M3	2.07 (0.93, 4.58)	0.073	1177|7|0	420917|1401|0	1.66E-03
IRF7	M3	2.15 (0.8, 5.74)	0.129	1179|5|0	421275|1043|0	1.24E-03
IFNAR1	M3	0.37 (0.08, 1.68)	0.195	1184|0|0	421811|507|0	5.99E-04
UNC93B1	M1	3.4 (0.43, 26.93)	0.247	1066|2|0	403926|374|0	4.64E-04
IRF7	M1	2.07 (0.6, 7.12)	0.247	1181|3|0	421674|644|0	7.64E-04
TLR7	M3	0.52 (0.15, 1.77)	0.297	1184|0|0	421856|334|128	6.97E-04
IFNAR1	M1	0.37 (0.04, 3.77)	0.404	1139|0|0	412664|235|0	2.84E-04
TLR3	M1	0.37 (0.02, 6.06)	0.488	1184|0|0	422141|177|0	2.09E-04
STAT1	M3	0.37 (0.02, 7.77)	0.520	1184|0|0	422137|181|0	2.14E-04
STAT2	M1	0.36 (0.01, 8.76)	0.529	1068|0|0	404195|105|0	1.30E-04
TRAF3	M3	0.37 (0.01, 10.13)	0.553	1139|0|0	412750|149|0	1.80E-04
IRF3	M1	1.5 (0.29, 7.66)	0.625	1182|2|0	422001|317|0	3.77E-04
TICAM1	M1	0.37 (0.0, 28.86)	0.652	1068|0|0	404205|95|0	1.17E-04
TICAM1	M3	0.37 (0.0, 28.86)	0.652	1068|0|0	404205|95|0	1.17E-04
TBK1	M1	0.36 (0.0, 31.85)	0.658	1068|0|0	404231|69|0	8.51E-05
IKBKG	M3	0.43 (0.0, 64.35)	0.744	1113|0|0	413640|70|9	1.06E-04
STAT1	M1	0.37 (0.0, 375.39)	0.776	1068|0|0	404268|32|0	3.95E-05
TBK1	M3	0.84 (0.24, 2.97)	0.787	1182|2|0	421450|867|1	1.03E-03
TLR3	M3	0.9 (0.39, 2.06)	0.803	1179|5|0	420322|1995|1	2.36E-03
IRF9	M1	0.37 (0.0, 1526.67)	0.813	1068|0|0	404275|25|0	3.08E-05
IRF9	M3	0.37 (0.0, 1526.67)	0.813	1068|0|0	404275|25|0	3.08E-05
IKBKG	M1	0.49 (0.0, 369.27)	0.834	1113|0|0	413679|31|9	5.91E-05
TRAF3	M1	0.37 (0.0, 10210.6)	0.847	1068|0|0	404284|16|0	1.97E-05
TLR7	M1	0.51 (0.0, 617.09)	0.851	1068|0|0	404282|12|6	2.96E-05
IRF3	M3	1.12 (0.26, 4.83)	0.874	1182|2|0	421885|433|0	5.14E-04
IFNAR2	M1	1.04 (0.14, 7.7)	0.968	1183|1|0	421965|353|0	4.18E-04
IFNAR2	M3	1.01 (0.14, 7.18)	0.996	1183|1|0	421949|369|0	4.37E-04

RR: individuals who were homozygote for the reference allele for all variants included in the burden test. RA: individuals who were heterozygote for at least one variant included in the burden test. AA: individuals who were homozygote for the alternative allele for at least one variant included in the burden test. The genes *TLR7* and *IKBKG* are located on the X chromosome; individuals counted as homozygote for the alternative allele include hemizygous males.

* M1: burden of rare (MAF<0.1%) pLoF variants. M3: burden of rare (MAF<0.1%) pLoF or missense variants that are predicted to be deleterious by 5 out of 5 algorithms.

SUPPLEMENTARY TABLES

Supplementary Tables 1 to 6 are provided in a separate document.

Supplementary Table 1. Genomic inflation factor (λ_{GC}) observed in the analysis of imputed variants for each of the eight phenotypes tested.

Supplementary Table 2. Association between COVID-19 phenotypes and both cardiovascular disease and Alzheimer's disease.

Supplementary Table 3. Genomic inflation factor (λ_{GC}) observed in the analysis of exome sequence variants for each of the eight phenotypes tested.

Supplementary Table 4. Results from burden association tests for 14 genes related to interferon signaling and recently reported to contain rare (MAF<0.1%), deleterious variants in patients with severe COVID-19 [14, 23].

Supplementary Table 5. Association between the phenotype COVID-19 positive and hospitalized (N=1,184) vs COVID-19 negative or unknown (N=422,318) and 36 genes located in two loci identified in a previous GWAS of severe COVID-19 [20].

Supplementary Table 6. Association between the phenotype COVID-19 positive and hospitalized (N=1,184) vs COVID-19 negative or unknown (N=422,318) and 31 genes that are

501 involved in the etiology of SARS-CoV-2, encode therapeutic targets or have been implicated in
502 other immune or infectious diseases through GWAS.

References

1. Zhu, N., et al., *A Novel Coronavirus from Patients with Pneumonia in China, 2019.* New England Journal of Medicine, 2020. **382**(8): p. 727-733.
2. Coronaviridae Study Group of the International Committee on Taxonomy of, V., *The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2.* Nat Microbiol, 2020. **5**(4): p. 536-544.
3. Guan, W.J., et al., *Clinical Characteristics of Coronavirus Disease 2019 in China.* N Engl J Med, 2020. **382**(18): p. 1708-1720.
4. Kimball, A., et al., *Asymptomatic and Presymptomatic SARS-CoV-2 Infections in Residents of a Long-Term Care Skilled Nursing Facility - King County, Washington, March 2020.* MMWR Morb Mortal Wkly Rep, 2020. **69**(13): p. 377-381.
5. Bai, Y., et al., *Presumed Asymptomatic Carrier Transmission of COVID-19.* JAMA, 2020. **323**(14): p. 1406-1407.
6. Richardson, S., et al., *Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area.* Jama, 2020. **323**(20): p. 2052-2059.
7. Atkins, J.L., et al., *PREEXISTING COMORBIDITIES PREDICTING SEVERE COVID-19 IN OLDER ADULTS IN THE UK BIOBANK COMMUNITY COHORT.* medRxiv, 2020: p. 2020.05.06.20092700.
8. Zhou, F., et al., *Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study.* Lancet, 2020. **395**(10229): p. 1054-1062.

9. Cummings, M.J., et al., *Epidemiology, clinical course, and outcomes of critically ill adults with COVID-19 in New York City: a prospective cohort study.* Lancet, 2020. **395**(10239): p. 1763-1770.
10. Ciancanelli, M.J., et al., *Infectious disease. Life-threatening influenza and impaired interferon amplification in human IRF7 deficiency.* Science, 2015. **348**(6233): p. 448-53.
11. Casanova, J.L., *Severe infectious diseases of childhood as monogenic inborn errors of immunity.* Proc Natl Acad Sci U S A, 2015. **112**(51): p. E7128-37.
12. Dupuis, S., et al., *Impairment of mycobacterial but not viral immunity by a germline human STAT1 mutation.* Science, 2001. **293**(5528): p. 300-3.
13. Jouanguy, E., et al., *Interferon-gamma-receptor deficiency in an infant with fatal bacille Calmette-Guerin infection.* N Engl J Med, 1996. **335**(26): p. 1956-61.
14. van der Made, C.I., et al., *Presence of Genetic Variants Among Young Men With Severe COVID-19.* JAMA, 2020. **324**(7): p. 663-673.
15. Zhang, S.Y., et al., *Severe COVID-19 in the young and healthy: monogenic inborn errors of immunity?* Nat Rev Immunol, 2020. **20**(8): p. 455-456.
16. Hansen, J., et al., *Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody cocktail.* Science, 2020. **369**(6506): p. 1010-1014.
17. Baum, A., et al., *Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies.* Science, 2020. **369**(6506): p. 1014-1018.
18. Samson, M., et al., *Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene.* Nature, 1996. **382**(6593): p. 722-5.

19. Thorven, M., et al., *A homozygous nonsense mutation (428G-->A) in the human secretor (FUT2) gene provides resistance to symptomatic norovirus (GGII) infections.* J Virol, 2005. **79**(24): p. 15351-5.
20. Ellinghaus, D., et al., *Genomewide Association Study of Severe Covid-19 with Respiratory Failure.* New England Journal of Medicine, 2020.
21. Shelton, J.F., et al., *Trans-ethnic analysis reveals genetic and non-genetic associations with COVID-19 susceptibility and severity.* medRxiv, 2020: p. 2020.09.04.20188318.
22. Pairo-Castineira, E., et al., *Genetic mechanisms of critical illness in Covid-19.* medRxiv, 2020: p. 2020.09.24.20200048.
23. Zhang, Q., et al., *Inborn errors of type I IFN immunity in patients with life-threatening COVID-19.* Science, 2020. **370**(6515): p. eabd4570.
24. Bycroft, C., et al., *The UK Biobank resource with deep phenotyping and genomic data.* Nature, 2018. **562**(7726): p. 203-209.
25. Van Hout, C.V., et al., *Exome sequencing and characterization of 49,960 individuals in the UK Biobank.* Nature, 2020.
26. Zerbino, D.R., et al., *Ensembl 2018.* Nucleic Acids Research, 2017. **46**(D1): p. D754-D761.
27. Vaser, R., et al., *SIFT missense predictions for genomes.* Nat Protoc, 2016. **11**(1): p. 1-9.
28. Adzhubei, I., D.M. Jordan, and S.R. Sunyaev, *Predicting functional effect of human missense mutations using PolyPhen-2.* Curr Protoc Hum Genet, 2013. **7**(1): p. 7.20.1-7.20.41.
29. Chun, S. and J.C. Fay, *Identification of deleterious mutations within three human genomes.* Genome research, 2009. **19**(9): p. 1553-1561.

30. Schwarz, J.M., et al., *MutationTaster evaluates disease-causing potential of sequence alterations.* Nat Methods, 2010. **7**(8): p. 575-6.
31. Mbatchou, J., et al., *Computationally efficient whole genome regression for quantitative and binary traits.* bioRxiv, 2020: p. 2020.06.19.162354.
32. Mathieson, I. and G. McVean, *Differential confounding of rare and common variants in spatially structured populations.* Nature Genetics, 2012. **44**(3): p. 243-246.
33. Zaidi, A.A. and I. Mathieson, *Demographic history impacts stratification in polygenic scores.* bioRxiv, 2020: p. 2020.07.20.212530.
34. *W.H.O. Rolling updates on coronavirus disease (COVID-19)*. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen.
35. Chudasama, Y.V., et al., *Multimorbidity and SARS-CoV-2 infection in UK Biobank.* Diabetes Metab Syndr, 2020. **14**(5): p. 775-776.
36. Gold, J.A.W., et al., *Characteristics and Clinical Outcomes of Adult Patients Hospitalized with COVID-19 - Georgia, March 2020.* MMWR Morb Mortal Wkly Rep, 2020. **69**(18): p. 545-550.
37. Niedzwiedz, C.L., et al., *Ethnic and socioeconomic differences in SARS-CoV-2 infection: prospective cohort study using UK Biobank.* BMC Med, 2020. **18**(1): p. 160.
38. Price-Haywood, E.G., et al., *Hospitalization and Mortality among Black Patients and White Patients with Covid-19.* N Engl J Med, 2020. **382**(26): p. 2534-2543.
39. Hao, K., et al., *Lung eQTLs to help reveal the molecular underpinnings of asthma.* PLoS Genet, 2012. **8**(11): p. e1003029.

40. Lonsdale, J., et al., *The Genotype-Tissue Expression (GTEx) project.* Nature Genetics, 2013. **45**(6): p. 580-585.
41. Vuille-dit-Bille, R.N., et al., *Human intestine luminal ACE2 and amino acid transporter expression increased by ACE-inhibitors.* Amino Acids, 2015. **47**(4): p. 693-705.
42. Zhao, J., et al., *Relationship between the ABO Blood Group and the COVID-19 Susceptibility.* Clin Infect Dis, 2020.
43. Li, J., et al., *Association between ABO blood groups and risk of SARS-CoV-2 pneumonia.* Br J Haematol, 2020. **190**(1): p. 24-27.
44. Thomson, G. and W.F. Bodmer, *Letter: Population stratification as an explanation of IQ and ABO association.* Nature, 1975. **254**(5498): p. 363-4.
45. Mourant, A.E., K. Domaniewska-Sobczak, and A.C. Kopec, *The distribution of the human blood groups and other polymorphisms*. 2nd ed. ed. 1976: London : Oxford university press.
46. Kuo, C.L., et al., *APOE e4 genotype predicts severe COVID-19 in the UK Biobank community cohort.* J Gerontol A Biol Sci Med Sci, 2020.

SUPPLEMENTARY TEXT

Regeneron Genetics Center (RGC) Research Team and Contribution Statements

All authors/contributors are listed in alphabetical order.

RGC Management and Leadership Team

Goncalo Abecasis, Ph.D., Aris Baras, M.D., Michael Cantor, M.D., Giovanni Coppola, M.D., Aris Economides, Ph.D., Luca A. Lotta, M.D., Ph.D., John D. Overton, Ph.D., Jeffrey G. Reid, Ph.D., Alan Shuldiner, M.D.

Contribution: All authors contributed to securing funding, study design and oversight. All authors reviewed the final version of the manuscript.

Sequencing and Lab Operations

Christina Beechert, Caitlin Forsythe, M.S., Erin D. Fuller, Zhenhua Gu, M.S., Michael Lattari, Alexander Lopez, M.S., John D. Overton, Ph.D., Thomas D. Schleicher, M.S., Maria Sotiropoulos Padilla, M.S., Louis Widom, Sarah E. Wolf, M.S., Manasi Pradhan, M.S., Kia Manoochehri, Ricardo H. Ulloa.

Contribution: C.B., C.F., A.L., and J.D.O. performed and are responsible for sample genotyping. C.B, C.F., E.D.F., M.L., M.S.P., L.W., S.E.W., A.L., and J.D.O. performed and are responsible for exome sequencing. T.D.S., Z.G., A.L., and J.D.O. conceived and are responsible for laboratory automation. M.P., K.M., R.U., and J.D.O are responsible for sample tracking and the library information management system.

Clinical Informatics

Nilanjana Banerjee, Ph.D., Michael Cantor, M.D. M.A., Dadong Li, Ph.D., Deepika Sharma, MHI

Contribution: All authors contributed to the development and validation of clinical phenotypes used to identify study subjects and (when applicable) controls.

Genome Informatics

Xiaodong Bai, Ph.D., Suganthi Balasubramanian, Ph.D., Andrew Blumenfeld, Gisu Eom, Lukas Habegger, Ph.D., Alicia Hawes, B.S., Shareef Khalid, Jeffrey G. Reid, Ph.D., Evan K. Maxwell, Ph.D., William Salerno, Ph.D., Jeffrey C. Staples, Ph.D.

Contribution: X.B., A.H., W.S. and J.G.R. performed and are responsible for analysis needed to produce exome and genotype data. G.E. and J.G.R. provided compute infrastructure development and operational support. S.B., and J.G.R. provide variant and gene annotations and their functional interpretation of variants. E.M., J.S., A.B., L.H., J.G.R. conceived and are responsible for creating, developing, and deploying analysis platforms and computational methods for analyzing genomic data.

Analytical Genetics

Gonçalo R. Abecasis, Ph.D., Joshua Backman, Ph.D., Manuel A. Ferreira, Ph.D., Lauren Gurski, Jack A. Kosmicki, Ph.D., Alexander Li, Ph.D., Adam Locke, Ph.D., Anthony Marcketta, Jonathan Marchini, Ph.D., Joelle Mbatchou, Ph.D., Shane McCarthy, Ph.D., Colm O'Dushlaine, Ph.D., Dylan Sun, Kyoko Watanabe, Ph.D.

Contribution: J.A.K. and M.A.F. performed association analyses and led manuscript writing group. J.B. identified low-quality variants in exome sequence data using machine learning. L.G. and K.W. helped with visualization of association results. A.Li., A.L., A.M. and D.S. prepared the analytical pipelines to perform association analyses. J.M. and J.M. developed and helped deploy REGENIE. S.M. and C.O'D. helped defined COVID-19 phenotypes. G.R.A. supervised all analyses. All authors contributed to and reviewed the final version of the manuscript.

Immune, Respiratory, and Infectious Disease Therapeutic Area Genetics

Julie E. Horowitz, PhD.

Contribution: J.E.H. helped defined COVID-19 phenotypes, interpret association results and led the manuscript writing group.

Research Program Management

Marcus B. Jones, Ph.D., Michelle LeBlanc, Ph.D., Jason Mighty, Ph.D., Lyndon J. Mitnaul, Ph.D.

Contribution: All authors contributed to the management and coordination of all research activities, planning and execution. All authors contributed to the review process for the final version of the manuscript.

UK Biobank Exome Sequencing Consortium Research Team

[1]Bristol Myers Squibb

Oleg Moiseyenko, Carlos Rios, Saurabh Saha

[2]Regeneron Pharmaceuticals Inc.

Listed in pages 38 to 40.

[3]Biogen Inc.

Sally John, Chia-Yen Chen, David Sexton, Paola G. Bronson, Christopher D. Whelan, Varant Kupelian, Eric Marshall, Timothy Swan, Susan Eaton, Jimmy Z. Liu, Stephanie Loomis, Megan Jensen, Saranya Duraisamy, Ellen A. Tsai, Heiko Runz

[4]Alnylam Pharmaceuticals

Aimee M. Deaton, Margaret M. Parker, Lucas D. Ward, Alexander O. Flynn-Carroll, Greg Hinkle, Paul Nioi

[5]AstraZeneca

Caroline Austin (Business Development); Ruth March (Precision Medicine & Biosamples); Menelas N. Pangalos (BioPharmaceuticals R&D); Adam Platt (Translational Science & Experimental Medicine, Research and Early Development, Respiratory and Immunology); Mike Snowden (Discovery Sciences); Athena Matakidou, Sebastian Wasilewski, Quanli Wang, Sri

693 Deevi, Keren Carss, Katherine Smith (Centre for Genomics Research, Discovery Sciences,
694 BioPharmaceuticals R&D), Carolina Haefliger, Slavé Petrovski
695
696 [1]Bristol Myers Squibb, Route 206 and Province Line Road, Princeton, NJ 08543
697 [2]Regeneron Pharmaceuticals Inc., 777 Old Saw Mill River Road, Tarrytown, New York 10591
698 [3]Biogen Inc., 225 Binney Street, Cambridge, MA 02139
699 [4]Alnylam Pharmaceuticals, 675 West Kendall St, Cambridge, MA 02142
700 [5]AstraZeneca Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D,
701 Cambridge, UK